STATISTICAL AND ALGORITHM ASPECTS OF OPTIMAL PORTFOLIOS

DISSERTATION SUBMITTED TO THE INSTITUTE OF COMPUTATIONAL AND MATHEMATICAL ENGINEERING OF STANFORD UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Howard Howan Stephen Shek

March 2011

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Walter Murray, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Tze Lai, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Peter Hansen

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

We address three key aspects of optimal portfolio construction: expected return, variance-covariance modeling and optimization in presence of cardinality constraints. On expected return modeling, we extend the self-excited point process framework to model conditional arrival intensities of bid and ask side market orders of listed stocks. The cross-excitation of market orders is modeled explicitly such that the ask side market order size and bid side probability weighted order book cumulative volume can affect the ask side order intensity, and vice versa. Different variations of the framework are estimated by using method of maximum likelihood estimation, based on a recursive application of the log-likelihood functions derived in this thesis. Results indicate that the self-excited point process framework is able to capture a significant amount of the underlying trading dynamics of market orders, both in-sample and out-of-sample.

A new framework is introduced, *Realized GARCH*, for the joint modeling of returns and realized measures of volatility. A key feature is a measurement equation that relates the realized measure to the conditional variance of returns. The measurement equation facilitates a simple modeling of the dependence between returns and future volatility. *Realized GARCH* models with a linear or log-linear specification have many attractive features. They are parsimonious, simple to estimate, and imply an ARMA structure for the conditional variance and the realized measure. An empirical application with DJIA stocks and an exchange traded index fund shows that a simple *Realized GARCH* structure leads to substantial improvements in the empirical fit over standard GARCH models.

Finally we describe a novel algorithm to obtain the solution of the optimal portfolio problem with *NP*-hard cardinality constraints. The algorithm is based on a local relaxation that exploits the inherent structure of the objective function. It solves a sequence of small, local, quadratic-programs by first projecting asset returns onto a reduced metric space, followed by clustering in this space to identify sub-groups of assets that best accentuate a suitable measure of similarity amongst different assets. The algorithm can either be cold started using the centroids of initial clusters or be warm started based on the output of a previous result. Empirical result, using baskets of up to 3,000 stocks and with different cardinality constraints, indicates that the algorithm is able to achieve significant performance gain over a sophisticated branch-and-cut method. One key application of this *local relaxation* algorithm is in dealing with large scale cardinality constrained portfolio optimization under tight time constraint, such as for the purpose of index tracking or index arbitrage at high frequency.

Contents

List of Tables

List of Figures

1 Introduction

Portfolio optimization in the classical mean-variance optimal sense is a well studied topic (see Markowitz, 1952, 1987; Sharpe, 1964). It is relevant at all trading frequencies, whenever the agent has an exponential utility function, and whenever multiple return forecasts are aggregated into a portfolio. In high frequency finance, when transaction cost plays a key part in the eventual return of a strategy and when speed of transaction depends on the number of stocks in the portfolio, we need a framework to form a mean-variance optimal portfolio that bounds the total number of names to execute. Ultimately, the problem we aim to solve is a *cardinality-constrained quadratic program* (CCQP) with linear constraints, which can be expressed as

$$\min_{x,\widetilde{x}} \quad f(x) = c^\top x + \tfrac{1}{2} x^\top H x \tag{1.1}$$

$$\text{s.t.} \quad Ax = b \tag{1.2}$$

$$e^\top \widetilde{x} = K \tag{1.3}$$

where $c, x, e, \widetilde{x} \in \mathbb{R}^{n\times 1}$, $H \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^{m\times 1}$, $m \leq n$ and e is a vector of 1's. Let x^* be our solution set, the indicator function in (1.3) is given by

$$\widetilde{x}_i = \begin{cases} 1 & x_i \in x^* \\ 0 & \text{o.w.}. \end{cases}$$

The three key components in the objective function are the prediction of expected return, c, and variance-covariance matrix, H, and the specific algorithm for solving the optimization problem. This thesis addresses these three key aspects of optimal portfolio construction.

Based on Shek (2010), we investigate a self-excited point process approach to forecast expected return. Until recently, the majority of time series analyses related to financial data has been carried out using regularly spaced price data, with the goal of modeling and forecasting key distributional characteristics of future returns, such as expected mean and variance. These time series data mainly consist of daily closing prices, where comprehensive data are widely available for a large set of asset classes. With the recent rapid development of high-frequency finance, the focus has shifted to intra-day tick data, which record every transaction during market hours, and come with irregularly spaced time-stamps. We could resample the dataset and apply the same analyses as before, or we could try to explore additional information that the inter-arrival times may convey in terms of likely future trade direction. In order to properly take into account these irregular occurrences of transactions, we can adopt the framework of point process. In a doubly stochastic framework (see Bartlett, 1963), both the counting process and the driving intensity are stochastic. A point process is called self-excited if the current intensity of the events is determined by events in the past, see Hawkes (1971). It is widely accepted and observed that volatility of price returns tends to cluster. That is, a period of elevated volatility is likely to be followed by periods of similar levels of volatility. Trade arrivals also exhibit such clustering effect, for example a buy order is likely to be followed closely by another buy order. These orders tend to cluster in time. There has been a growing amount of literature on the application of point process to model inter-arrival trade durations, see Bowsher (2003) for a

comprehensive survey of the latest modeling frameworks. The proposed framework extends previous work on the application of self-excited process to model high frequency financial data. The extension comes in the form of a marked version of the process in order to take into account trade size influence on the underlying arrival intensity. In addition, by incorporating information from the *limit order book* (LOB), the proposed framework takes into account a measure of supply-demand imbalance of the market by parametrize the underlying based intensity as a function of this imbalance measure. The intensities, λ_{1t}, λ_{2t} for market bid and ask side orders, respectively, are given by the following,

$$\begin{cases} \lambda_{1t} = \mu_1 \bar{v}_{2t} + \frac{1}{\bar{w}_1} \sum_{t_i<t} \alpha_{11} w_{1i} e^{-\beta_{11}(t-t_i)} + \frac{1}{\bar{w}_2} \sum_{t_j<t} \alpha_{12} w_{2j} e^{-\beta_{12}(t-t_j)} \\ \\ \lambda_{2t} = \mu_2 \bar{v}_{1t} + \frac{1}{\bar{w}_2} \sum_{t_j<t} \alpha_{22} w_{2j} e^{-\beta_{22}(t-t_j)} + \frac{1}{\bar{w}_1} \sum_{t_i<t} \alpha_{21} w_{1i} e^{-\beta_{21}(t-t_i)}, \end{cases}$$

where t_i and t_j are $\mathcal{F}_t$-adapted jump times for bid and ask side market orders, respectively. This exponential parametrization is in reasonable agreement with empirical findings. The probability weighted volume for bid side orders (similarly for ask side orders), $\bar{v}_{1t} = \bar{v}(t, \tau, L; 1)$ is defined below,

$$\bar{v}(t, \tau, L; i) = \frac{1}{\sum_{i,l} v_{t,l;i}} \sum_{l=0}^{L} v_{t,l;i} p_{l,i,\tau},$$

where $p_{l,i,\tau} = \mathbb{P}(t_f < t + \tau | l, i)$ is the probability of an order of type $i \in \{1, 2\}$ submitted at layer l getting completely filled at time t_f, which is within τ seconds from order submission at time t. $v_{t,l;i}$ is the queue size at time t, at the l-th layer and on side i of the limit order book. Different variations of the framework are estimated by using method of maximum likelihood estimation, using a recursive application of the log-likelihood functions derived in this thesis. Results indicate that the self-excited point process framework is able to capture a significant amount of the underlying trading dynamics of market orders, both in-sample and out-of-sample.

Based on Hansen, Huang, and Shek (2011), we explore an innovative framework, *Realized GARCH*, that incorporates high frequency information in making better covariance forecast, by jointly model returns and realized measures of volatility. The latent volatility process of asset returns are relevant to a wide variety of applications, such as option pricing and risk management, and *generalized autoregressive conditional heteroskedasticity* (GARCH) models are widely used to model the dynamic features of volatility. This has sparked the development of a large number of *autoregressive conditional heteroskedasticity* (ARCH) and GARCH models since the seminal paper by Engle (1982). Within the GARCH framework, the key element is the specification for the conditional variance. GARCH models utilize daily returns (typically squared returns) to extract information about the current level of volatility, and this information is used to form expectations about the next period's volatility. A single return is unable to offer more than a weak signal about the current level of volatility. The implication is that GARCH models are poorly suited for situations where volatility changes rapidly to a new level, because the GARCH model is slow at "catching up" and it will take many periods for the conditional variance (implied by the GARCH model) to reach its new level. High-frequency financial data are now readily available and the literature has recently introduced a number of realized measures of volatility, including the realized variance, the bipower variation, the realized kernel, and many related quantities, see Andersen and Bollerslev (1998), Andersen, Bollerslev, Diebold, and Labys (2001b), Barndorff-Nielsen and Shephard (2002), Barndorff-Nielsen

and Shephard (2004), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008b), Hansen and Horel (2010), and references therein. Any of these measures is far more informative about the current level of volatility than is the squared return. This makes realized measures very useful for modeling and forecasting future volatility. Estimating a GARCH-X model that includes a realized measure in the GARCH equation provides a good illustration of this point. Such models were estimated by Engle (2002b) who used the realized variance. Within the GARCH-X framework no effort is paid to explain the variation in the realized measures, so these GARCH-X models are partial (incomplete) models that have nothing to say about returns and volatility beyond a single period into the future. Engle and Gallo (2006) introduced the first "complete" model in this context. Their model specifies a GARCH structure for each of the realized measures, so that an additional latent volatility process is introduced for each realized measure in the model. The model by Engle and Gallo (2006) is known as the *multiplicative error model* (MEM), because it builds on the MEM structure proposed by Engle (2002b). Another complete model is the HEAVY model by Shephard and Sheppard (2010) that, in terms of its mathematical structure, is nested in the MEM framework. Unlike the traditional GARCH models, these models operate with multiple latent volatility processes. For instance, the MEM by Engle and Gallo (2006) has a total of three latent volatility processes and the HEAVY model by Shephard and Sheppard (2010) has two (or more) latent volatility processes. Within the context of stochastic volatility models, Takahashi et al. (2009) were the first to propose a joint model for returns and a realized measure of volatility. The *Realized GARCH* framework introduced here combines a GARCH structure for returns with a model for realized measures of volatility. Models within our framework are called *Realized GARCH* models, a name that transpires both the objective of these models (similar to GARCH) and the means by which these models operate (using realized measures). A *Realized GARCH* model maintains the single volatility-factor structure of the traditional GARCH framework. Instead of introducing additional latent factors, we take advantage of the natural relationship between the realized measure and the conditional variance, and we will argue that there is no need for additional factors in many cases. Consider the case where the realized measure, x_t, is a consistent estimator of the integrated variance. Now write the integrated variance as a linear combination of the conditional variance and a random innovation, and we obtain the relation $x_t = \xi + \varphi h_t + \epsilon_t$. We do not impose $\varphi = 1$ so that this approach also applies when the realized measure is compute from a shorter period (e.g. 6.5 hours) than the interval that the conditional variance refers to (e.g. 24 hours). Having a measurement equation that ties x_t to h_t has several advantages. First, it induces a simple and tractable structure that is similar to that of the classical GARCH framework. For instance, the conditional variance, the realized measure, and the squared return, all have *autoregressive moving average* (ARMA) representations. Second, the measurement equation makes it simple to model the dependence between shocks to returns and shocks to volatility, that is commonly referred to as a leverage effect. Third, the measurement equation induces a structure that is convenient for prediction. Once the model is estimated it is simple to compute distributional predictions for the future path of volatilities and returns, and these predictions do not require us to introduce auxiliary future values for the realized measure. To illustrate our framework and fix ideas, consider a canonical version of the *Realized GARCH* model that will be referred to as the RealGARCH(1,1) model with a linear specification. This model is given by the following three

equations

$$\begin{aligned} r_t &= \sqrt{h_t} z_t, \\ h_t &= \omega + \beta h_{t-1} + \gamma x_{t-1}, \\ x_t &= \xi + \varphi h_t + \tau(z_t) + u_t, \end{aligned}$$

where r_t is the return, $z_t \sim \text{iid}(0,1)$, $u_t \sim \text{iid}(0, \sigma_u^2)$, and $h_t = \text{var}(r_t|\mathcal{F}_{t-1})$ with the filtration defined as $\mathcal{F}_t = \sigma(r_t, x_t, r_{t-1}, x_{t-1}, \ldots)$. The last equation relates the observed realized measure to the latent volatility, and is therefore called the *measurement equation*. It is easy to verify that h_t is an autoregressive process of order one, $h_t = \mu + \pi h_{t-1} + w_{t-1}$, where $\mu = \omega + \gamma\xi$, $\pi = \beta + \varphi\gamma$, and $w_t = \gamma\tau(z_t) + \gamma u_t$. So it is natural to adopt the nomenclature of GARCH models. The inclusion of the realized measure in the model and the fact that x_t has an ARMA representation motivate the name *Realized GARCH*. A simple, yet potent specification of the leverage function is $\tau(z) = \tau_1 z + \tau_2(z^2 - 1)$, which can generate an asymmetric response in volatility to return shocks. The simple structure of the model makes the model easy to estimate and interpret, and leads to a tractable analysis of the quasi maximum likelihood estimator. We apply the *Realized GARCH* framework to the DJIA stocks and an exchange traded index fund, SPY. We find, in all cases, substantial improvements in the log-likelihood function when benchmarked to a standard GARCH model. Substantial improvements are found in-sample as well as out-of-sample. The empirical evidence also strongly favors inclusion of the leverage function, and the parameter estimates are remarkably similar across stocks.

Based on Murray and Shek (2011), we combined the result from the first and second parts of the thesis and form the inputs to a cardinally constrained portfolio optimization problem, and explore two proposed innovative ways to solve this *NP*-hard problem in the most efficient way. The presence of cardinality constraint changes the complexity of the problem from that of an inequality constrained convex *quadratic program* (QP) to that of a non-convex QP in which the feasible region is a mixed-integer set with potentially many local optima. Shaw et al. (2008) has reduced a *3-partitioning* problem to a CCQP, hence establishing the *NP*-hardness of the problem. For these type of problems, even at modest sizes, computationally effective algorithms do not exist and, up until recently, there has been relatively little work presented in the literature. One of the current state of the art commercial solvers, the built-in *mixed integer QP* (MIQP) solver in CPLEX, uses branch-and-cut algorithm together with heuristics for solving large scale problems. The branch-and-cut algorithm is a combination of a branch-and-bound algorithm which uses a sophisticated divide and conquer approach to solve the problems by building a pruned tree, and a cutting plan method that improves the relaxation of the sub-problems to more closely approximate the integer programming problem. The proposed *global smoothing* algorithm is a prototype algorithm that introduces penalty and global smoothing functions to the original problem, then iteratively increase the penalty factor while decrease the amount of smoothing introduced, until convergence. The proposed *local relaxation* algorithm solves a sequence of small, local, quadratic-programs by first projecting asset returns onto a reduced metric space, followed by clustering in this space to identify sub-groups of assets that best accentuate a suitable measure of similarity amongst different assets. The algorithm can either be cold started using the centroids of initial clusters or be warm started based on the output of a previous result. Empirical result, using baskets of up to 3,000 stocks and with different cardinality constraints, indicates that the proposed *local relaxation* algorithm is able to achieve significant performance gain

over a sophisticated branch-and-cut method. One key application of this algorithm is in dealing with large scale cardinality constrained portfolio optimization under tight time constraint, such as for the purpose of index tracking or index arbitrage at high frequency.

For the empirical parts of this thesis, ideally we would like to use a single dataset that covers the necessary sample durations for the different components of our overall framework. However, availability of data that spans all dimensions (order book detail, transaction prices, tick sampling) is limited at the time of writing, which necessitates parts of the empirical study using different datasets. Throughout the thesis it will be made precise which dataset is used for empirical work in each section.

This thesis is organized as follows. Section 2 focuses on estimation and prediction of the covariance matrix, H, by incorporating high frequency asset return information. The key model is RealGARCH(1,1). Section 3 focuses on a prediction framework for the expected return, c, again by using high frequency information contents, in the form of tick-by-tick return series and limit order book dynamics. The key model here is a multivariate self-excited stochastic intensity process. Section 4 combines the frameworks from Sections 2 and 3 to give the distributional parameters in the objective function for the final optimization problem. This final chapter introduces two new approaches, *global smoothing* and *local relaxation* with factor projection, to solve the underlying cardinality constrained optimization problem.

2 Covariance Matrix Estimation and Prediction

For the most part of this section, our focus will be on the single variate case, i.e. on the variance of some random process. The extension from single to multivariate, hence the covariance matrix, will be introduced towards the end of the section, after fixing some key ideas and addressing the challenges of univariate variance estimation and prediction.

2.1 Literature Review

Define a filtered probability space $(\Omega, \mathcal{F}_t, P)$, and let the true latent arbitrage free log-price process, $X_t \in \mathbb{R}$, be a any $\mathcal{F}_t$-adapted semi-martingale in this space. One such semi-martingale is the simple jump diffusion process given by,

$$dX_t = \mu_t dt + \sigma_t dW_t + \kappa_t dJ_t. \tag{2.1}$$

The *quadratic variation* (QV) for this jump diffusion process is defined to be

$$\langle X, X \rangle_T = \underbrace{\int_0^T \sigma_t^2 dt}_{\text{integrated volatility}} + \underbrace{\sum_{0<t\leq T} \kappa_t^2}_{\text{jump volatility}}.$$

For the most parts of the subsequent analysis, we ignore the jump part and focus on the *geometric Brownian motion* (GBM),

$$dX_t = \mu_t dt + \sigma_t dW_t \tag{2.2}$$

with the corresponding definition below.

Definition 1. The *integrated variation* (IV) or *quadratic variation* (QV) of the GBM process of (2.2) is defined to be

$$\langle X, X \rangle_T = \int_0^T \sigma_t^2 dt. \tag{2.3}$$

When the underlying process is observed in discrete time, we have two sets of measures for QV, depending on our assumption of market microstructure noise. Section 2.1.1 studies a set of QV estimators assuming no noise, followed by Section 2.1.2 that introduces a number of estimators that takes into account various specifications of the market microstructure noise.

2.1.1 High Frequency Volatility Estimators in Absence of Noise

Realized variance at tick frequency An intuitive practice is to estimate the variance from the sum of the frequently sampled squared returns. The estimator is defined to be

$$[X, X]_T = \sum_{t_i} \left(X_{t_{i+1}} - X_{t_i} \right)^2. \tag{2.4}$$

Under model (2.2), the approximation in (2.4) has the following property

$$\text{plim} \sum_{t_i} \left(X_{t_{i+1}} - X_{t_i} \right)^2 \to \int_0^T \sigma_t^2 dt$$

as the sampling frequency increases. Although this approach is justified under the assumption of a continuous stochastic model in an idealized world, it runs into a number of challenges when in presence of market microstructure in practical applications. It has been found empirically (see for example Hansen and Lunde, 2003) that the realized volatility estimator is not robust when the sampling interval is small, which gives raise to issues such as large bias in the estimate and non-robustness to changes in the sampling interval. In other words, since the observation noise is not necessarily *cadlag* or have bounded variation, the observed log return is not in fact a semi-martingale.

Realized variance (RV) at sub tick frequency with regularly spaced sampling (Andersen et al., 2001c) When the sampling frequency over the period $[0, T]$ is equally spaced at T/Δ , the realized variance becomes

$$[X, X]_T = \sum_{j=1}^{T/\Delta} \left(X_{t_{\Delta j}} - X_{t_{\Delta(j-1)}} \right)^2 .$$

Here, we essentially throw away a large fraction of the available data by sampling less frequently from the underlying high-frequency tick prices. This approach reduces the impact of microstructure noise, without quantifying and correcting its effect for volatility estimation.

Bi-power variation (BV) (Barndorff-Nielsen and Shephard, 2004) To isolate and measure the integrated volatility in a process with possible jumps as in (2.1), Barndorff-Nielsen and Shephard (2004) proposed the *bi-power variance* (BV) estimator in the form

$$BV_t(\Delta) = \frac{\pi}{2} \sum_{j=1}^{T/\Delta - 1} \left| X_{t_{j+\Delta}} - X_{t_j} \right| \left| X_{t_j} - X_{t_{j-\Delta}} \right| ,$$

which converges to QV in absence of noise.

2.1.2 High Frequency Volatility Estimators in Presence of Market Microstructure Noise

Measures introduced in Section 2.1.1 do not take into account the fact that true asset price dynamics are often not directly obtainable. Suppose that the log-price process as observed at the sampling times is of the form

$$Y_{t_i} = X_{t_i} + \epsilon_{t_i}, \tag{2.5}$$

where X_t is a latent true, or efficient, log-price process and ϵ_{t_i} is known as the market microstructure noise process. The source of this noise, ϵ_{t_i}, could come from

- ▷ Bid-ask bounce, where the traded price alternate between the bid and ask price multiple times in a short period of time, which in term induces quadratic variation not inherent in the true price process;
- ▷ Aggregation across different *Electronic Communication Networks* (ECN) that leads to synchronicity issues;
- ▷ Delay of recording, closely related to above, where the timestamp for a transaction can lag behind the true transaction time due to latency issues at the data processing plant;

- ▷ Difference in trade size or information content of price changes;
- ▷ Gradual response to block trades;
- ▷ Strategic component of order flow and inventory control effects;
- ▷ Miss-recording, where price quote is erroneously recorded (i.e. zero prices, misplaced decimal);
- ▷ Post processing adjustments, where adjustments are introduced at the exchange or at the data vendor, such as extrapolation of last period's price during period of subdued activity.

In general, and throughout this analysis, the word "noise" usually refers to the noise induced by ϵ, and the word "discretization noise" for randomness due to the discretization (i.e. rounding error) effect in $[X,X]_T$ in evaluating $\langle X,X\rangle_T$.

We are interested in the implications of such a data generating process for the estimation of the volatility of the efficient log-price process as given in (2.2), using discretely sampled data on the transaction price process. At sampling frequency measured in seconds rather than minutes or hours, the drift, μ_t, is irrelevant, both economically and statistically, and so we shall focus on the functional form of σ_t and set $\mu_t \equiv 0$ throughout.

The observable return over a small time interval, $\Delta = t_2 - t_1$, can be expressed as

$$Y_{t_2} - Y_{t_1} = \int_{t_1}^{t_2} \sigma_t dW_t + \epsilon_{t_2} - \epsilon_{t_1} \longrightarrow X_{t_2} - X_{t_1} + \epsilon_{t_2} - \epsilon_{t_1}.$$

Unlike the return process, there is no reason to believe that noise should approach zero with increasing sampling frequency, so we see that the noise to signal ratio increases with sampling frequency.

A question that naturally arises is why we are interested in the quadratic variation of X rather then the observed process Y, given that the process Y is the one that we actually see, and therefore trade on, in practice. The following key arguments are given by Aït-Sahalia et al. (2005),

- ▷ the variation of ϵ's is tied to each transaction, as opposed to the price process of the underlying security. From the standpoint of trading, the ϵ's represent trading costs, which are different from the costs created by the volatility of the underlying process. This cost varies between different market agents;
- ▷ continuous finance would be difficult to implement if we were to use the QV estimated by $[Y,Y]$, which depends on the data frequency;

Throughout the subsequent subsections on non-parametric estimators, we adopt the following notation, with additional, or slight change of, notations introduced as and when necessary.

Definition 2. *Estimator for QV* as defined in (2.3), based on model Φ is denoted by $\widehat{\langle X,X\rangle}_T^{(\Phi)}$.

Definition 3. *Realized Variance* based on every j'th observed log-price process, starting with observation number r is denoted by:

$$[Y,Y]_T^{(j,r)} = \sum_{0\le j(i-1)\le n-r-j} \left(Y_{t_{ji+t}} - Y_{t_{j(i-1)+t}}\right)^2.$$

Tick sampling estimator If we use all the data at the highest sampling frequency, i.e. tick-by-tick, then under the assumption of *iid* noise, we have

$$\widehat{\langle X,X\rangle}_T^{(tick)} = [Y,Y]_T^{(\Delta_{tick},1)},$$

where Δ_{tick} is the stochastic inter-arrival time between each tick arrival. Under the assumption of serially correlated noise, we obtain the asymptotic distribution (Aït-Sahalia et al., 2009)

$$\begin{aligned}
\widehat{\langle X,X\rangle}_T^{(tick)} &\overset{\mathcal{L}}{\approx} \underbrace{[X,X]_T}_{\text{QV}} + \underbrace{2n\left(E\left[\epsilon^2\right] + E\left[\epsilon_{t_0}\epsilon_{t_1}\right]\right)}_{\text{bias due to noise}} + \left(\underbrace{\underbrace{4n\Omega_\infty}_{\text{due to noise}} + \underbrace{\frac{2T}{n}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}}_{\text{total variance}} \right)^{1/2} Z_{total} \\
Z_{total} &\sim \mathcal{N}(0,1)
\end{aligned}$$

where the asymptotic variance, from standard formula for mixing sums, is given by

$$\Omega_\infty = Var\left\{(\epsilon_1-\epsilon_0)^2\right\} + 2\sum_{i=1}^{\infty} Cov\left\{(\epsilon_1-\epsilon_0)^2, (\epsilon_{i+1}-\epsilon_i)^2\right\}.$$

Note that in the case of *iid* noise, we have

$$\widehat{\langle X,X\rangle}_T^{(tick)} = \underbrace{[X,X]_T}_{\text{object of interest}} + \underbrace{2nE\left[\epsilon^2\right]}_{\text{due to noise}}.$$

That is, our estimator is not unbiased and this bias is significant compare to the signal, with the noise to signal ratio increasing linearly with n. Thus this realized variance estimator does not give the true integrated volatility $\langle X,X\rangle_T$, but rather the variance of the microstructure noise $E\left[\epsilon^2\right]$ scaled by $(2n)^{-1}$.

Sparse sampling estimator (Andersen et al., 2001a) This is based on a trade off between sampling more frequently to obtain more data points and less frequently to avoid data being overwhelmed by noise. Andersen et al. (2001a) suggest a sampling interval in the range from 5 to 30 minutes, so that the interval is short enough for the asymptotic of the measure to work well, and long enough that the market microstructure noise can be neglected.

Definition 4. The *sparse sampling estimator* is given by

$$\widehat{\langle X,X\rangle}_T^{(sparse)} = [Y,Y]_T^{(\Delta_{sparese},1)},$$

where $\Delta_{sparse} = T/n_{sparse}$.

For example, with $T = 1$ day, or 6.5 hours of open trading on the NYSE, and we start with data sampled on average $\overline{\Delta t} = 1$ second, then, for the full dataset, $n = T/\overline{\Delta t} = 23,400$; but once we sample sparsely every 5 minutes, then we sample every 300th observation, and $n_{sparse} = 78$.

Under the assumption of dependent noise, we obtain the asymptotic distribution (Aït-Sahalia et al., 2009)

$$\begin{aligned}\widehat{\langle X,X\rangle}_T^{(sparse)} &\overset{\mathcal{L}}{\approx} \underbrace{[X,X]_T}_{\text{QV}} + \underbrace{2n_{sparse}E\left[\epsilon^2\right]}_{\text{bias due to noise}} + \left(\underbrace{\underbrace{4n_{sparse}E\left[\epsilon^4\right]}_{\text{due to noise}} + \underbrace{\frac{2T}{n_{sparese}}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}}_{\text{total variance}}\right)^{1/2} Z_{total} \\ Z_{total} &\sim \mathcal{N}(0,1). \end{aligned} \tag{2.6}$$

It is possible to determine an optimal sampling frequency that minimize the MSE (Aït-Sahalia et al., 2005), which gives

$$n_{sparse}^* = \left(\frac{T}{4E\left[\epsilon^2\right]^2}\int_0^T \sigma_t^4 dt\right)^{1/3}.$$

Based on (2.6), we might be tempted to conclude that the optimal choice of n_{sparse} is to make it as small as possible. But that would overlook the fact that the bigger the $n_{sparese}$, the closer the $[X,X]_T$ to the target integrated variance $\langle X,X\rangle_T$, i.e. the smaller the discretization noise. An excessively sparse n_{sparse} has the effect of increasing the variance of the estimator via the discretization effect, which is proportional to n_{sparse}^{-1}, as indicated in (2.6).

Two scale realized volatility (TSRV) estimator (Aït-Sahalia et al., 2009) The TSRV estimator is based on a three step approach to ensure asymptotic unbiasness and efficiency:

1. Sub-sampling by partitioning the original grid of observation times, $G = \{t_0,\ldots,t_n\}$ into sub-samples, $G^{(k)}$, $k = 1,\ldots,K$ where $n/K \to \infty$ as $n \to \infty$. For example, for $G^{(1)}$ start at the the first observation and take an observation every Δ_{sparse} minutes, etc. This gives $[Y,Y]_T^{(\Delta_{sparse},k)}$;

2. Averaging the estimators obtained on the sub-samples, which gives

$$[Y,Y]_T^{(avg)} = \frac{1}{K}\sum_{k=1}^K [Y,Y]_T^{(\Delta_{sparse},k)},$$

   constructed by averaging the estimators $[Y,Y]_T^{(\Delta_{sparse},k)}$ obtained by sampling sparely on each of the K grids of average size $\overline{n} = n/K$.

3. Bias correction is obtained by the estimator for noise term $\widehat{E\left[\epsilon^2\right]} = \frac{1}{2n}[Y,Y]_T^{(all)}$.

Definition 5. The unbiased *small sampled adjusted TSRV estimator* with iid noise is given by

$$\widehat{\langle X,X\rangle}_T^{(tsrv,iid,adj)} = \underbrace{\left(1-\frac{\overline{n}}{n}\right)^{-1}}_{\text{small sample adj.}} \left\{\underbrace{[Y,Y]_T^{(avg)}}_{\text{slow time scale}} - \underbrace{\frac{\overline{n}}{n}[Y,Y]_T^{(all)}}_{\text{fast time scale}}\right\}.$$

With the number of sub-samples optimally selected as $K^* = cn^{2/3}$, then has the following distribution

$$\widehat{\langle X,X\rangle}_T^{(tsrv,iid,adj)} \overset{\mathcal{L}}{\approx} \underbrace{\langle X,X\rangle_T}_{\text{QV}} + \frac{1}{n^{1/6}} \left(\underbrace{\underbrace{\frac{8}{c^2} E\left[\epsilon^2\right]^2}_{\text{due to noise}} + \underbrace{c\frac{4T}{3}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}}_{\text{total variance}} \right)^{1/2} Z_{total}, \quad Z_{total} \sim \mathcal{N}(0,1).$$

We see that the estimator is unbiased and consistent.

Definition 6. The *unbiased TSRV estimator with dependent noise* is given by[1]

$$\widehat{\langle X,X\rangle}_T^{(tsrv,aa)} = \frac{n}{(K-J)\,\overline{n}_K} \left(\underbrace{[Y,Y]_T^{(K)}}_{\text{slow time scale}} - \frac{\overline{n}_K}{\overline{n}_J} \underbrace{[Y,Y]_T^{(J)}}_{\text{fast time scale}} \right),$$

for $1 \leq J < K \leq n$, where $[Y,Y]_T^{(J)} = \frac{1}{J}\sum_{r=0}^{J-1}[Y,Y]_T^{(j,r)} = \frac{1}{J}\sum_{i=0}^{n-J}\left(Y_{t_{i+J}} - Y_{t_i}\right)^2$ and $\overline{n}_K = (n-K+1)/K$.

With the number of sub-samples optimally selected as $K^* = cn^{2/3}$, then the estimator has the following distribution

$$\widehat{\langle X,X\rangle}_T^{(tsrv,aa)} \overset{\mathcal{L}}{\approx} \underbrace{\langle X,X\rangle_T}_{\text{QV}} + \frac{1}{n^{1/6}} \left(\underbrace{\underbrace{\frac{1}{c^2}\xi^2}_{\text{due to noise}} + \underbrace{c\frac{4T}{3}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}}_{\text{total variance}} \right)^{1/2} Z_{total}, \quad Z_{total} \sim N(0,1)$$

where ξ^2 depends on the rate the J and K approaches ∞ as $n \to \infty$. The two cases are

- ▷ $\limsup_{n\to\infty} \frac{J}{K} = 1$: $\xi^2 = \xi_\infty^2 = 8Var(\epsilon)^2 + 16\sum_{i=1}^{\infty} Cov(\epsilon_{t_o}, \epsilon_{t_i})^2$;
- ▷ $\limsup_{n\to\infty} \frac{J}{K} = 0$: $\xi^2 = \xi_\infty^2 + 4\alpha_0 + 8\sum_{i=1}^{\infty}\alpha_i$, where $\alpha_i = Cov\left(\epsilon_{t_0}, \epsilon_{t_{i+J}}\right) Cov\left(\epsilon_{t_i}, \epsilon_{t_J}\right) + Cum\left(\epsilon_{t_0}, \epsilon_{t_i}, \epsilon_{t_J}, \epsilon_{t_{i+J}}\right)$.

Multiple scale realized volatility (MSRV) estimator (Zhang, 2006) This improves upon the TRSV estimator's convergence rate of $n^{-1/6}$ to $n^{-1/4}$ at the cost of higher complexity. It essentially generalizes TSRV to multiple time scale, by averaging not on two time scales (J,K) but on multiple time scales.

[1] Assume strong mixing, such that $\exists \rho < 1$, such that $\left|Cov\left(\epsilon_{t_i}, \epsilon_{t_{i+L}}\right)\right| \leq \rho^l Var(\epsilon)$

Definition 7. The *MSRV estimator* is given by

$$\widehat{\langle X,X\rangle}_T^{(msrv)} = \underbrace{\sum_{i=1}^{M} a_i \left[Y,Y\right]_T^{(K_i)}}_{\text{slow time scale}} - \underbrace{\frac{1}{n}\left[Y,Y\right]_T^{(all)}}_{\text{fast time scale}},$$

where the weights are given by

$$a_i = \frac{i}{M} h\left(\frac{i}{M}\right) - \frac{1}{2M^2}\frac{i}{M} h'\left(\frac{i}{M}\right),$$

where $h \in C^1$ and satisfying $\int_0^1 xh\left(x\right)dx = 1$ and $\int_0^1 h\left(x\right)dx = 0$.

The asymptotic distribution is cumbersome and readers are referred to derivations given in the paper. Two points worth noting are

- ▷ Convergence of the MSRV remains $O_p\left(n^{-1/4}\right)$ even for dependent noise;
- ▷ Optimizing the overall variance of MSRV estimator leads to *realized kernel* (RK) estimator, discussed below.

Alternation (ALT) estimator (Large, 2007) Consider the observed price, Y_t, as a pure jump process with constant jump size, k, and whose deviation from the true latent process, X_t, are stationary in business time[2]. The proposed semi-parametric ALT estimator essentially scales the inconsistent but simple estimator nk^2, where n is the number of jumps in the quote, by a factor that takes into account Y_t's propensity to alternate.

The underlying assumptions of the proposed estimator are

- ▷ Y_t has uncorrelated alternations, where alternations are jumps whose direction is a reversal of the last jump;
- ▷ Y_t always jumps by a constant $\pm k$;
- ▷ $\epsilon_t = Y_t - X_t$ has no leverage effect, is stationary in business time and is *weakly mixing*;
- ▷ Y_t always jumps towards X_t;
- ▷ The *identification assumption* holds, such that

$$\left\{E\left[Y_{t_i}\middle|H_1\right] = E\left[Y_{t_i}\middle|H_2\right]\right\} \leftrightarrow \left\{E\left[X_{t_i}\middle|H_1\right] = E\left[X_{t_i}\middle|H_2\right]\right\},$$

  where $H_1 \in F_{t_i^-}$, $H_2 \in F_{t_i^-}$ and $H_2 \subset H_1$;

- ▷ The *buy-sell symmetry* holds, such that $(V - V_0, W) \overset{\mathcal{L}}{=} -(V - V_0, W)$.

Definition 8. The *ALT* estimator is given by

$$\widehat{\langle X,X\rangle}_T^{(alt)} = k^2 N_T \frac{C_T}{A_T},$$

[2]Business time: time of discrete observations; Calendar time: continue time in the normal sense.

where N_T, A_T and C_T are the number of jumps, number of alternations and number of continuations in $[0,T]$, where $N_T = A_T + C_T$. The asymptotic distribution is given by

$$\lim_{\alpha\to 0}\sqrt{N_t}\left(\frac{\widehat{\langle X,X\rangle}_T^{(alt)}}{\langle X,X\rangle_T}-1\right)\sim\mathcal{N}\left(0,UMU^\top\right),$$

where $U=\left(1,\frac{(1+R)^2}{R}\right)$, M is the long-run variance of $\Pi=\left\{\left(\frac{\langle X,X\rangle_{t_i}-\langle X,X\rangle_{t_{i-1}}}{E\left[\langle X,X\rangle_{t_i}-\langle X,X\rangle_{t_{i-1}}\right]},\frac{Q_i+1}{2}\right):i\in\mathbb{N}\right\}$, $Q_i=\{dA_{t_i}-dC_{t_i}:i\in\mathbb{N}\}\in\{-1,1\}$, and $R=\frac{\langle X,X\rangle_T}{E\left[\langle Y,Y\rangle_T\right]}$.

Realized kernel (RK) estimator (Barndorff-Nielsen et al., 2008b)

Definition 9. *Realized kernel* (RK) estimator for the quadratic variation of the observed log-price process Y_t, sampled at time $t_0, t_\Delta, \ldots, t_{n\Delta}$ is given by

$$\widehat{\langle X,X\rangle}_T^{(rk)}=\sum_{h=-H}^{H}\mathcal{K}\left(\frac{h}{H+1}\right)\sum_{j=|h|+1}^{T/\Delta}\left(Y_{\Delta j}-Y_{\Delta(j-1)}\right)\left(Y_{\Delta(j-h)}-Y_{\Delta(j-h-1)}\right),$$

where the *Parzen kernel* is often used, given by

$$\mathcal{K}(x)=\begin{cases}1-6x^2+6x^3 & 0\le x\le 1/2\\ 2(1-x)^3 & 1/2\le x\le 1\\ 0 & x>1.\end{cases}$$

It can be shown that the following asymptotic result holds (Barndorff-Nielsen et al., 2008b)

$$\widehat{\langle X,X\rangle}_T^{(rk)}\xrightarrow{p}\int_0^T\sigma_s ds,$$

where $H=cn^{3/5}$ gives the best trade-off between asymptotic bias and variance. The main advantage of the RK estimator is that it allow high frequency sampling by mitigating the noise effect by using a kernel smoother over the sampled period.

Bayesian filtering (BF) estimator (Zeng, 2003, 2004) Let $X(t)$ be the latent continuous value process for our assets. We have the following model setup:

- ▷ Trading times $t_1, t_2, \ldots, t_i \ldots$ are modeled by a conditional Poisson process, with intensity function denoted by $a(\theta(t), x(t), t)$, where $\theta(t)$ is the parameter set at time t;
- ▷ The observed prices $Y(t_i)$ at time t_i, is constructed via a random function $Y(t_i)=F(X_{t_i})$, where $y=F(x)$ is a random transformation with transition probability $p(y|x)$. Note the random functions takes into account the noise around the true value process X_t, plus other stylized facts such as clustering phenomenon (e.g. prices with integers and halves are most likely and odd quarters are less so, etc).

With the above framework, information affects X_t, and has a permanent influence on the price; while noise affects $F(x)$, the random transition function, and only has a transient impact on price.

Next, instead of viewing the prices $Y(t_i)$ in the order of trading occurrence over time, we model them as a collection of counting processes:

$$\vec{Y}(t) \triangleq \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} N_1\left\{\int_0^t \lambda_1(\theta(s), X(s), s)\, ds\right\} \\ N_2\left\{\int_0^t \lambda_2(\theta(s), X(s), s)\, ds\right\} \\ \vdots \\ N_n\left\{\int_0^t \lambda_n(\theta(s), X(s), s)\, ds\right\} \end{pmatrix} \tag{2.7}$$

where $y_k = N_k\left\{\int_0^t \lambda_k(\theta(s), X(s), s)\, ds\right\}$ is the counting process recording the cumulative number of trades that have occurred at the k-th price level up to time t, and $\lambda_k(\theta(s), X(s), s)$ is the corresponding intensity.

Under this representation, $(\theta(t), X(t))$ becomes the signal process, which cannot be observed directly, and $\vec{Y}(t)$ becomes the observation process, which is corrupted by market microstructure noise, modeled by $p(y|x)$. Hence $\left(\theta(t), X(t), \vec{Y}\right)$ is framed as a filtering problem with counting process observations.

Four mild assumption are necessary for the framework:

1. N_k's are unit Poisson processes under physical P-measure.
2. (θ, X), $N_1, N_2, \ldots, N_n$ are independent under P-measure, which implies that under suitable Q-measure, (θ, X), $Y_1, Y_2, \ldots, Y_n$ are independent.
3. The intensity can be expressed as $\lambda_k(\theta, X, t) = a(\theta, x, t)\, p(y_k|x)$, where $a(\theta, x, t)$ is the total intensity at time t and $p(y_k|x)$ is our previously defined transition probability from x to y_k, the k-th price level in (2.7). In other words, the total intensity determines the overall rate of trade occurrence at time t and $p(y_k|x)$ determines the proportional intensity of trade at the price level y_k, when the value is x.
4. The total intensity, $a(\theta, x, t)$, is uniformly bounded from above.

The core of the Bayesian estimation via filtering is constructing an algorithm to compute the conditional distribution, which becomes a posterior after a prior is assigned.

Let π_t be the conditional distribution of (θ, X) given $\mathcal{F}_t^{\vec{Y}}$ and let

$$\pi(f,t) = E^P\left[f(\theta(t), X(t))|\,\mathcal{F}_t^{\vec{Y}}\right] = \int f(\theta(t), X(t))\,\pi_t(d\theta, dx)$$

be the conditional expectation of $f(\theta(t), X(t))$ given $\mathcal{F}_t^{\vec{Y}}$. Then if we assume deterministic total intensity $a(\theta, x, t) = a(t)$, the normalized filtering equation is implied as

$$\pi(f,t) = \pi(f,0) + \int_0^t \pi(\boldsymbol{A}f, s)\, ds + \sum_{k=1}^n \int_0^t \left[\frac{\pi(fp_{y_k}, s-)}{\pi(p_k, s-)} - \pi(f, s-)\right] dY_k(s) \tag{2.8}$$

where $\boldsymbol{A}$ is the infinitesimal generator for the underlying stochastic equation of the value process X_t, and $p_k = p(y_k|x)$.

The filtering problem is solved by a recursive algorithm, outlined below:

1. Split parameters set $\theta(t)$ into constant and time-dependent parts, i.e. $\theta(t) = (\xi, \eta(t))$.

2. Discretize state space (ξ, η, X) with mesh $(\epsilon_\xi, \epsilon_\eta, \epsilon_x)$ and set $\epsilon = \max(|\epsilon_\xi|, |\epsilon_\eta|, |\epsilon_x|)$.

3. Then (2.8) can be expressed as

$$\pi_\epsilon(f,t) = \pi_\epsilon(f,0) + \int_0^t \pi_\epsilon(\boldsymbol{A_\epsilon} f, s)\, ds + \sum_{k=1}^{n} \int_0^t \left[\frac{\pi_\epsilon(f p_{p_k}, s-)}{\pi_\epsilon(p_k, s-)} - \pi(f, s-) \right] dY_{\epsilon,k}(s). \quad (2.9)$$

4. Given index spaces $\{\xi_j : j \in \mathcal{J}\}$, $\{\eta_m : m \in \mathcal{M}\}$ and $\{x_l : l \in \mathcal{L}\}$, we let

$$P_\epsilon(\xi_j, \eta_m, x_l; t) \triangleq P\left\{ \xi_\epsilon = \xi_j, \eta_\epsilon(t) = \eta_m, x_\epsilon(t) = x_l \middle| \mathcal{F}_t^{\vec{Y}_\epsilon} \right\}$$

and

$$\mathbf{1}_{\{\xi_\epsilon = \xi_j, \eta_\epsilon = \eta_m, x_\epsilon = x_l\}}(\xi_\epsilon, \eta_\epsilon, x_\epsilon) \triangleq \mathbf{1}(\xi_j, \eta_m, x_l)$$

then we have

$$\begin{cases} \pi(\mathbf{1}(\xi_j, \eta_m, x_l), t) & = P_\epsilon(\xi_j, \eta_m, x_l; t) \\ \pi(\mathbf{1}(\xi_j, \eta_m, x_l) p_k, t) & = P_\epsilon(\xi_j, \eta_m, x_l; t)\, p(y_k | x_l; \xi_j, \eta_m) \\ \pi_\epsilon(p_k, t) & = \sum_{j', m', l'} P_\epsilon(\xi_{j'}, \eta_{m'}, x_{l'}; t)\, p(y_k | x_{l'}; \xi_{j'}, \eta_{m'}) \end{cases}$$

5. Given the derivation in the previous step, we substitute the indicator function $\mathbf{1}$ for f in (2.9) to give us the final recursive equations for the posterior density of our parameter set.

Markov chain (MC) estimator (Hansen and Horel, 2010) Let X_t denote the observed process and Y_t the latent process[3], such that

$$X_t = Y_t + U_t,$$

where U_t is due to market microstructure noise plus the latent finite variation process, such as a *cadlag* jump process, inherent in Y_t. Define two filtrations, $G_t = \sigma(Y_s, U_s, s \le t)$ and $F_t = \sigma(X_s, s \le t)$, such that $F_t \subset G_t$ and Y_t is assumed to be a G_t-martingale.

Lemma 10. *If U_t is stationary with $E[|U_t|] < \infty$ and ϕ-mixing with respect to G_t, that is*

$$\phi(m) = \sup\{|P(A|B) - P(B)| : A \in \sigma(U_{t+s}, s \ge m), B \in G_t\} \to 0, \text{ as } m \to 0,$$

then $E[U_{t+h}|G_t] \xrightarrow{L_1} E[U_t]$ as $h \to \infty$.

Consider the G-filtered process

$$E\left[X_{T_{i+h}} \middle| G_{T_i}\right] = E\left[Y_{T_{i+h}} \middle| G_{T_i}\right] + E\left[U_{T_{i+h}} \middle| G_{T_i}\right]$$

which, under the G_{T_i}-martingale Y_t and zero mean U_t assumptions, can be simplified to give

$$E\left[X_{T_{i+h}} \middle| G_{T_i}\right] = Y_{T_i}. \quad (2.10)$$

[3]Note: this notational convention is opposite to our normal nomenclature. We maintain notations used in the original paper for ease of reference.

Define the standard in-fill asymptotic scheme

$$0 = T_0 < T_1 < \cdots < T_n = T$$

where $\sup_{1\leq i\leq n} |T_i - T_{i-1}| \to 0$ as $n \to \infty$. The h-steps filtered realized variance is given by

$$\begin{aligned} RV_F^{(h)} &= \sum_{i=1}^{n} \left\{ E\left[X_{T_{i+h}} \middle| G_{T_i} \right] - E\left[X_{T_{i+h-1}} \middle| G_{T_{i-1}} \right] \right\}^2 \\ &= \sum_{i=1}^{n} \left\{ \triangle X_{T_i} + \sum_{j=1}^{h} E\left[\triangle X_{T_{i+j}} \middle| G_{T_i} \right] - \sum_{j=1}^{h} E\left[X_{T_{i+j-1}} \middle| G_{T_{i-1}} \right] \right\}^2 . \end{aligned} \tag{2.11}$$

Note that under the natural, feasible, filtration for $\{X_t\}$, instead of (2.10) we obtain the result via tower property of expectation

$$\begin{aligned} E\left[X_{T_{i+h}} \middle| F_{T_i} \right] &= E\left[E\left[Y_{T_{i+h}} \middle| G_{T_i} \right] \middle| F_{T_i} \right] + E\left[E\left[U_{T_{i+h}} \middle| G_{T_i} \right] \middle| F_{T_i} \right] \\ &= E\left[Y_{T_i} \middle| F_{T_i} \right] + E\left[\widetilde{U}_{T_i} \middle| F_{T_i} \right] \\ &= E\left[X_{T_i} - U_{T_i} \middle| F_{T_i} \right] + E\left[\widetilde{U}_{T_i} \middle| F_{T_i} \right] \\ &= Y_{T_i} + U_{T_i} - E\left[U_{T_i} \middle| F_{T_i} \right] + E\left[\widetilde{U}_{T_i} \middle| F_{T_i} \right] \end{aligned}$$

where $\widetilde{U}_t = \lim_{h\to\infty} E\left[U_{t+h} \middle| G_t \right]$. Note that quadratic variations of Y_t and $E\left[Y_{T_i} \middle| F_{T_i} \right]$ will coincide only if the quadratic variation of $U_{T_i} - E\left[U_{T_i} \middle| F_{T_i} \right]$ is zero, which is key for this problem.

The central idea is to filter observed price process by Markov Chain framework, using the natural filtration for $\{X_t\}$, $F_t = \sigma\left(X_s,\, s \leq t \right)$. The realized variance of this filtered price then defines a novel estimator of the quadratic variation estimator. The estimator takes advantage of the fact that price increments are confined to a grid[4]. We initially assume that the observed price increments follow a homogeneous chain, then later show that the estimator is robust to inhomogeneity when k, the order of the Markov Chain, grows with sample size at a suitable rate.

Consider sample of high frequency prices, $X_{T_0}, \ldots, X_{T_n}$, where price increments,

$$\Delta X_{T_i} \in \{x_1, \ldots, x_S\},$$

are distributed as a homogeneous Markov chain of order k with the number of states given by S . Then consider the k-tuple, $\triangle \mathcal{X}_{T_i} = (\Delta X_{T-k+1}, \ldots, \triangle X_{T_i})$, and index the possible values for $\triangle \mathcal{X}_{T_i}$ by $\mathbf{x}_s$, $s = 1, \ldots, S^k$, where $\mathbf{x}_s \in \{x_1, \ldots, x_S\}^k \subset \mathbb{R}^k$. The transition matrix, P, is given by

$$P_{r,s} = Pr\left(\triangle \mathcal{X}_{T_{i+1}} = \mathbf{x}_s \middle| \triangle \mathcal{X}_{T_i} = \mathbf{x}_r \right).$$

We use the vector $f \in \mathbb{R}^{S^k}$ to keep track of the value of ΔX_{T_i}, with f_s being the last element of x_s, $s = 1, \ldots, S^k$. For a particular realization of $\triangle \mathcal{X}_{T_i}$, the conditional expectation of $\Delta X_{T_{i+1}}$ can be expressed as

$$E\left[\Delta X_{T_{i+h}} \middle| \triangle \mathcal{X}_{T_i} = \mathbf{x}_r \right] = \sum_{s=1}^{S^k} P_{r,s}^h f_s = \left(P^h F \right)_r .$$

[4]Note: the logarithm of price is not on the grid

Define the return of the h-steps filter log-price,

$$y^{(h)}\left(\triangle\mathcal{X}_{T_{i-1}},\triangle\mathcal{X}_{T_i}\right):=E\left[X_{T_{i+h}}\middle|\triangle\mathcal{X}_{T_i}\right]-E\left[X_{T_{i+h-1}}\middle|\triangle\mathcal{X}_{T_{i-1}}\right],$$

such that

$$\begin{aligned}y^{(h)}\left(\triangle\mathcal{X}_{T_{i-1}},\triangle\mathcal{X}_{T_i}\right) &= \triangle X_{T_i}+\sum_{j=1}^{h}E\left[\Delta X_{T_{i+j}}\middle|\triangle\mathcal{X}_{T_i}\right]-\sum_{j=1}^{h}E\left[\Delta X_{T_{i+j-1}}\middle|\triangle\mathcal{X}_{T_{i-1}}\right]\\ &= \triangle X_{T_i}+\sum_{j=1}^{h}\sum_{r=1}^{S}\left(P^jF\right)_r 1_{\left\{\triangle\mathcal{X}_{T_i}=\mathbf{x}_r\right\}}-\sum_{j=1}^{h}\sum_{r=1}^{S}\left(P^jF\right)_r 1_{\left\{\triangle\mathcal{X}_{T_{i-1}}=\mathbf{x}_r\right\}}.\end{aligned}$$

The contribution to RV_F, when $\left(\triangle\mathcal{X}_{T_{i-1}},\triangle\mathcal{X}_{T_i}\right)=(\mathbf{x}_r,\mathbf{x}_s)$, is simply given by $\left\{y^{(h)}(\mathbf{x}_r,\mathbf{x}_s)\right\}$. Let $n_{r,s}=\sum_{i=1}^{n}1_{\left\{\triangle\mathcal{X}_{T_{i-1}}=\mathbf{x}_r,\,\triangle\mathcal{X}_{T_i}=\mathbf{x}_s\right\}}$, then the Markov filtered realized variance is then given by

$$RV_F^{(h)}=\sum_{r,s}n_{r,s}\left\{y^{(h)}(\mathbf{x}_r,\mathbf{x}_s)\right\}^2.$$

We have the following expression for the filtered returns.

Lemma 11. *Let e_r denote the r-th unit vector. Then*

$$y^{(h)}(\mathbf{x}_r,\mathbf{x}_s)=e_r^\top\left(I-Z^{(h)}\right)f+e_s^\top Z^{(h)}f$$

with $Z^{(h)}=I+\sum_{j=1}^{h}\left(P^j-I\right)$ and

$$y(\mathbf{x}_r,\mathbf{x}_s)=\lim_{h\to\infty}y^{(h)}(\mathbf{x}_r,\mathbf{x}_s)=e_r^\top(I-Z)f+e_s^\top Zf$$

where Z is the fundamental matrix of the underlying Markov chain.

Focusing on the $h=\infty$ case and define $y_{(r,s)}=y(\mathbf{x}_r,\mathbf{x}_s)$, then the filtered realized variance (the infeasible estimator) is given by

$$RV_F=\sum_{r,s}n_{r,s}y_{(r,s)}^2.$$

The empirical transition matrix $\hat{P}$ is given by $\hat{P}_{r,s}={}^{n_{r,s}}/_{n_{r,\bullet}}$, where $n_{r,\bullet}=\sum_s n_{r,s}$. Then we have the feasible estimator

$$RV_{\hat{F}}=\sum_{r,s}n_{r,s}\hat{y}_{(r,s)}^2$$

where $\hat{y}_{(r,s)}^2=e_r^\top\left(I-\hat{Z}\right)f+e_r^\top\hat{Z}f$ and $\hat{Z}=\left(I-\hat{P}-\hat{\Pi}\right)^{-1}$.

Definition 12. Markov Chain estimator in price level is defined as

$$MC^{\#}:=n\left\langle f,\left(2\hat{Z}-I\right)f\right\rangle_{\hat{\pi}}$$

where the inner product is defined to be $\langle a,b\rangle_\pi=a^\top\Lambda_\pi b$ with $\Lambda_\pi=\mathrm{diag}(\pi_1,\pi_2,\ldots,\pi_{S^k})$. It can be shown (Theorem 4 in paper) that $RF_{\hat{F}}-MC^{\#}=O_p\left(n^{-1}\right)$, and if the first observe state coincides with the last observed state then $RV_{\hat{F}}=MC^{\#}$.

Markov Chain estimator in log-price is to a good approximation

$$MC = \frac{n^2 \left\langle f, \left(2\hat{Z} - I\right) f \right\rangle_{\hat{\pi}}}{\sum_{i=1}^{n} X_{T_i}^2}.$$

This approximation is good as long as X_{T_i} does not fluctuate dramatically. MC allows for faster computation and it preserves the asymptotic theory derived for $MC^{\#}$.

Robustness to inhomogeneity (for both $MC^{\#}$ and its standard deviation) is dealt by artificially increasing the order of the Markov chain. For example, a simulation study shows that estimation for a data generating process with a Markov chain of order one, a Markov chain of order two yields a consistent estimate for $MC^{\#}$, but the asymptotic variance of estimator increases with k. Note, although misspecification of the order two Markov chain leads to poor estimates of many population quantities, but it will accurately estimate the quantity we seek.

One special feature of the MC estimator is that it is a generalization of the ALT estimator proposed by Large (2007) (see Section 2.1.2). It can be shown that this is identical to $MC^{\#}$, with $k = 1$, and $S = 2$ with $f = (+\kappa, -\kappa)^{\top}$.

2.1.3 Volatility Prediction

Volatility is more predictable then the mean of the underlying return process. Stylized effects such as diurnal intraday pattern, and clustering at different time duration are prevalent in many assets and are well known and studied. In the subsections that follows, some key features of a number of forecasting frameworks are introduced. Table 1 on page 19 summaries their key features.

GARCH (Bollerslev, 1986; Engle, 1982) The latent volatility process of asset returns are relevant to a wide variety of applications, such as option pricing and risk management, and GARCH models are widely used to model the dynamic features of volatility. This has sparked the development of a large number of ARCH and GARCH models since the seminal paper by Engle (1982). Within the GARCH framework, the key element is the specification for the conditional variance.

GARCH models utilize daily returns (typically squared returns) to extract information about the current level of volatility, and this information is used to form expectations about the next period's volatility. A single return is unable to offer more than a weak signal about the current level of volatility. The implication is that GARCH models are poorly suited for situations where volatility changes rapidly to a new level, because the GARCH model is slow at "catching up" and it will take many periods for the conditional variance (implied by the GARCH model) to reach its new level.

Partial GARCH and MEM discussed in this Section, and the *Realized GARCH* in Section 2.2 are second generation of GARCH models that address these shortcomings, but still retain the essential features of the original framework.

UHF-GARCH (Engle, 2000) The proposed procedure is to model the associated variables, such as observed (i.e. latent value plus any measurement noise) returns (or *marks*) conditional on the arrival times, and then to *separately* model the arrival times. Below, the notations used in the original paper are retained for ease of comparison.

Define inter-trade duration

$$x_i = t_i - t_{i-1}$$

	Latent Variables†	Observable	Distribution‡
GARCH(1,1) (Bollerslev, 1986)	$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$	$r_t = \sqrt{h_t} z_t$	$z_t \sim \text{iid}\,\mathcal{N}(0,1)$
UHF-GARCH (Engle, 2000)	$\sigma_i^2 = \omega + \alpha e_{i-1}^2 + \beta \sigma_{i-1}^2 + \gamma x_{i-1}^{-1}$ $\psi_i = \widetilde{\omega} + \widetilde{\alpha} x_{i-1} + \widetilde{\beta} \psi_{i-1}$	$\frac{r_i}{\sqrt{x_i}} = \rho \frac{r_{i-1}}{\sqrt{x_{i-1}}} + e_i + \phi e_{i-1}$ $x_t = \psi_t \varepsilon_t$ $\sigma_i^2 = V_{i-1}\left(\frac{r_i}{\sqrt{x_i}} \middle\vert x_i \right)$	$e_i \sim \text{iid}\,\mathcal{N}(0,1)$ $\varepsilon_i \sim \text{iid}\,Exp(1)$
MEM (Engle & Gallo, 2006)	$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \delta r_{t-1} + \varphi R_{t-1}^2$ $h_{R,t} = \omega_R + \alpha_R R_{t-1}^2 + \beta_R h_{R,t-1} + \delta_R r_{t-1}$ $h_{RV,t} = \omega_{RV} + \alpha_{RV} RV_{t-1} + \beta_{RV} h_{RV,t-1} + \delta_{RV} r_{t-1} + \vartheta_{RV} RV_{t-1} 1_{(r_{t-1}<0)} + \varphi_{RV} r_{t-1}^2$	$r_t^2 = h_t z_t^2$ $R_t^2 = h_{R,t} z_{R,t}^2$ $RV_t = h_{RV,t} z_{RV,t}^2$	$\begin{pmatrix} z_t \\ z_{R,t} \\ z_{RV,t} \end{pmatrix} \sim \text{iid}\,\mathcal{N}(0, I)$
HEAVY (Shephard & Sheppard, 2009)	$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \gamma x_{t-1}$ $\mu_t = \omega_R + \alpha_R x_{t-1} + \beta_R \mu_{t-1}$	$r_t = \sqrt{h_t} z_t$ $x_t = \mu_t z_{RK,t}^2$	$\begin{pmatrix} z_t \\ z_{RK,t} \end{pmatrix} \sim \text{iid}\,\mathcal{N}(0, I)$
Realized GARCH (Hansen, Huang and Shek, 2009)	$h_t = \exp\{\omega + \beta \log h_{t-1} + \gamma \log x_{t-1}\}$	$r_t = \sqrt{h_t} z_t$ $\log x_t = \xi + \varphi \log h_t + \tau(z_t) + u_t$	$\begin{pmatrix} z_t \\ \frac{u_t}{\sigma_u} \end{pmatrix} \sim \text{iid}\,\mathcal{N}(0, I)$

Table 1: Key model features at a glance: The realized measures, R_t, RV_t, and x_t denote the intraday range, the realized variance, and the realized kernel, respectively. In the Realized GARCH model, the dependence between returns and innovations to the volatility (leverage effect) is modeled with $\tau(z_t)$, such as $\tau(z) = \tau_1 z + \tau_2(z^2 - 1)$, so that $\text{E}\tau(z_t) = 0$, when $z_t \sim (0,1)$. †The MEM specification listed here is that selected by Engle & Gallo (2006) using BIC (see their Table 4). ‡The distributional assumptions listed here are those used to specify the quasi log-likelihood function. (Gaussian innovations are not essential for any of the models).

and the corresponding k-marks, $y_i \in \mathbb{R}^k \subset \Xi$. The joint conditional density is given by

$$(x_i, y_i) | F_{i-1} \sim f(x_i, y_i | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_i),$$

where $\widetilde{x}_{i-1} = \{x_{i-1}, x_{i-2}, \ldots, x_1\}$, $\widetilde{y}_{i-1} = \{y_{i-1}, y_{i-2}, \ldots, y_1\}$ and θ is the parameters set for the model. Consider the underlying arrival as a non-homogeneous Poisson process, with conditional intensity

$$\begin{aligned} \lambda_i(t, \widetilde{x}_{i-1}, \widetilde{y}_{i-1}) &= \lim_{\Delta t \to 0} \frac{Pr(N(t+\Delta t) > N(t) | \widetilde{x}_{i-1}, \widetilde{y}_{i-1})}{\Delta t} \\ &= \frac{\int_{u \in \Xi} f(t - t_{i-1}, u | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_i)\, du}{\int\int_{s \geq t,\, u \in \Xi} f(s - t_{i-1}, u | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_i)\, du ds}. \end{aligned}$$

Define the component conditional densities

$$f(x_i, y_i | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_i) = g(x_i | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_{1i})\, q(y_i | x_i, \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_{2i}).$$

Then we obtain

$$\lambda_i(t, \widetilde{x}_{i-1}, \widetilde{y}_{i-1}) = \frac{g(t - t_{i-1} | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_{1i})}{\int_{s \geq t} g(s - t_{i-1} | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_{1i})\, ds}. \tag{2.12}$$

Model for inter-arrival time, x_i To model the inter-arrival time, x_i, Engle and Russell (1998) adopted the following framework:

Definition 13. *Autoregressive Conditional Duration (ACD)* (Engle and Russell, 1998) is given by

$$x_i = \psi_i \varepsilon_i \quad \varepsilon_i \sim iid(1,1) \tag{2.13}$$

$$\psi_i = \psi(\widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_i) = E[x_i | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_{1i}] = \int_\Omega x_i g(x_i | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_{1i})\, dx_i. \tag{2.14}$$

Note that (2.13) is powerful because it nests

- ▷ log linear model: $\log x_i = z_i \beta + w_i$, and
- ▷ Cox proportional hazard model (Cox, 1972): $\lambda(t, z) = \lambda_0(t) e^{-z\beta}$; so if $\lambda_0 \equiv$ constant, then we have (2.13).

Example. Conditional mean specification

$$\begin{aligned} x_i &= \psi_i \varepsilon_i,\ \varepsilon_i \sim Exp(1) \\ \psi_i &= \omega + \alpha x_{i-1} + \beta \psi_{i-1}. \end{aligned}$$

From (2.13) we have

$$\begin{aligned} g(x_i | \widetilde{x}_{i-1}, \widetilde{y}_{i-1}; \theta_{1i}) &= g(x_i = \varepsilon_i \psi_i | \psi_i; \theta_{1i}) \\ &= g\left(\varepsilon_i = \frac{x_i}{\psi_i} \middle| \psi_i; \theta_{1i}\right) \\ &= p_0\left(\varepsilon_i = \frac{x_i}{\psi_i} \middle| \theta_{11}\right). \end{aligned}$$

The density and survivor function ε, and the base intensity can be expressed

$$\begin{aligned} \varepsilon &\sim p_0(\bullet;\theta_{11}) \\ S_0(t;\theta_{11}) &= \int_{s>t} p_0(s;\theta_{11})\,ds \\ \lambda_0(t;\theta_{11}) &= \frac{p_0(t;\theta_{11})}{S_0(t;\theta_{11})}. \end{aligned}$$

Then (2.12) can be expressed by

$$\begin{aligned} \lambda_i(t,\widetilde{x}_{i-1},\widetilde{y}_{i-1}) &= \frac{p_0\left(\frac{t-t_{i-1}}{\psi_i};\theta_{11}\right)}{\int_{s\geq t} p_0\left(\frac{s-t_{i-1}}{\psi_i};\theta_{11}\right)ds} \\ &= \frac{p_0\left(\frac{t-t_{i-1}}{\psi_i};\theta_{11}\right)}{\psi_i \int_{\widetilde{s}\geq \frac{t-t_{i-1}}{\psi_i}} p_0(\widetilde{s};\theta_{11})\,d\widetilde{s}} \\ &= \lambda_0\left(\frac{t-t_{i-1}}{\psi_i};\theta_{11}\right)\frac{1}{\psi_i}. \end{aligned}$$

Note that once we have specified the density of ε and the functional form for ψ, then the conditional intensity is fully specified. For example, if we have $Pr(\varepsilon = x) = e^{-x}$, i.e. standard exponential density, then we have

$$\lambda_i = \frac{1}{\psi_i^2}\left(\frac{1}{e^{\frac{t-t_{i-1}}{\psi_i}}-1}\right).$$

Model for price volatility, σ_i Recall that the UHF-GARCH framework is a two step process of modeling arrival time, x_i, and price volatility, σ_i, separately. With x_i specified by (2.13) and (2.14), we are now in a position to model σ_i.

The return process, r_i, in terms of the observed price process, p_i, is given by

$$r_i = p_i - p_{i-1}.$$

The observed price process is related to the latent process, m_i, via

$$p_i = m_i + \zeta_i,$$

where ζ_i is error from truncation. Assume that the latent price, p_i, to be a Martingale with respect to public information with innovation that can be written, without loss of generality, proportional to the square root of the time. We have a duration adjusted return given by

$$\frac{r_i}{\sqrt{x_i}} = \frac{\Delta m_i}{\sqrt{x_i}} + \frac{\Delta\zeta_i}{\sqrt{x_i}} = \nu_i + \eta_i,$$

such that all relevant quantities are expressed as per unit of time. Consider a serially correlated[5] specification for the second term

$$\eta_i = \rho\eta_{i-1} + \xi_i + \chi\xi_{i-1},$$

[5] Autocorrelated since the truncation at one point in time is likely to be the same as the truncation several seconds later

then we arrive at an ARMA(1,1) model for our duration adjusted return given by

$$\frac{r_i}{\sqrt{x_i}} = \rho \frac{r_{i-1}}{\sqrt{x_{i-1}}} + e_i + \phi e_{i-1}, \tag{2.15}$$

where $e_i = \nu_i + \xi_i$.

Definition 14. Define the conditional variance per unit of time

$$V_{i-1}\left(\frac{r_i}{\sqrt{x_i}} \middle| x_i \right) = \sigma_i^2, \tag{2.16}$$

where the filtration for $V_{i-1}(\bullet)$ is $\sigma((x_i, r_i) : i = 0, 1, \ldots, i-1)$.

Compare (2.16) to the definition in classic GARCH for return $V_{i-1}(r_i | x_i) = h_i$, we see that $h_i = x_i \sigma_i^2$, hence σ_i^2 is the variance per unit time.

Empirical observations There are a number of ways to specify σ_i^2, once such specification (Eqn (40) in their paper) is given below

$$\sigma_i^2 = \omega + \alpha e_{i-1}^2 + \beta \sigma_{i-1}^2 + \gamma_i x_i^{-1} + \gamma_2 \frac{x_i}{\psi_i} + \gamma_3 \xi_{i-1} + \gamma_4 \psi_i^{-1}.$$

With the above fitted to data for IBM, a number of interesting observations is observed:

- ▷ *mean of* (2.15) *is negative* ⇒ "no news is bad news," (Diamond and Verrecchia, 1987);
- ▷ γ_1 *is positive* ⇒ "no trade means no news," hence low volatility (Easley and O'Hara, 1992);
- ▷ γ_2 *is negative* ⇒ mean reversion after surprise of trade;
- ▷ γ_4 *is positive* ⇒ high transaction means high volatility.

Partial GARCH (Engle, 2002b) High-frequency financial data are now readily available and the literature has recently introduce a number of realized measures of volatility, including the realized variance, the bipower variation, the realized kernel, and many related quantities, see Sections 2.1.1 and 2.1.2, and Barndorff-Nielsen and Shephard (2002), Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen et al. (2008b), Hansen and Horel (2010), and references therein. Any of these measures is far more informative about the current level of volatility than is the squared return. This makes realized measures useful for modeling and forecasting future volatility.

Let x_t be some observable time series that is referred to as the realized measure. For instance, x_t could be the realized variance computed from high-frequency data on day t. The realized variance is, like r_t^2, related to h_t, and this holds true for many high-frequency based measures. So it is natural to augment the standard GARCH model with x_t. A simple extension is to add the realized measure to the dynamic equation for the conditional variance,

$$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \gamma x_{t-1}. \tag{2.17}$$

This is known as the GARCH-X model, where the label "X" is due to x_t being treated as an exogenous variable. Engle (2002b) was the first to estimate this model with the realized variance

as the "exogenous" variable x_t. The GARCH-X model is referred to as a *partial-model* because it treats x_t as an exogenous variable. Within the GARCH-X framework no effort is paid to explain the variation in the realized measures, so these are partial (incomplete) models that have nothing to say about returns and volatility beyond a single period into the future. In order to complete the model we need a specification for x_t, which is given in Section 2.2. This model was also estimated by Barndorff-Nielsen and Shephard (2007) who used both the realized variance and the bi-power variation as the exogenous variable. The condition variance, h_t, is, by definition, adapted to $\mathcal{F}_{t-1}$. Thus, if $\gamma \neq 0$ then (2.17) implies that x_t is adapted to $\mathcal{F}_t$. A filtration that would satisfy this requirement is $\mathcal{F}_t = \sigma(r_t, x_t, r_{t-1}, x_{t-1}, \ldots)$, but $\mathcal{F}_t$ could in principle be an even richer σ-field.

MEM Engle (2002b) The Multiplicative Error Model (MEM) by Engle (2002b) was the first "complete" model in this context, see also Engle and Gallo (2006). This model specifies a GARCH structure for each of the realized measures, so that an additional latent volatility process is introduced for each realized measure in the model. Another complete model is the HEAVY model by Shephard and Sheppard (2010) that, in terms of its mathematical structure, is nested in the MEM framework. Unlike the traditional GARCH models, these models operate with multiple latent volatility processes. For instance, the MEM by Engle and Gallo (2006) has a total of three latent volatility processes and the HEAVY model by Shephard and Sheppard (2010) has two (or more) latent volatility processes.

2.1.4 Covariance Prediction

Model-free predictor A simple estimator would be to use the trailing 255 daily close-to-close return history to estimate the *variance-covariance* (VCV) matrix, with monthly update, i.e. the estimated VCV is then kept constant throughout the month, until the next update. An alternative would be to incorporate intraday information, using hourly open-to-close returns, where market-microstructure noise is less prevalent. For this alternative, we need to address the contribution of variance from activities outside of market hours. One obvious way to account for this systemic bias is to scale the open-to-close VCV. In Figure 2.1 we plot the monthly variance of daily close-to-close return against the monthly variance of hourly open-to-close return. Using simple *ordinary least squares* (OLS) regression, together with the fact that trading spans 6.5 hours during the day, the estimated the average per-hour ratios for open-to-close to overnight variance is in the range of 4.4 to 16.1[6]. This is in-line with estimated value of 12.78 by Oldfield and Rogalski (1980) and 13.2 by French and Roll (1986). We see that the value of this bias could vary significantly depending on the stock and, not shown here, it could also be a function of the sample period. This can be explained by considering the following:

- ▷ on one end of the spectrum, for an *American depositary receipt* (ADR) that is linked to a stock mainly traded on an Asian exchange, most of the information content would be captured in the close-to-open returns. Hence, we expect a smaller (likely to be less than 1) hourly open-to-close to overnight variance ratio;
- ▷ on the other end of the spectrum, for the stock of a company that generates its income mainly in the US, we would expect most of the price movement coming from trading during market

[6] Note that the higher the ratio, the more volatile the open-to-close return relative to close-to-open returns.

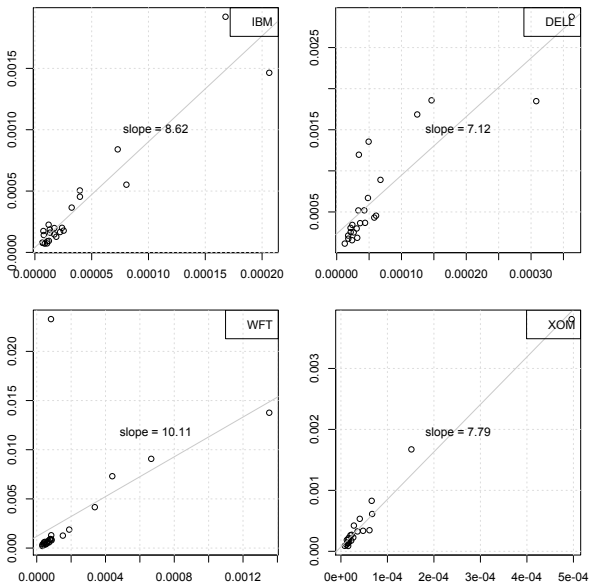

Figure 2.1: Monthly variance using daily close-to-close vs using hourly open-to-close. Sample period: 2008-02-02 to 2009-01-01. Slope is from simple OLS.

hours, hence a more volatile open-to-close return variance which leads to a higher variance ratio.

The simplest form of model free estimator incorporating higher frequency data is as follows

- ▷ calculate hourly open-to-close returns (based on tick or higher frequency data);
- ▷ calculate VCV matrix over a suitably chosen number of observations as the *look-back period*;
- ▷ scale VCV by stock specific scaling parameter, calculated over the same look-back period, to give an unbiased estimator of daily close-to-close VCV.

Moving average predictor The RiskMetrics conditional *exponentially weighted moving average* (EWMA) covariance estimator, based on daily observations, is given by

$$H_t = \alpha r_{t-1} r_{t-1}^\top + (1-\alpha) H_{t-1} \tag{2.18}$$

where the initial recursion value, H_0, can be taken to be the unconditional sample covariance matrix. RiskMetrics (1996) suggests a value for the smoothing parameter of $\alpha = 0.06$. The primary advantage of this estimator is clear: it has no parameters to estimate and is simple to implement. The obvious drawback is that it forces all assets to have the same smoothing coefficient irrespective of the asset's volatility dynamics. This will be addressed by the following section, where, with added complexity, we aim to address some of these asset idiosyncratic behaviors.

When applied to high-frequency data, it has often been observed that (2.18) gives a noisy estimator of the VCV matrix. Therefore, the RiskMetric EWMA estimator is adapted to high frequency observations as follows. Let the *daily* latent covariance estimator, H_τ, be given by

$$H_\tau = \alpha r_\tau r_\tau^\top + (1-\alpha) H_{\tau-1}$$

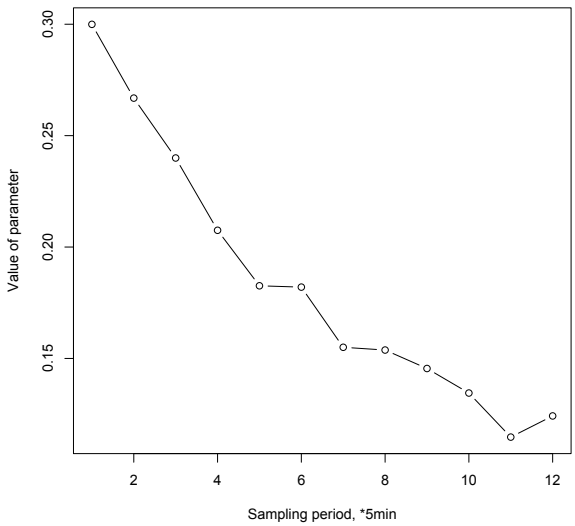

Figure 2.2: MLE fitted EMWA parameter as a function of sampling period; Gaussian density.

where $r_\tau r_\tau^\top$ is the variance-covariance of the return over the period, $\tau \in [t_{1,\tau}, t_{n,\tau}]$, $r_{1,\tau}$ is the first observed return for day τ and $r_{n,\tau}$ is the n-th and last observed return for day τ. In other words, $r_\tau r_\tau^\top$ is simply the high-frequency realized VCV for day τ. So we have one latent VCV estimator, H_τ, per day.

To estimate the EWMA parameters explicitly, we use MLE to maximize the joint Gaussian density. The fitted result gives $\alpha = 0.30$, i.e. significantly more weight on the high-frequency estimator, $r_{\tau-1} r_{\tau-1}^\top$, than that suggested by RiskMetric[7]. To quantify robustness of the EWMA estimator to the sampling frequency, the dataset is down-sampled to frequencies corresponding to sampling every 10min, 15min, 20min, 25min and 30min. Figure 2.2 shows the MLE fitted parameter as a function of sampling period. It can be seen that as sampling frequency decreases, there is a corresponding decrease in the persistency parameter. This is attributable to the decreasing information content in less frequently sampled return data.

DCC(1,1)-GARCH predictor (Engle, 2002c) The DCC-GARCH framework relies on two latent processes: conditional variance D_t^2, and conditional correlation R_t, where both $D_t, R_t \in \mathbb{R}^{n \times n}$ and D_t is diagonal, so that $H_t = D_t R_t D_t$. The conditional variance is given by

$$D_t^2 = \operatorname{diag}(\omega_i) + \operatorname{diag}(\kappa_i) \otimes r_{t-1} r_{t-1}^\top + \operatorname{diag}(\lambda_i) \otimes D_{t-1}^2, \tag{2.19}$$

where $r_t \in \mathbb{R}^n$ and $\otimes$ is the Hadamard product. The conditional correlation is given by

$$\begin{aligned} R_t &= \operatorname{diag}(Q_t)^{-1/2} Q_t \operatorname{diag}(Q_t)^{-1/2} \\ Q_t &= (1 - \alpha - \beta)\overline{Q} + \alpha \epsilon_{t-1} \epsilon_{t-1}^\top + \beta Q_{t-1} \end{aligned}$$

[7] The MLE fitted parameters assuming a *t(5)-distribution* are *(0.28, 0.72)*.

where $\overline{Q} = E\left[\epsilon_{t-1}\epsilon_{t-1}^{\top}\right]$ is the unconditional correlation matrix of the standardized residual. The log-likelihood for Gaussian innovation can be expressed as

$$\begin{aligned} L &= -\frac{1}{2}\sum_t n\log 2\pi + \log|D_t|^2 + r_t^{\top} D_t^{-2} r_t - \epsilon_t^{\top}\epsilon_t \\ &\quad + \log|R_t| + \epsilon_t^{\top} R_t^{-1}\epsilon_t. \end{aligned}$$

Parameter estimation is done via a two stage optimization procedure.

Stage One In the first stage, we maximize the log-likelihood of the conditional variance process,

$$\widehat{\theta} = \arg\max_{\omega,\kappa,\lambda}\left\{-\frac{1}{2}\sum_t\left(n\log 2\pi + \log|D_t|^2 + r_t^{\top} D_t^{-2} r_t\right)\right\}.$$

Stage Two In the second stage, we maximize the conditional correlation process, given the stage one result,

$$\max_{\alpha,\beta}\left\{-\frac{1}{2}\sum_t\left(\log|R_t| + \epsilon_t^{\top} R_t^{-1}\epsilon_t - \epsilon_t^{\top}\epsilon_t\right)\right\}.$$

In the next Section, we will extend this framework to take into account high frequency observations based on a new proposed univariate model for the conditional variance.

2.2 Complete Framework: Realized GARCH

2.2.1 Framework

We propose a new framework that combines a GARCH structure for returns with a model for realized measures of volatility. Models within this framework are called *Realized GARCH* models, a name that transpires both the objective of these models (similar to GARCH) and the means by which these models operate (using realized measures). A *Realized GARCH* model maintains the single volatility-factor structure of the traditional GARCH framework. Instead of introducing additional latent factors, we take advantage of the natural relationship between the realized measure and the conditional variance, and we will argue that there is no need for additional factors in many cases.

The general structure of the RealGARCH(p,q) model is given by

$$r_t = \sqrt{h_t} z_t, \tag{2.20}$$

$$h_t = v(h_{t-1}, \ldots, h_{t-p}, x_{t-1}, \ldots, x_{t-q}), \tag{2.21}$$

$$x_t = m(h_t, z_t, u_t), \tag{2.22}$$

where $z_t \sim iid\,(0,1)$ and $u_t \sim iid\,(0,\sigma_u^2)$, with z_t and u_t being mutually independent.

We refer to the first two equations as the *return equation* and the *GARCH equation*, and these define a class of GARCH-X models, including those that were estimated by Engle (2002a) and Barndorff-Nielsen and Shephard (2007). Recall that the GARCH-X acronym refers to the the fact that x_t is treated as an exogenous variable.

We shall refer to (2.22) as the *measurement equation*, because the realized measure, x_t, can often be interpreted as a measurement of h_t. The simplest example of a measurement equation is: $x_t = h_t + u_t$. The measurement equation is an important component because it "completes" the model. Moreover, the measurement equation provides a simple way to model the joint dependence between r_t and x_t, which is known to be empirically important. This dependence is modeled though the presence of z_t in the measurement equation, which we find to be highly significant in our empirical analysis.

It is worth noting that most (if not all) variants of ARCH and GARCH models are nested in the *Realized GARCH* framework. See Bollerslev (2009) for a comprehensive list of such models. The nesting can be achieved by setting $x_t = r_t$ or $x_t = r_t^2$, and the measurement equation is redundant for such models, because it is reduced to a simple identity.

The *Realized GARCH* model with a simple log-linear specification is characterized by the following GARCH and measurement equations.

$$\log h_t = \omega + \sum\nolimits_{i=1}^{p} \beta_i \log h_{t-i} + \sum\nolimits_{j=1}^{q} \gamma_j \log x_{t-j}, \tag{2.23}$$

$$\log x_t = \xi + \varphi \log h_t + \tau(z_t) + u_t, \tag{2.24}$$

where $z_t = r_t/\sqrt{h_t} \sim iid\,(0,1)$, $u_t \sim iid\,(0,\sigma_u^2)$, and $\tau(z)$ is called the *leverage function.*

Remark 15. A logarithmic specification for the measurement equation seems natural in this context. The reason is that (2.20) implies that

$$\log r_t^2 = \log h_t + \log z_t^2, \tag{2.25}$$

and a realized measure is in many ways similar to the squared return, r_t^2, albeit a more accurate measure of h_t. It is therefore natural to explore specifications where $\log x_t$ is expressed as a function of $\log h_t$ and z_t, such as (2.24). A logarithmic form for the measurement equation makes it convenient to specify the GARCH equation with a logarithmic form, because this induces a convenient ARMA structure, as we shall see below.

Remark 16. In our empirical application we adopt a quadratic specification for the leverage function, $\tau(z_t)$. The identity (2.25) motivated us to explore expressions that involves $\log z_t^2$, but these were inferior to the quadratic expression, and resulted in numerical issues because zero returns are occasionally observed in practice.

Remark 17. The conditional variance, h_t is, by definition, adapted to $\mathcal{F}_{t-1}$. Therefore, if $\gamma \neq 0$ then x_t must also be adapted to $\mathcal{F}_t$. A filtration that would satisfy this requirement is $\mathcal{F}_t = \sigma(r_t, x_t, r_{t-1}, x_{t-1}, \ldots)$, but $\mathcal{F}_t$ could in principle be an even richer σ-field.

Remark 18. Note that the measurement equation does not require x_t to be an unbiased measure of h_t. For instance, x_t could be a realized measure that is computed with high-frequency data from a period that only spans a fraction of the period that r_t is computed over. E.g. x_t could be the realized variance for a 6.5 hour long period whereas the return, r_t, is a close-to-close return that spans 24 hours. When x_t is roughly proportional to h_t, then we should expect $\varphi \approx 1$, and that is indeed what we find empirically. Both when we use open-to-close returns and close-to-close returns.

An attractive feature of the log-linear *Realized GARCH* model is that it preserves the ARMA structure that characterizes some of the standard GARCH models. This shows that the "ARCH" nomenclature is appropriate for the *Realized GARCH* model. For the sake of generality we derive the result for the case where the GARCH equation includes lagged squared returns. Thus consider the following GARCH equation,

$$\log h_t \quad = \quad \omega + \sum\nolimits_{i=1}^{p} \beta_i \log h_{t-i} + \sum\nolimits_{j=1}^{q} \gamma_j \log x_{t-j} + \sum\nolimits_{j=1}^{q} \alpha_j \log r_{t-j}^2, \tag{2.26}$$

where $q = \max_i\{(\alpha_i, \gamma_i) \neq (0,0)\}$.

Proposition 19. *Define* $w_t = \tau(z_t) + u_t$ *and* $v_t = \log z_t^2 - \kappa$, *where* $\kappa = \mathrm{E} \log z_t^2$. *The Realized GARCH model defined by (2.24) and (2.26) implies*

$$\begin{aligned}
\log h_t &= \mu_h + \sum_{i=1}^{p\vee q} (\alpha_i + \beta_i + \varphi\gamma_i) \log h_{t-i} + \sum_{j=1}^{q} (\gamma_j w_{t-j} + \alpha_j v_{t-j}), \\
\log x_t &= \mu_x + \sum_{i=1}^{p\vee q} (\alpha_i + \beta_i + \varphi\gamma_i) \log x_{t-i} + w_t + \sum_{j=1}^{p\vee q} \left\{ -(\alpha_j + \beta_j) w_{t-j} + \varphi\alpha_j v_{t-j} \right\}, \\
\log r_t^2 &= \mu_r + \sum_{i=1}^{p\vee q} (\alpha_i + \beta_i + \varphi\gamma_i) \log r_{t-i}^2 + v_t + \sum_{j=1}^{p\vee q} \left\{ \gamma_i (w_{t-j} - \varphi v_{t-j}) - \beta_j v_{t-j} \right\},
\end{aligned}$$

where $\mu_h = \omega + \gamma_\bullet \xi + \alpha_\bullet \kappa$, $\mu_x = \varphi(\omega + \alpha_\bullet \kappa) + (1 - \alpha_\bullet - \beta_\bullet)\xi$, *and* $\mu_r = \omega + \gamma_\bullet \xi + (1 - \beta_\bullet - \varphi\gamma_\bullet)\kappa$, *with*

$$\alpha_\bullet = \sum\nolimits_{j=1}^{q} \alpha_j, \qquad \beta_\bullet = \sum\nolimits_{i=1}^{p} \beta_i, \qquad \text{and} \qquad \gamma_\bullet = \sum\nolimits_{j=1}^{q} \gamma_j,$$

using the conventions $\beta_i = \gamma_j = \alpha_j = 0$ *for* $i > p$ *and* $j > q$.

So the log-linear Realized GARCH model implies that $\log h_t$ is ARMA$(p \vee q, q-1)$, whereas $\log r_t^2$ and $\log x_t$ are ARMA$(p \vee q, p \vee q)$. If $\alpha_1 = \cdots = \alpha_q = 0$, then $\log x_t$ is ARMA$(p \vee q, p)$.

From Proposition 19 we see that the persistence of volatility is summarized by a *persistence parameter*

$$\pi = \sum_{i=1}^{p\vee q} (\alpha_i + \beta_i + \varphi\gamma_i) = \alpha_\bullet + \beta_\bullet + \varphi\gamma_\bullet.$$

Example 20. By setting $x_t = r_t^2$, it is easy to verify that the RealGARCH(p,q) nests the GARCH(p,q) model. For instance with $p = q = 1$ we obtain the GARCH(1,1) structure with

$$\begin{aligned}
v(h_{t-1}, r_{t-1}^2) &= \omega + \alpha r_{t-1}^2 + \beta h_{t-1}, \\
m(h_t, z_t, u_t) &= h_t z_t^2.
\end{aligned}$$

The measurement equation is simply an identity in this case, i.e. we can take $u_t = 0$, for all t.

Example 21. If we set $x_t = r_t$, then we obtain the EGARCH(1,1) model by Nelson (1991) with

$$\begin{aligned}
v(h_{t-1}, r_{t-1}) &= \exp\left\{ \omega + \alpha |z_{t-1}| + \theta z_{t-1} + \beta \log h_{t-1} \right\}, \quad \text{since} \quad z_{t-1} = r_{t-1}/\sqrt{h_{t-1}}, \\
m(h_t, z_t, u_t) &= \sqrt{h_t} z_t.
\end{aligned}$$

Naturally, the interesting case is when x_t is a high-frequency based realized measure, or a vector containing several realized measures. Next we consider some particular variants of the *Realized GARCH* model.

Consider the case where the realized measure, x_t, is a consistent estimator of the integrated variance. Now write the integrated variance as a linear combination of the conditional variance and a random innovation, and we obtain the relation $x_t = \xi + \varphi h_t + \epsilon_t$. We do not impose $\varphi = 1$ so that this approach also applies when the realized measure is computed from a shorter period (e.g. 6.5 hours) than the interval that the conditional variance refers to (e.g. 24 hours). Having a measurement equation that ties x_t to h_t has several advantages. First, it induces a simple and tractable structure that is similar to that of the classical GARCH framework. For instance, the conditional variance, the realized measure, and the squared return, all have ARMA representations. Second, the measurement equation makes it simple to model the dependence between shocks to returns and shocks to volatility, that is commonly referred to as a leverage effect. Third, the measurement equation induces a structure that is convenient for prediction. Once the model is estimated it is simple to compute distributional predictions for the future path of volatilities and returns, and these predictions do not require us to introduce auxiliary future values for the realized measure.

To illustrate the framework and fix ideas, consider a canonical version of the *Realized GARCH* model that will be referred to as the RealGARCH(1,1) model with a log-linear specification. This model is given by the three equations

$$\begin{aligned} r_t &= \sqrt{h_t} z_t, \\ \log h_t &= \omega + \beta \log h_{t-1} + \gamma \log x_{t-1}, \\ \log x_t &= \xi + \varphi \log h_t + \tau(z_t) + u_t, \end{aligned}$$

where r_t is the return, $z_t \sim \text{iid}(0,1)$ and $u_t \sim \text{iid}(0, \sigma_u^2)$, and $h_t = \text{var}(r_t | r_{t-1}, x_{t-1}, r_{t-2}, x_{t-2}, \ldots)$. The last equation relates the observed realized measure to the latent volatility, and is therefore called the measurement equation. It is easy to verify that $\log h_t$ is an autoregressive process of order one, $\log h_t = \mu + \pi \log h_{t-1} + w_{t-1}$, where $\mu = \omega + \gamma\xi$, $\pi = \beta + \varphi\gamma$, and $w_t = \gamma\tau(z_t) + \gamma u_t$. So it is natural to adopt the nomenclature of GARCH models. The inclusion of the realized measure in the model and the fact that $\log x_t$ has an ARMA representation motivate the name *Realized GARCH*. A simple, yet potent specification of the leverage function is $\tau(z) = \tau_1 z + \tau_2(z^2 - 1)$, which can generate an asymmetric response in volatility to return shocks. The simple structure of the model makes the model easy to estimate and interpret, and leads to a tractable analysis of the quasi maximum likelihood estimator.

The MEM by Engle and Gallo (2006) utilizes two realized measures in addition to the squared returns. These are the intraday range (high minus low) and the realized variance, whereas the HEAVY model by Shephard and Sheppard (2010) uses the realized kernel (RK) by Barndorff-Nielsen et al. (2008b). These models introduce an additional latent volatility process for each of the realized measures. So the MEM and the HEAVY digress from the traditional GARCH models that only have a single latent volatility factor.

Unlike the MEM by Engle and Gallo (2006) and the HEAVY model by Shephard and Sheppard (2010), the Realized GARCH has the following characteristics.

▷ Maintains the single factor structure of latent volatility.

▷ Ties the realized measure directly to the conditional variance.

▷ Explicit modeling of the return-volatility dependence (leverage effect).

Key model features are given in Table 1.

Brownless and Gallo (2010) estimates a restricted MEM model that is closely related to the Realized GARCH with the linear specification. They utilize a single realized measure, the realized kernel by Barndorff-Nielsen et al. (2008b), so that they have two latent volatility processes, $h_t = \mathrm{E}(r_t^2|\mathcal{F}_{t-1})$ and $\mu_t = \mathrm{E}(x_t|\mathcal{F}_{t-1})$. However, their model is effectively reduced to a single factor model as they introduce the constraint, $h_t = c + d\mu_t$, see Brownless and Gallo (2010, Eqns. (6)-(7)). This structure is also implied by the linear version of our measurement equation. However, they do not formulate a measurement equation or relate $x_t - \mu_t$ to a leverage function. Instead they, for the purpose of simplifying the prediction problem, adopt a simple time-varying ARCH structure, $\mu_t = a_t + b_t x_{t-1}$, where a_t and b_t are defined by spline methods. Spline methods were introduced in this context by Engle and Rangel (2008) to capture the low-frequency variation in the volatility.

One of the main advantages of *Realized GARCH* framework is the simplicity by which dependence between return-shocks and volatility shocks is modeled with the leverage function. The MEM is formulated with a general dependence structure for the innovations that drive the latent volatility processes. The usual MEM formulation is based on a vector of non-negative random innovations, η_t, that are required to have mean $\mathrm{E}(\eta_t) = (1, \ldots, 1)'$. The literature has explored distributions with this property such as certain multivariate Gamma distributions, and Cipollini, Engle, and Gallo (2009) use copula methods that entail a very flexible class of distributions with the required structure. Some drawbacks of this approach include: estimation is complex; and a rigorous analysis of the asymptotic properties of these estimators seems intractable. A simpler way to achieve the structure in the multiplicative error distribution is by setting $\eta_t = Z_t \odot Z_t$, and work with the vector of random variables random variables, Z_t, instead. The required structure can be obtained with a more traditional error structure, where each element of Z_t is required to have zero mean and unit variance. This alternative formulation can be adopted without any loss of generality, since the dependence between the elements of Z_t is allow to take any form. The estimates in Engle and Gallo (2006) and Shephard and Sheppard (2010) are based on a likelihood where the elements of η_t are independent χ^2-distributed random variables with one degree of freedom. We have used the alternative formulation in Table 1 where $(z_t^2, z_{R,t}^2, z_{RV,t}^2)'$ corresponds to η_t in the MEM by Engle and Gallo (2006).

Leverage function and news impact The function $\tau(z)$ is called the leverage function because it captures the dependence between returns and future volatility, a phenomenon that is referred to as the leverage effect. We normalized such functions by $\mathrm{E}\tau(z_t) = 0$, and we focus on those that have the form

$$\tau(z_t) = \tau_1 a_1(z_t) + \cdots + \tau_k a_k(z_t), \qquad \text{where} \quad \mathrm{E}a_k(z_t) = 0, \quad \text{for all } k,$$

so that the function is linear in the unknown parameters. We shall see that the leverage function induces an EGARCH type structure in the GARCH equation, and we note that the functional form used in Nelson (1991), $\tau(z_t) = \tau_1 z + \tau_+(|z_t| - \mathrm{E}|z_t|)$ is within the class of leverage functions we

consider. We shall mainly consider leverage functions that are constructed from Hermite polynomials

$$\tau(z) = \tau_1 z + \tau_2(z^2 - 1) + \tau_3(z^3 - 3z) + \tau_4(z^4 - 6z^2 + 3) + \cdots ,$$

and our baseline choice for the leverage function is a simple quadratic form $\tau(z_t) = \tau_1 z_t + \tau_2(z_t^2 - 1)$. This choice is convenient because it ensures that $\mathrm{E}\tau(z_t) = 0$, for any distribution of z_t, so long as $\mathrm{E}z_t = 0$ and $\mathrm{var}(z_t) = 1$. The polynomial form is also convenient in our quasi likelihood analysis, and in our derivations of the kurtosis of returns generated by this model.

The leverage function $\tau(z)$ is closely related to the *news impact curve*, see Engle and Ng (1993), that maps out how positive and negative shocks to the price affect future volatility. We can define the news impact curve by

$$\nu(z) = \mathrm{E}(\log h_{t+1} | z_t = z) - \mathrm{E}(\log h_{t+1}),$$

so that $100\nu(z)$ measures the percentage impact on volatility as a function of the studentized return. From the ARMA representation it follows that $\nu(z) = \gamma_1 \tau(z)$.

Quasi-Maximum Likelihood Estimation (QMLE) Analysis In this section we discuss the asymptotic properties of the quasi-maximum likelihood estimator within the RealGARCH(p, q) model. The structure of the QMLE analysis is very similar to that of the standard GARCH model, see Bollerslev and Wooldridge (1992), Lee and Hansen (1994), Lumsdaine (1996), and Jensen and Rahbek (2004b,a). Both Engle and Gallo (2006) and Shephard and Sheppard (2010) justify the standard errors they report, by referencing existing QMLE results for GARCH models. This argument hinges on the fact that the joint log-likelihood in Engle and Gallo (2006) and Shephard and Sheppard (2010) is decomposed into a sum of univariate GARCH-X models, whose likelihood can be maximized separately. The factorization of the likelihood is achieved by two facets of these models: One is that all observables (i.e. squared return and each of the realized measures) are being tied to their individual latent volatility process. The other is that the primitive innovations in these models are taken to be independent in the formulation of the likelihood function. The latter inhibits a direct modeling of leverage effect with a function such as $\tau(z_t)$, which is one of the traits of the Realized GARCH model.

In this section we will derive the underlying QMLE structure for the log-linear *Realized GARCH* model. The structure of the linear *Realized GARCH* model is similar. We provide detailed expressions for the first and second derivatives of the log-likelihood function. These expressions facilitate direct computation of robust standard errors, and provide insight about regularity conditions that would justify QMLE inference. For instance, the first derivative will unearth regularity conditions that enables a central limit theorem be applied to the score function.

For the purpose of estimation, we adopt a Gaussian specification, so that the log likelihood function is given by

$$\ell(r, x; \theta) = -\frac{1}{2} \sum_{t=1}^{n} [\log(h_t) + r_t^2/h_t + \log(\sigma_u^2) + u_t^2/\sigma_u^2].$$

We write the leverage function as $\tau' a_t = \tau_1 a_1(z_t) + \cdots + \tau_k' a_k(z_t)$, and denote the parameters in the

model by

$$\theta = (\lambda', \psi', \sigma_u^2)', \qquad \text{where} \quad \lambda = (\omega, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)', \quad \psi = (\xi, \varphi, \tau')'.$$

To simplify the notation we write $\tilde{h}_t = \log h_t$ and $\tilde{x}_t = \log x_t$, and define

$$g_t = (1, \tilde{h}_{t-1}, \ldots, \tilde{h}_{t-p}, \tilde{x}_{t-1}, \ldots, \tilde{x}_{t-q})', \qquad m_t = (1, \tilde{h}_t, a_t')'.$$

So the GARCH and measurement equations can be expresses as

$$\tilde{h}_t = \lambda' g_t \qquad \text{and} \qquad \tilde{x}_t = \psi' m_t + u_t.$$

The dynamics that underlies the score and Hessian are driven by h_t and its derivatives with respect to λ. The properties of these derivatives are stated next.

Lemma 22. *Define* $\dot{h}_t = \frac{\partial \tilde{h}_t}{\partial \lambda}$ *and* $\ddot{h}_t = \frac{\partial^2 \tilde{h}_t}{\partial \lambda \partial \lambda'}$. *Then* $\dot{h}_s = 0$ and $\ddot{h}_s = 0$ for $s \leq 0$, and

$$\dot{h}_t = \sum_{i=1}^{p} \beta_i \dot{h}_{t-i} + g_t \qquad \textit{and} \qquad \ddot{h}_t = \sum_{i=1}^{p} \beta_i \ddot{h}_{t-i} + (\dot{H}_{t-1} + \dot{H}_{t-1}'),$$

where $\dot{H}_{t-1} = \left(0_{1+p+q\times 1}, \dot{h}_{t-1}, \ldots, \dot{h}_{t-p}, 0_{1+p+q\times q}\right)$ *is an* $p+q+1 \times p+q+1$ *matrix.*

(ii) When $p = q = 1$ *we have with* $\beta = \beta_1$ that

$$\dot{h}_t = \sum_{j=0}^{t-1} \beta^j g_{t-j} \qquad \textit{and} \qquad \ddot{h}_t = \sum_{k=1}^{t-1} k\beta^{k-1}(G_{t-k} + G_{t-k}'),$$

where $G_t = (0_{3\times 1}, g_t, 0_{3\times 1})$.

Proposition 23. *(i) The score,* $\frac{\partial \ell}{\partial \theta} = \sum_{t=1}^{n} \frac{\partial \ell_t}{\partial \theta}$, *is given by*

$$\frac{\partial \ell_t}{\partial \theta} = -\frac{1}{2} \begin{pmatrix} (1 - z_t^2 + \frac{2u_t}{\sigma_u^2}\dot{u}_t)\dot{h}_t \\ -\frac{2u_t}{\sigma_u^2} m_t \\ \frac{\sigma_u^2 - u_t^2}{\sigma_u^4} \end{pmatrix},$$

where $\dot{u}_t = \partial u_t / \partial \log h_t = -\varphi + \frac{1}{2} z_t \tau' \dot{a}_t$ *with* $\dot{a}_t = \partial a(z_t)/\partial z_t$.

(ii) The second derivative, $\frac{\partial^2 \ell}{\partial \theta \partial \theta'} = \sum_{t=1}^{n} \frac{\partial^2 \ell_t}{\partial \theta \partial \theta'}$, *is given by*

$$\frac{\partial^2 \ell_t}{\partial \theta \partial \theta'} = \begin{pmatrix} -\frac{1}{2}\left\{z_t^2 + \frac{2(\dot{u}_t^2 + u_t \ddot{u}_t)}{\sigma_u^2}\right\}\dot{h}_t \dot{h}_t' - \frac{1}{2}\left\{1 - z_t^2 + \frac{2u_t \dot{u}_t}{\sigma_u^2}\right\}\ddot{h}_t & \bullet & \bullet \\ \frac{\dot{u}_t}{\sigma_u^2} m_t \dot{h}_t' + \frac{u_t}{\sigma_u^2} b_t \dot{h}_t' & -\frac{1}{\sigma_u^2} m_t m_t' & \bullet \\ \frac{u_t \dot{u}_t}{\sigma_u^4} \dot{h}_t' & \frac{u_t}{\sigma_u^4} m_t' & \frac{1}{2}\frac{\sigma_u^2 - 2u_t^2}{\sigma_u^6} \end{pmatrix},$$

where $b_t = (0, 1, -\frac{1}{2} z_t \dot{a}_t')'$ and $\ddot{u}_t = -\frac{1}{4}\tau'\left\{z_t \dot{a}_t + z_t^2 \ddot{a}_t\right\}$ *with* $\ddot{a}_t = \partial^2 a(z_t)/\partial z_t^2$.

An advantage of our framework is that we can draw upon results for generalized hidden Markov models. Consider the case $p = q = 1$: From Carrasco and Chen (2002, Proposition 2) it follows that $\tilde{h}_t$ has a stationary representation provided that $\pi = \beta + \varphi\gamma \in (-1, 1)$, and if we assign $\tilde{h}_0$

its invariant distribution, then $\tilde{h}_t$ is strictly stationary and β-mixing with exponential decay, and $\mathrm{E}|\tilde{h}_t|^s < \infty$ if $\mathrm{E}|\tau(z_t) + u_t|^s < \infty$. Moreover, $\{(r_t, x_t), t \geq 0\}$ is a generalized hidden Markov model, with hidden chain $\{\tilde{h}_t, t \geq 0\}$, and so by Carrasco and Chen (2002, Proposition 4) it follows that also $\{(r_t, x_t)\}$ is stationary β-mixing with exponentially decay rate.

The robustness of the QMLE as defined by the Gaussian likelihood is, in part, reflected by the weak assumptions that make the score a martingale difference sequence. These are stated in the following Proposition.

Proposition 24. *(i) Suppose that* $\mathrm{E}(u_t|z_t, \mathcal{F}_{t-1}) = 0$, $\mathrm{E}(z_t^2|\mathcal{F}_{t-1}) = 1$, *and* $\mathrm{E}(u_t^2|\mathcal{F}_{t-1}) = \sigma_u^2$. *Then* $s_t(\theta) = \frac{\partial \ell_t(\theta)}{\partial \theta}$ *is a martingale difference sequence.*

(ii) Suppose, in addition, that $\{(r_t, x_t, \tilde{h}_t)\}$ *is stationary and ergodic. Then*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \ell_t}{\partial \theta} \xrightarrow{d} N(0, \mathcal{J}_\theta) \qquad \textit{and} \qquad -\frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \ell_t}{\partial \theta \partial \theta'} \xrightarrow{p} \mathcal{I}_\theta,$$

provided that

$$\mathcal{J}_\theta = \begin{pmatrix} \frac{1}{4}\mathrm{E}(1 - z_t^2 + \frac{2u_t}{\sigma_u^2}\dot{u}_t)^2 \mathrm{E}\left(\dot{h}_t \dot{h}_t'\right) & \bullet & \bullet \\ -\frac{1}{\sigma_u^2}\mathrm{E}\left(\dot{u}_t m_t \dot{h}_t'\right) & \frac{1}{\sigma_u^2}\mathrm{E}(m_t m_t') & \bullet \\ \frac{-E(u_t^3)E(\dot{u}_t)}{2\sigma_u^6} E(\dot{h}_t') & \frac{E(u_t^3)}{2\sigma_u^6} E(m_t') & \frac{E(u_t^2/\sigma_u^2 - 1)^2}{4\sigma_u^4} \end{pmatrix},$$

and

$$\mathcal{I}_\theta = \begin{pmatrix} \left\{\frac{1}{2} + \frac{\mathrm{E}(\dot{u}_t^2)}{\sigma_u^2}\right\} \mathrm{E}(\dot{h}_t \dot{h}_t') & \bullet & 0 \\ -\frac{1}{\sigma_u^2}\mathrm{E}\left\{(\dot{u}_t m_t + u_t b_t)\, \dot{h}_t'\right\} & \frac{1}{\sigma_u^2}\mathrm{E}(m_t m_t') & 0 \\ 0 & 0 & \frac{1}{2\sigma_u^4} \end{pmatrix},$$

are finite.

Note that in the stationary case we have $\mathcal{J}_\theta = \mathrm{E}\left(\frac{\partial \ell_t}{\partial \theta} \frac{\partial \ell_t}{\partial \theta'}\right)$, *so a necessary condition for* $|\mathcal{J}_\theta| < \infty$ *is that* z_t *and* u_t *have finite forth moments. Additional moments may be required for* z_t, *depending on the complexity of the leverage function* $\tau(z)$, *because* $\dot{u}_t$ *depends on* $\tau(z_t)$.

Theorem 25. *Under suitable regularity conditions, we have the asymptotic result.*

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to N\left(0, \mathcal{I}_\theta^{-1} \mathcal{J}_\theta \mathcal{I}_\theta^{-1}\right).$$

It is worth noting that the estimator of the parameters in the GARCH equation, λ, and those of the measurement equation, ψ, are not asymptotically independent. This asymptotic correlation is induced by the leverage function in our model, and the fact that we link the realized measure, x_t, to h_t with a measurement equation.

In the context of ARCH and GARCH models, it has been shown that the QMLE estimator is consistent with an Gaussian limit distribution regardless of the process being stationary or non-stationary. The latter was established in Jensen and Rahbek (2004b,a). So unlike the case for autoregressive processes, we do not have a discontinuity of the limit distribution at the knife-edge in the parameter space that separates stationary and non-stationary processes. This is an important result for empirical applications, because the point estimates are typically found to be very close to the boundary.

Notice that the martingale difference result for the score, Proposition 24(i), does not rely on stationarity. So it is reasonable to conjecture that the central limit theorem for martingale difference processes is applicable to the score even if the process is non-stationary.

While standard errors for $\hat{\theta}$ may be compute from numerical derivatives, these can also be computed directly using the following expressions

$$\hat{\mathcal{J}} = \frac{1}{n}\sum_{t=1}^{n} \hat{s}_t \hat{s}_t', \qquad \text{where} \qquad \hat{s}_t = \left\{ \tfrac{1}{2}(1 - \hat{z}_t^2 + \frac{2\hat{u}_t}{\hat{\sigma}_u^2}\hat{u}_t)\hat{\dot{h}}_t', -\frac{u_t}{\sigma_u^2}\hat{m}_t', \frac{\hat{\sigma}_u^2 - \hat{u}_t^2}{2\hat{\sigma}_u^4}) \right\}',$$

and

$$\begin{aligned}
\hat{\mathcal{I}} &= \frac{1}{n}\sum_{t=1}^{n} \begin{pmatrix} \frac{1}{2}\left\{\hat{z}_t^2 + \frac{2(\hat{u}_t^2 + \hat{u}_t\hat{\ddot{u}}_t)}{\hat{\sigma}_u^2}\right\}\hat{\dot{h}}_t\hat{\dot{h}}_t' + \frac{1}{2}\left\{1 - \hat{z}_t^2 + \frac{2\hat{u}_t\hat{\dot{u}}_t}{\hat{\sigma}_u^2}\right\}\hat{\ddot{h}}_t & \bullet & \bullet \\ -\hat{\sigma}_u^{-2}\left(\hat{\dot{u}}_t\hat{m}_t + \hat{u}_t\hat{b}_t\right)\hat{\dot{h}}_t' & \frac{1}{\hat{\sigma}_u^2}\hat{m}_t\hat{m}_t' & \bullet \\ -\frac{\hat{u}_t\hat{\dot{u}}_t}{\hat{\sigma}_u^4}\hat{\dot{h}}_t' & -\frac{\hat{u}_t}{\hat{\sigma}_u^4}\hat{m}_t' & \frac{1}{2}\frac{2\hat{u}_t^2 - \hat{\sigma}_u^2}{\hat{\sigma}_u^6} \end{pmatrix} \\
&= \frac{1}{n}\sum_{t=1}^{n} \begin{pmatrix} \frac{1}{2}\left\{\hat{z}_t^2 + \frac{2(\hat{u}_t^2 + \hat{u}_t\hat{\ddot{u}}_t)}{\hat{\sigma}_u^2}\right\}\hat{\dot{h}}_t\hat{\dot{h}}_t' + \frac{1}{2}\left\{1 - \hat{z}_t^2 + \frac{2\hat{u}_t\hat{\dot{u}}_t}{\hat{\sigma}_u^2}\right\}\hat{\ddot{h}}_t & \bullet & \bullet \\ -\hat{\sigma}_u^{-2}\left(\hat{\dot{u}}_t\hat{m}_t + \hat{u}_t\hat{b}_t\right)\hat{\dot{h}}_t' & -\frac{1}{\hat{\sigma}_u^2}\hat{m}_t\hat{m}_t' & \bullet \\ -\frac{\hat{u}_t\hat{\dot{u}}_t}{\hat{\sigma}_u^4}\hat{\dot{h}}_t' & 0 & \frac{1}{\hat{\sigma}_u^4} \end{pmatrix},
\end{aligned}$$

where the zero follows from the first order condition: $\sum_{t=1}^{n} \hat{u}_t\hat{m}_t' = 0$. Moreover, the first-order conditions for λ implies that $-\frac{\hat{u}_t\hat{\dot{u}}_t}{\hat{\sigma}_u^4}\hat{\dot{h}}' = \frac{1-\hat{z}_t^2}{2\hat{\sigma}_u^2}\hat{\ddot{h}}_t$.

For our baseline leverage function, $\tau_1 z_t + \tau_2(z_t^2 - 1)$, we have

$$m_t = \begin{pmatrix} 1 \\ \log h_t \\ z_t \\ z_t^2 - 1 \end{pmatrix}, \; b_t = \begin{pmatrix} 0 \\ 1 \\ -\frac{1}{2}z_t \\ -z_t^2 \end{pmatrix}, \quad \dot{u}_t = -\varphi + \frac{1}{2}\tau_1 z_t + \tau_2 z_t^2, \quad \ddot{u}_t = -\frac{1}{4}\tau_1 z_t - \tau_2 z_t^2.$$

Decomposition of likelihood function The log-likelihood function is (conditionally on $\mathcal{F}_0 = \sigma(\{r_t, x_t, h_t\}, t \leq 0)$) given by

$$\log L(\{r_t, x_t\}_{t=1}^n; \theta) = \sum_{t=1}^{n} \log f(r_t, x_t | \mathcal{F}_{t-1}).$$

Standard GARCH models do not model x_t, so the log-likelihood we obtain for these models cannot be compared to those of the *Realized GARCH* model. However, we can factorize the joint conditional density for (r_t, x_t) by

$$f(r_t, x_t | \mathcal{F}_{t-1}) = f(r_t | \mathcal{F}_{t-1}) f(x_t | r_t, \mathcal{F}_{t-1}),$$

and compare the partial log-likelihood, $\ell(r) := \sum_{t=1}^{n} \log f(r_t | \mathcal{F}_{t-1})$, with that of a standard GARCH model. Specifically for the Gaussian specification for z_t and u_t, we split the joint likelihood, into the

sum

$$\ell(r,x) \quad = \quad \underbrace{-\frac{1}{2}\sum_{t=1}^{n}[\log(2\pi)+\log(h_t)+r_t^2/h_t]}_{=\ell(r)} + \underbrace{-\frac{1}{2}\sum_{t=1}^{n}[\log(2\pi)+\log(\sigma_u^2)+u_t^2/\sigma_u^2]}_{=\ell(x|r)}.$$

Asymmetries in the leverage function are summarized by the following two statistics,

$$\rho^- = \text{corr}\{\tau(z_t)+u_t, z_t | z_t < 0\} \qquad \text{and} \qquad \rho^+ = \text{corr}\{\tau(z_t)+u_t, z_t | z_t > 0\}.$$

These capture the slope of a piecewise linear news impact curve for negative and positive returns, such as that implied by the EGARCH model.

Multiperiod forecasts One of the main advantages of having a complete specification, i.e., a model that fully describes the dynamic properties of x_t is that it multi-period ahead forecasting is feasible. In contrast, the GARCH-X model can only be used to make one-step ahead predictions. Multi-period ahead predictions are not possible without a model for x_t. Multi-period ahead predictions with the *Realized GARCH* model is straightforward. We let $\tilde{h}_t$ denote either h_t or $\log h_t$, such that the results presented in this section apply to both the linear and log-linear variants of the *Realized GARCH* model.

Consider first the case where $p = q = 1$. By substituting the GARCH equation into measurement equation we obtain the VARMA(1,1) structure

$$\begin{bmatrix} \tilde{h}_t \\ \tilde{x}_t \end{bmatrix} = \begin{bmatrix} \beta & \gamma \\ \varphi\beta & \varphi\gamma \end{bmatrix} \begin{bmatrix} \tilde{h}_{t-1} \\ \tilde{x}_{t-1} \end{bmatrix} + \begin{bmatrix} \omega \\ \xi + \varphi\omega \end{bmatrix} + \begin{bmatrix} 0 \\ \tau(z_t) + u_t \end{bmatrix},$$

that can be used to generate the predictive distribution of future values of $\tilde{h}_t$, $\tilde{x}_t$, as well as returns r_t, using

$$\begin{bmatrix} \tilde{h}_{t+h} \\ \tilde{x}_{t+h} \end{bmatrix} = \begin{bmatrix} \beta & \gamma \\ \varphi\beta & \varphi\gamma \end{bmatrix}^h \begin{bmatrix} \tilde{h}_t \\ \tilde{x}_t \end{bmatrix} + \sum_{j=0}^{h-1} \begin{bmatrix} \beta & \gamma \\ \varphi\beta & \varphi\gamma \end{bmatrix}^j \left\{ \begin{bmatrix} \omega \\ \xi + \varphi\omega \end{bmatrix} + \begin{bmatrix} 0 \\ \tau(z_{t+h-j}) + u_{t+h-j} \end{bmatrix} \right\}.$$

This is easily extended to the general case ($p, q \geq 1$) where we have

$$Y_t = AY_{t-1} + b + \epsilon_t,$$

with the conventions

$$Y_t = \begin{bmatrix} \tilde{h}_t \\ \vdots \\ \tilde{h}_{t-p+1} \\ \tilde{x}_t \\ \vdots \\ \tilde{x}_{t-q+1} \end{bmatrix}, \quad A = \begin{pmatrix} (\beta_1, \ldots, \beta_p) & (\gamma_1, \ldots, \gamma_q) \\ (I_{p-1\times p-1}, 0_{p-1\times 1}) & 0_{p-1\times q} \\ \varphi(\beta_1, \ldots, \beta_p) & \varphi(\gamma_1, \ldots, \gamma_q) \\ 0_{q-1\times p} & (I_{q-1\times q-1}, 0_{q-1\times 1}) \end{pmatrix},$$

$$b = \begin{pmatrix} \omega \\ 0_{p-1\times 1} \\ \xi + \varphi\omega \\ 0_{q-1\times 1} \end{pmatrix}, \quad \epsilon_t = \begin{bmatrix} 0_{p\times 1} \\ \tau(z_t) + u_t \\ 0_{q\times 1} \end{bmatrix},$$

so that

$$Y_{t+h} \quad = \quad A^h Y_t + \sum_{j=0}^{h-1} A^j (b + \epsilon_{t+h-j}).$$

The predictive distribution for $\tilde{h}_{t+h}$ and/or $\tilde{x}_{t+h}$, is given from the distribution of $\sum_{i=0}^{h-1} A^i \epsilon_{t+h-i}$, which also enable us to compute a predictive distribution for r_{t+h}, and cumulative returns $r_{t+1} + \cdots + r_{t+h}$.

Induced properties of cumulative returns The skewness for single period returns is non-zero, if and only if the studentized return, z_t, has non-zero skewness. This follows directly from the identity $r_t = \sqrt{h_t} z_t$, and the assumption that $z_t \perp\!\!\!\perp h_t$, that shows that,

$$\mathrm{E}(r_t^d) = \mathrm{E}(h_t^{d/2} z_t^d) = \mathrm{E}\left\{\mathrm{E}(h_t^{d/2} z_t^d | \mathcal{F}_{t-1})\right\} = \mathrm{E}(h_t^{d/2})\mathrm{E}(z_t^d),$$

and in particular that $\mathrm{E}(r_t^3) = \mathrm{E}(h_t^{3/2})\mathrm{E}(z_t^3)$. So a symmetric distribution for z_t implies that r_t has zero skewness, and this is property that is shared by standard GARCH model and *Realized GARCH* model alike.

For the skewness and kurtosis of cumulative returns, $r_t + \cdots + r_{t+k}$, the situation is very different, because the leverage function induces skewness.

Proposition 26. *Consider RealGARCH(1,1) model and define $\pi = \beta + \varphi\gamma$ and $\mu = \omega + \varphi\xi$, so that*

$$\log h_t = \pi \log h_{t-1} + \mu + \gamma w_{t-1}, \qquad \textit{where} \qquad w_t = \tau_1 z_t + \tau_2(z_t^2 - 1) + u_t,$$

with $z_t \sim \mathrm{iid}\,\mathcal{N}(0,1)$ and $u_t \sim \mathrm{iid}\,\mathcal{N}(0, \sigma_u^2)$. The kurtosis of the return $r_t = \sqrt{h_t} z_t$ is given by

$$\frac{\mathrm{E}(r_t^4)}{\mathrm{E}(r_t^2)^2} = 3 \left(\prod_{i=0}^{\infty} \frac{1 - 2\pi^i\gamma\tau_2}{\sqrt{1 - 4\pi^i\gamma\tau_2}} \right) \exp\left\{ \sum_{i=0}^{\infty} \frac{\pi^{2i}\gamma^2\tau_1^2}{1 - 6\pi^i\gamma\tau_2 + 8\pi^{2i}\gamma^2\tau_2^2} \right\} \exp\left\{ \frac{\gamma^2\sigma_u^2}{1 - \pi^2} \right\}. \tag{2.27}$$

There does not appear to be a way to further simplify the expression (2.27), however when $\gamma\tau_2$ is small, as we found it to be empirically, we have the approximation (see the appendix for details)

$$\frac{\mathrm{E}(r_t^4)}{\mathrm{E}(r_t^2)^2} \simeq 3 \exp\left\{ \frac{\gamma^2\tau_2^2}{-\log\pi} + \frac{\gamma^2(\tau_1^2 + \sigma_u^2)}{1 - \pi^2} \right\}. \tag{2.28}$$

Simple extension to multivariate The simplest extension to multivariate case is via the DCC(1,1)-GARCH framework, by replacing the conditional variance equation (2.19) by

$$D_t^2 = \mathrm{diag}\,(\omega_i) + \mathrm{diag}\,(\beta_i) \otimes D_{t-1}^2 + \mathrm{diag}\,(\gamma_i) \otimes X_{t-1}$$

where the diagonal matrix D_t^2 consists of $\log h_{1,t}, \ldots, \log h_{N,t}$, and X_t of $\log x_{1,t}, \ldots, \log x_{N,t}$, and is related to D_t^2 by the usual measurement equation (cf. (2.24))

$$\log x_{i,t} = \xi_i + \varphi_i \log h_{i,t} + \tau(z_{i,t}) + u_{i,t}.$$

Essentially, the proposed extension replaces the first stage of DCC-GARCH by RealGarch(1,1), the output of which is then fed into the second stage to estimate the conditional correlation given by

$$\begin{aligned} R_t &= \operatorname{diag}(Q_t)^{-1/2} Q_t \operatorname{diag}(Q_t)^{-1/2} \\ Q_t &= \left(1 - \widetilde{\alpha} - \widetilde{\beta}\right) \overline{Q} + \widetilde{\alpha} \epsilon_{t-1} \epsilon_{t-1}^{\top} + \widetilde{\beta} Q_{t-1}, \end{aligned}$$

where $\overline{Q} = E\left[\epsilon_{t-1}\epsilon_{t-1}^{\top}\right]$ is the unconditional correlation matrix of the standardized residual. The key assumption for this extension is that we conjecture that correlation varies at a much slower time scale than does the volatility for each asset. This enables us to simply plug-in the variance estimations based on *Realized GARCH*, normalize our return series, than estimate the correlation component exactly as before.

2.2.2 Empirical Analysis

Data Our sample spans the period from 2002-01-01 to 2008-08-31, which we divide into an in-sample period: 2002-01-01 to 2007-12-31; leaving the eight months, 2008-01-02 and 2008-08-31, for out-of-sample analysis. We adopt the realized kernel as the realized measure, x_t, using a Parzen kernel function. This estimator is similar to the well known realized variance, but is robust to market microstructure noise and is a more accurate estimator of the quadratic variation. Our implementation of the realized kernel follows Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a) that guarantees a positive estimate, which is important for our log-linear specification. The exact computation is explained in great details in Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009). When we estimate a *Realized GARCH* model using open-to-close returns we should expect $x_t \approx h_t$, whereas with close-to-close returns we should expect x_t to be smaller than h_t on average.

To avoid outliers that would result from half trading days, we removed days where high-frequency data spanned less than 90% of the official 6.5 hours between 9:30am and 4:00pm. This removes about three daily observation per year, such as the day after Thanksgiving and days around Christmas. When we estimate a model that involves $\log r_t^2$, we deal with zero returns by substituting $\min_{\{s<t:r_s^2>0\}} \log r_t^2$ for $\log 0$.

Result In this section we present empirical results using returns and realized measures for 28 stocks and and exchange-traded index fund, SPY, that tracks the S&P 500 index. We adopt the realized kernel, introduced by Barndorff-Nielsen et al. (2008b), as the realized measure, x_t. We estimate the realized GARCH models using both open-to-close returns and close-to-close returns. High-frequency prices are only available between "open" and "close", so the population quantity that is estimated by the realized kernel is directly related to the volatility of open-to-close returns, but only captures some of the volatility of close-to-close returns.

Next we consider *Realized GARCH* models with a log-linear specification of the GARCH and measurement equations. For the purpose of comparison we estimate an LGARCH(1,1) model in

addition to the six *Realized GARCH* models. Again we report results for both open-to-close returns and close-to-close returns for SPY. The results are presented in Table 2.

The main results in Table 2 can be summarized by:

▷ ARCH parameter, α, is insignificant.

▷ Consistent with the log-linearity we find $\varphi \simeq 1$ for both open-to-close and close-to-close returns.

We report empirical results for all 29 assets in Table 3 and find the point estimates to be remarkable similar across the many time series. In-sample and out-of-sample likelihood ratio statistics are computed in Table 4.

Table 4 shows the likelihood ratios for both in-sample and out-of-sample period. The statistics are based on our analysis with open-to-close returns. The statistics in Panel A are the conventional likelihood ratio statistics, where each of the five smaller models are benchmarked against the largest model. The largest model is labeled (2,2). This is the log-linear RealGARCH(2,2) model that has the squared return r_t^2 in the GARCH equation in addition to the realized measure. Thus the likelihood ratio statistics are in Panel A are defined by

$$LR_i = 2\left\{\ell_{RG(2,2)^*}(r,x) - \ell_i(r,x)\right\},$$

where i represents one of the five other *Realized GARCH* models. In the QMLE framework the limit distribution of likelihood ratio statistic, LR_i, is usually given as a weighted sum of χ^2-distributions. Thus comparing the LR_i to the usual critical value of a χ^2-distribution is only indicative of significance.

Comparing the RealGARCH(2,2)* to RealGARCH(2,2) leads to small LR statistics in most cases. So α tends to be insignificant in our sample. This is consistent with the existing literature that finds that squared returns adds little to the model, once a more accurate realized measure is used in the GARCH equation.

The leverage function, $\tau(z_t)$, is highly significant in all cases. The LR statistics associated with the hypothesis that $\tau_1 = \tau_2 = 0$ are well over 100 in all cases. These statistics can be computed by subtracting the statistics in the column labeled (2,2) from those in the column labeled (2,2)†. The joint hypothesis, $\beta_2 = \gamma_2 = 0$ is rejected in most cases, and so the empirical evidence does not support a simplification of the model to the RealGARCH(1,1). The results for the two hypotheses $\beta_2 = 0$ and $\gamma_2 = 0$ are less conclusive. The likelihood ratio statistics for the hypothesis, $\beta_2 = 0$ are, on average, 5.7, which would be borderline significant when compared to conventional critical values from a $\chi^2_{(1)}$-distribution. The LR statistics for the hypothesis, $\gamma_2 = 0$, tend to be larger and are on average 16.6. So the empirical evidence favors the RealGARCH(1,2) model over the RealGARCH(2,1) model.

Consider next the out-of-sample statistics in Panel B. These *likelihood ratio* (LR) statistics are computed as

$$\sqrt{\frac{n}{m}}\{\ell_{RG(2,2)}(r,x) - \ell_j(r,x)\},$$

where n and m denote the sample sizes, in-sample and out-of-sample, respectively. The in-sample parameter estimates are simply plugged into the out-of-sample log-likelihood, and the asymptotic distribution of these statistics are non-standard because the in-sample estimates do not solve the first-order conditions out-of-sample, see Hansen (2009). The RealGARCH(2,2) model nests, or is

	Open-to-Close Returns							Close-to-Close Returns						
Model	G(1,1)	RG(1,1)	RG(1,2)	RG(2,1)	RG(2,2)	RG(2,2)†	RG(2,2)*	G(1,1)	RG(1,1)	RG(1,2)	RG(2,1)	RG(2,2)	RG(2,2)†	RG(2,2)*
						Panel A: Point Estimates and Log-Likelihood (in-sample)								
ω	0.04 (0.00)	0.06	0.04	0.06	0.00	0.00	0.00	0.05	0.18	0.11	0.19	0.01	0.01	0.02
α	0.03	-	-	-	-	-	0.00	0.03	-	-	-	-	-	0.00
β_1	0.96	0.55	0.70	0.40	1.43	1.42	1.45	0.96	0.54	0.72	0.37	1.38	1.40	1.35
β_2		-	-	0.13	-0.44	-0.44	-0.46		-	-	0.15	-0.41	-0.43	-0.39
γ_1		0.41	0.45	0.43	0.46	0.40	0.42		0.43	0.48	0.46	0.45	0.42	0.46
γ_2		-	-0.18	-	-0.44	-0.38	-0.41		-	-0.21	-	-0.42	-0.40	-0.42
ξ		-0.18	-0.18	-0.18	-0.23	-0.16	-0.18		-0.42	-0.42	-0.42	-0.42	-0.41	-0.42
φ		1.04	1.04	1.04	0.96	1.07	1.03		0.99	1.00	0.99	1.02	1.03	0.99
σ_u		0.38	0.38	0.38	0.38	0.41	0.38		0.39	0.38	0.39	0.38	0.41	0.38
τ_1		-0.07	-0.07	-0.07	-0.07	-	-0.07		-0.11	-0.11	-0.11	-0.11	-	-0.11
τ_2		0.07	0.07	0.07	0.07	-	0.07		0.04	0.04	0.04	0.04	-	0.04
$\ell(r,x)$	-	-2395.6	-2388.8	-2391.9	-2385.1	-2495.7	-2382.9		-2576.86	-2567.15	-2571.67	-2564.16	-2661.73	-2563.53
						Panel B: Auxiliary Statistics (in-sample)								
π	.988	.975	.986	.976	.999	.999	.999	.988	.974	.987	.975	.999	.999	.999
ρ		-0.18	-0.18	-0.16	-0.19	-	-0.16		-0.27	-0.25	-0.25	-0.25	-	-0.28
ρ^-		-0.33	-0.32	-0.32	-0.35	-	-0.35		-0.31	-0.29	-0.28	-0.29	-	-0.33
ρ^+		0.12	0.12	0.13	0.13	-	0.14		-0.01	-0.03	-0.03	0.01	-	-0.05
$\ell(r)$	-1752.7	-1712.0	-1710.3	-1711.4	-1712.3	-1708.9	-1709.6	-1938.24	-1876.51	-1875.46	-1876.12	-1875.39	-1874.88	-1876.13

Table 2: Results for the logarithmic specification: G(1,1) denotes the LGARCH(1,1) model that does not utilize a realized measure of volatility. RG(2,2)† denotes the RealGARCH(2,2) model without the $\tau(z)$ function that captures the dependence between returns and innovations in volatility. RG(2,2)* is the (2,2) extended to include the ARCH-term $\alpha \log r_{t-1}^2$. The latter being insignificant.

	ω	β	γ_1	γ_2	ξ	φ	σ_u	τ_1	τ_2	$\ell(r)$	$\ell(r,x)$	π	ρ	ρ^-	ρ^+	$\ell(r)^{os}$	$\ell(r,x)^{os}$
AA	0.03	0.77	0.33	-0.14	-0.07	1.15	0.40	-0.04	0.09	-2776.4	-3519.9	0.98	-0.08	-0.32	0.24	-409.1	-493.8
AIG	0.02	0.74	0.45	-0.21	-0.06	1.02	0.45	-0.02	0.04	-2403.1	-3317.2	0.98	-0.06	-0.17	0.08	-438.0	-535.8
AXP	0.05	0.70	0.38	-0.12	-0.16	1.08	0.43	-0.02	0.10	-2371.1	-3217.9	0.99	-0.05	-0.30	0.25	-390.8	-477.8
BA	0.02	0.82	0.31	-0.17	-0.13	1.22	0.39	-0.03	0.09	-2536.0	-3260.0	0.99	-0.09	-0.36	0.26	-335.8	-440.2
BAC	0.00	0.78	0.51	-0.29	0.00	0.99	0.42	-0.04	0.08	-2016.9	-2823.4	0.99	-0.09	-0.31	0.21	-416.9	-517.4
C	-0.02	0.74	0.45	-0.19	0.09	0.99	0.39	-0.03	0.09	-2260.5	-2974.0	0.99	-0.07	-0.31	0.24	-427.6	-516.0
CAT	0.03	0.82	0.37	-0.22	-0.14	1.07	0.38	-0.03	0.09	-2621.1	-3279.4	0.99	-0.08	-0.32	0.27	-315.6	-394.0
CVX	0.03	0.71	0.33	-0.14	-0.09	1.32	0.39	-0.08	0.08	-2319.1	-3021.8	0.97	-0.19	-0.35	0.14	-306.1	-377.7
DD	-0.01	0.77	0.37	-0.17	0.08	1.08	0.40	-0.05	0.08	-2301.2	-3067.3	0.98	-0.13	-0.35	0.20	-307.9	-394.1
DIS	0.01	0.85	0.39	-0.25	-0.05	1.10	0.41	-0.04	0.09	-2518.5	-3289.6	1.00	-0.09	-0.35	0.22	-288.7	-364.8
GE	0.00	0.81	0.38	-0.19	0.01	0.98	0.41	-0.01	0.08	-2197.8	-2988.7	0.99	-0.02	-0.26	0.25	-300.8	-401.7
GM	0.06	0.84	0.39	-0.24	-0.32	1.02	0.47	-0.01	0.12	-2987.9	-3967.3	0.99	-0.01	-0.33	0.31	-384.7	-499.5
HD	0.01	0.79	0.39	-0.20	0.00	1.01	0.41	-0.05	0.09	-2538.4	-3318.4	0.99	-0.13	-0.37	0.20	-378.2	-452.0
IBM	0.00	0.74	0.41	-0.15	0.01	0.94	0.39	-0.04	0.08	-2192.6	-2896.7	0.98	-0.09	-0.32	0.24	-284.7	-360.8
INTC	0.02	0.87	0.46	-0.33	-0.11	1.03	0.36	-0.02	0.07	-2869.1	-3481.1	1.00	-0.05	-0.24	0.22	-345.2	-424.4
JNJ	-0.03	0.80	0.38	-0.19	0.13	1.04	0.44	0.02	0.10	-1874.8	-2777.3	0.99	0.04	-0.25	0.30	-192.9	-277.6
JPM	0.01	0.81	0.49	-0.30	-0.02	0.98	0.42	-0.04	0.09	-2463.0	-3276.8	0.99	-0.10	-0.30	0.22	-402.7	-496.7
KO	-0.05	0.76	0.45	-0.21	0.19	0.93	0.38	-0.02	0.07	-1886.7	-2573.6	0.99	-0.06	-0.28	0.19	-265.0	-360.3
MCD	0.00	0.88	0.37	-0.25	-0.01	0.98	0.45	-0.05	0.11	-2461.8	-3371.9	0.99	-0.09	-0.35	0.26	-290.7	-386.3
MMM	0.00	0.77	0.43	-0.23	0.02	0.98	0.41	-0.02	0.07	-2140.3	-2944.8	0.97	-0.04	-0.23	0.21	-261.7	-331.4
MRK	0.03	0.84	0.33	-0.21	-0.19	1.23	0.47	0.01	0.07	-2479.2	-3478.5	0.98	0.04	-0.13	0.18	-314.0	-437.1
MSFT	-0.01	0.79	0.44	-0.22	0.08	0.92	0.38	-0.03	0.08	-2330.7	-3021.1	0.99	-0.08	-0.31	0.24	-317.7	-402.9
PG	-0.04	0.78	0.43	-0.25	0.18	1.04	0.41	-0.05	0.08	-1850.7	-2646.6	0.98	-0.14	-0.32	0.14	-239.8	-315.2
T	0.00	0.86	0.53	-0.38	0.01	0.86	0.46	-0.03	0.10	-2560.4	-3512.7	0.99	-0.07	-0.32	0.25	-313.1	-397.3
UTX	-0.01	0.80	0.45	-0.24	0.06	0.88	0.40	-0.01	0.10	-2302.4	-3059.2	0.99	-0.05	-0.34	0.29	-289.5	-372.0
VZ	-0.01	0.79	0.40	-0.20	0.07	1.01	0.43	-0.03	0.09	-2343.4	-3196.7	0.99	-0.08	-0.31	0.23	-320.8	-404.9
WMT	-0.02	0.80	0.37	-0.19	0.12	1.04	0.39	-0.01	0.09	-2164.9	-2893.6	0.99	-0.02	-0.29	0.30	-275.9	-353.7
XOM	0.03	0.71	0.34	-0.12	-0.10	1.26	0.38	-0.08	0.08	-2334.7	-2994.1	0.98	-0.20	-0.37	0.15	-299.0	-364.9
SPY	0.04	0.70	0.45	-0.18	-0.18	1.04	0.38	-0.07	0.07	-1710.3	-2388.8	0.99	-0.17	-0.32	0.13	-260.0	-338.4
Average	0.01	0.79	0.41	-0.21	-0.02	1.04	0.41	-0.03	0.09	-	-	0.99	-0.08	-0.30	0.22	-	-

Table 3: Estimates for the RealGARCH$(1,2)$ model.

nested in, all other models. For nested and correctly specified models where the larger model has k additional parameters that are all zero (under the null hypothesis) the out-of-sample likelihood ratio statistic is asymptotically distributed as

$$\sqrt{\frac{n}{m}}\left\{\ell_i(r,x)-\ell_j(r,x)\right\} \xrightarrow{d} Z_1'Z_2, \qquad \text{as} \quad m,n\to\infty \quad \text{with} \quad m/n\to 0,$$

where Z_1 and Z_2 are independent $Z_i \sim \mathcal{N}_k(0,I)$. This follows from, for instance, Hansen (2009, corollary 2), and the (two-sided) critical values can be inferred from the distribution of $|Z_1'Z_2|$. For $k=1$ the 5% and 1% critical values are 2.25 and 3.67, respectively, and for two degrees of freedom ($k=2$), these are 3.05 and 4.83, respectively. When compared to these critical values we find, on average, significant evidence in favor of a model with more lags than RealGARCH(1,1). The statistical evidence in favor of a leverage function is very strong. Adding the ARCH parameter, α, will (on average) result in a worse out-of-sample log-likelihood. As for the choice between the RealGARCH(1,2), RealGARCH(2,1), and RealGARCH(2,2) the evidence is mixed.

In Panel C, we report partial likelihood ratio statistics, that are defined by $2\{\max_i \ell_i(r|x) - \ell_j(r|x)\}$, so each model is compared with the model that had the best out-of-sample fit in term of the partial likelihood. These statistics facilitate a comparison of the *Realized GARCH* models with the standard GARCH(1,1) model, and we see that the *Realized GARCH* models also dominate the standard GARCH model in this metric. This is made more impressive by the fact that *Realized GARCH* models are maximizing the joint likelihood, and not the partial likelihood that is used in these comparisons.[8]

The leverage function, $\tau(z)$ is closely related to the *news impact curve* that was introduced by Engle and Ng (1993). High frequency data enable are more more detailed study of the news impact curve than is possible with daily returns. A detailed study of the news impact curve that utilizes high frequency data is Ghysels and Chen (2010). Their approach is very different from ours, yet the shape of the news impact curve they estimate is very similar to ours. The news impact curve shows how volatility is impacted by a shock to the price, and our Hermite specification for the leverage function presents a very flexible framework for estimating the news impact curve. In the log-linear specification we define the new impact curve by

$$\nu(z) = \mathrm{E}(\log h_{t+1}|z_t = z) - \mathrm{E}(\log h_{t+1}),$$

so that $100\nu(z)$ measures the percentage impact on volatility as a function of return-shock measures in units of standard deviations. As shown in Section 2.2.1 we have $\nu(z) = \gamma_1\tau(z)$. We have estimated the log-linear RealGARCH(1,2) model for both IBM and SPY using a flexible leverage function based on the first four Hermite polynomials. The point estimates were $(\hat\tau_1,\hat\tau_2,\hat\tau_3,\hat\tau_4) = (-0.036, 0.090, 0.001, -0.003)$ for IBM and $(\hat\tau_1,\hat\tau_2,\hat\tau_3,\hat\tau_4) = (-0.068, 0.081, 0.014, 0.002)$ for SPY. Note that the Hermite polynomials of orders three and four add little beyond the first two polynomials. The news impact curves implied by these estimates are presented in Figure 2.3. The fact that $\nu(z)$ is smaller than zero for some (small) values of z is an implication of its definition that implies, $\mathrm{E}[\nu(z)] = 0$.

[8] There is not a well developed theory for the asymptotic distribution of these statistics, in part because we are comparing a model that maximizes the partial likelihood (the GARCH(1,1) model) with models that maximizes the joint likelihood (the *Realized GARCH* models).

	Panel A: In-Sample Likelihood Ratio						Panel B: Out-of-Sample Likelihood Ratio						Panel C: Out-of-Sample Partial Likelihood Ratio						
	(1,1)	(1,2)	(2,1)	(2,2)	$(2,2)^{\dagger}$	$(2,2)^{*}$	(1,1)	(1,2)	(2,1)	(2,2)	$(2,2)^{\dagger}$	$(2,2)^{*}$	G11	(1,1)	(1,2)	(2,1)	(2,2)	$(2,2)^{\dagger}$	$(2,2)^{*}$
AA	24.0	8.3	11.7	3.9	187.4	0	6.9	4.5	6.4	0	21.9	0.1	4.6	3.3	1.4	2.6	0.0	0.6	0.7
AIG	43.8	15.2	20.6	14.2	201.2	0	15.7	-0.2	6.0	0	25.5	12.5	56.1	9.3	5.9	7.4	6.0	0.0	7.2
AXP	30.5	25.0	26.0	0.0	197.0	0	1.2	2.7	1.5	0	13.3	0.2	24.0	0.0	0.3	0.1	1.3	1.7	1.4
BA	34.8	6.0	18.8	0.0	197.1	0	-1.4	0.4	-2.6	0	24.5	0.0	1.1	0.6	1.8	1.7	0.0	0.3	0.0
BAC	46.3	5.9	20.5	5.1	198.8	0	8.0	-0.7	-0.3	0	56.5	-2.8	147.9	4.1	0.6	0.0	1.5	19.7	0.9
C	26.8	9.5	16.0	6.3	228.4	0	1.3	-2.7	-2.5	0	-0.1	-7.2	26.9	0.3	0.9	0.5	1.1	0.0	0.4
CAT	39.8	2.3	13.8	1.6	227.3	0	1.7	-1.0	-4.2	0	9.0	2.2	47.3	0.1	0.8	0.0	1.0	1.6	1.5
CVX	17.7	0.1	3.5	0.0	194.6	0	5.3	0.0	2.2	0	20.8	-0.1	30.1	0.6	0.3	0.2	0.3	0.0	0.3
DD	31.1	10.8	15.9	6.1	167.0	0	4.6	4.1	3.0	0	-7.1	6.9	19.2	1.2	1.5	1.5	1.2	1.4	0.0
DIS	57.7	12.6	33.1	0.0	213.7	0	4.8	4.8	3.7	0	24.3	0.0	35.4	0.0	2.0	1.4	0.8	1.1	0.8
GE	37.7	12.6	20.2	12.2	200.3	0	14.1	0.6	5.9	0	9.3	0.6	41.6	0.9	0.0	0.6	0.0	0.3	0.7
GM	62.7	18.8	39.9	18.4	319.9	0	17.0	-0.2	6.8	0	7.2	14.6	57.5	1.4	0.0	0.2	0.0	0.3	0.9
HD	29.8	4.0	14.4	2.0	201.3	0	2.2	-1.0	-1.0	0	28.9	0.5	45.5	0.9	0.0	0.1	0.2	2.1	0.4
IBM	28.5	16.2	20.2	1.0	176.7	0	3.3	-0.1	1.0	0	25.0	-0.2	12.0	0.4	0.5	0.7	0.1	0.3	0.0
INTC	75.9	13.9	47.8	9.7	130.8	0	0.9	0.1	-3.4	0	36.8	19.1	133.1	1.7	0.5	2.0	0.0	0.1	1.8
JNJ	34.6	1.3	4.8	0.1	235.4	0	7.1	-0.8	2.0	0	30.7	0.2	21.8	0.4	0.0	0.1	0.1	0.1	0.0
JPM	50.7	3.5	23.6	1.9	213.5	0	0.3	-2.5	-5.4	0	-3.6	-0.6	34.9	0.0	1.3	0.1	1.6	2.0	1.6
KO	39.7	22.0	28.3	0.0	186.1	0	2.1	-2.8	-1.0	0	22.6	0.0	5.6	0.4	0.6	0.0	0.5	0.9	0.5
MCD	54.8	2.9	26.1	0.8	278.6	0	-13.0	-0.6	-9.3	0	47.4	1.4	6.8	0.0	1.5	0.2	1.8	2.0	1.7
MMM	32.0	6.7	20.4	0.1	183.5	0	-2.4	-0.2	-2.8	0	24.5	-0.1	14.5	0.0	1.9	0.9	1.4	1.0	1.4
MRK	64.6	12.0	10.6	2.9	309.0	0	-3.4	-2.5	-1.2	0	32.4	0.4	11.8	0.0	1.6	1.7	1.9	2.4	1.8
MSFT	37.3	11.5	20.5	10.2	186.1	0	8.3	4.2	7.1	0	17.0	4.1	37.1	0.1	0.5	0.4	0.0	0.2	0.1
PG	36.5	3.0	12.3	2.9	160.4	0	-1.3	0.3	-0.4	0	35.0	1.0	24.8	0.0	1.4	0.6	1.3	1.5	1.7
T	69.8	4.8	39.0	0.0	198.3	0	19.7	-0.3	13.5	0	35.1	-0.2	19.2	2.8	0.2	2.6	0.0	1.7	0.0
UTX	39.6	21.6	29.0	0.0	223.7	0	5.8	-2.0	1.2	0	33.1	-0.4	12.3	1.4	0.2	0.9	0.2	0.0	0.1
VZ	31.5	4.3	13.2	0.5	188.7	0	15.0	3.2	11.1	0	16.6	-1.3	8.7	3.5	0.7	2.8	0.3	0.0	0.1
WMT	36.2	12.0	23.0	8.3	190.0	0	-7.2	-3.8	-7.9	0	28.7	9.4	30.5	0.0	0.6	0.0	1.3	1.4	2.6
XOM	14.7	1.1	4.3	0.9	234.0	0	5.7	-0.1	1.0	0	27.9	0.6	21.6	0.0	0.5	0.3	0.5	0.5	0.5
SPY	25.3	11.6	17.9	4.2	225.6	0	6.3	-1.2	1.9	0	24.4	1.4	40.8	0.8	0.6	0.7	0.0	2.5	1.3
Average	39.8	9.6	20.5	3.9	208.8	0	4.4	0.1	1.1	0	23.0	2.1	33.5	1.2	1.0	1.0	0.8	1.6	1.0

Table 4: In-Sample and Out-of-Sample Likelihood Ratio Statistics

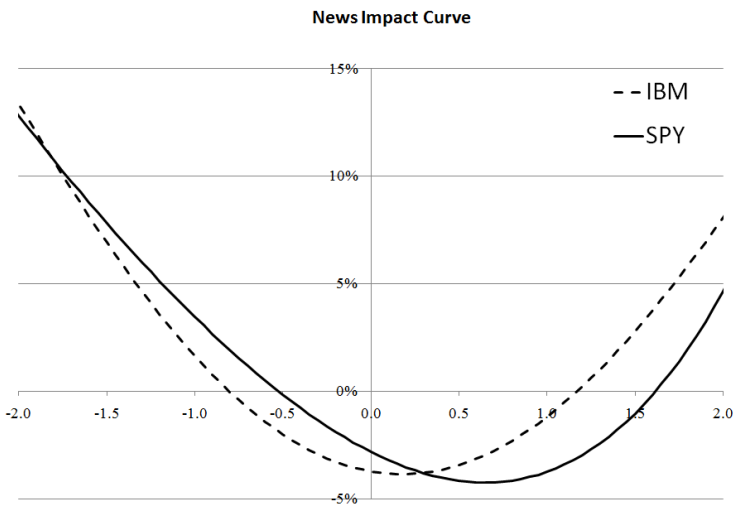

Figure 2.3: News impact curve for IBM and SPY

The estimated news impact curve for IBM is more symmetric about zero than that of SPY, and this empirical result is fully consistent with the existing literature. The most common approach to model the news impact curve is to adopt a specification with a discontinuity at zero, such as that used in the EGARCH model by Nelson (1991), $\tau(z) = \tau_1 z + \tau_+(|z| - \mathrm{E}|z|)$. We also estimated the leverage functions with the piecewise linear function that leads to similar empirical results. Specifically, the implied news impact curves have the most pronounced asymmetry for the index fund, SPY, and the two oil related stocks, CVX and XOM. However, the likelihood function tends to be larger with the polynomial leverage function, $\tau(z) = \tau_1 z + \tau_2(z^2 - 1)$, and the polynomial specification simplifies aspects of the likelihood analysis.

For the multivariate case, an application of the proposed DCC(1,1)-RealGARCH(1,1) extension to the univariate *Realized GARCH* model is illustrated by using four stocks, IBM, XOM, SPY and WMT, with fitted result shown in Table 5. Note that this simple extension is a proof of concept of how the RealGARCH framework can be extended to deal with covariance estimation, as a straightforward plug-in to an existing establish framework. The assumption that volatility changes more frequently than the underlying correlation of stocks is justifiable for typical portfolio holding horizon measured in days or longer. Here we can draw on the result of empirical analysis for the univariate case and expect similar predictive gain for this proposed extension. More sophisticated frameworks would be beneficial when we need to deal with cases where both volatility and correlation are highly non-stationary, and change on a comparable time scale. This is an area for future research.

	IBM	XOM	SPY	WMT
h_0	0.76	1.17	0.28	2.39
ω	-0.00	0.04	0.06	-0.03
β	0.64	0.59	0.55	0.66
γ	0.36	0.30	0.41	0.30
ξ	0.01	-0.10	-0.18	0.12
φ	0.94	1.27	1.04	1.04
τ_1	-0.04	-0.08	-0.07	-0.01
τ_2	0.08	0.08	0.07	0.09
σ_u	0.39	0.38	0.38	0.40

$\widetilde{\alpha}$	$\widetilde{\beta}$
0.018	0.952

Table 5: Fitted result for DCC(1,1)-RGARCH(1,1)

3 Asset Return Estimation and Prediction

In the mean-variance portfolio optimization framework, *mean* often refers to some prediction of future asset return. Essentially, the framework requires a forecast for the first two moments of the joint distribution of an universe of assets under consideration. Of course, in a jointly Gaussian setting, these are the only two moments needed for deriving some objective value based on a suitably chosen utility function. As it is now commonly acknowledged, asset returns are far some Gaussian and often exhibit characteristic tail and asymmetric behavior that are distinctly different to a simple Gaussian distribution. Whether it is essential to take these added complexity into account depends on the problem at hand and on the reliability of whether these stylized facts can be adequately estimated, and important signals extracted, from the underlying inherently noisy observations.

Given the significant amount of improvement in profitability from being able to increase just a small fraction of the investor's edge over a random estimation, a lot of effort have been spent on modeling this first moment of the asset return dynamics.

3.1 Outline of Some Commonly Used Temporal Models

Random Walk If we assume that the underlying return process is a symmetric random walk, then the one step ahead prediction is simply equal to the current return. That is,

$$r_{t+1} | \mathcal{F}_t = r_t + \epsilon_{t+1},$$

where $r_t \in \mathbb{R}^N$ is the vector of returns often defined as the difference in the logarithm of prices, and ϵ_t is a $\mathcal{F}_t$-adapted, zero-mean, *iid* random variable. This is not a model per se, but a benchmark that are often used to compare the predictive gain of different modeling frameworks.

Auto-Regressive-Moving-Average (ARMA) models ARMA model (see for example Box and Jenkins (1976)) combines the ideas of *auto-regressive* (AR) and *moving average* (MA) models into a compact and more parsimonious form so that the number of parameters used is kept small. Note that GARCH, mentioned in Section (2), can be regarded as an ARMA model. There are a number of variations of this versatile framework.

Example 27. Simple ARMA

The classic ARMA(p,q) framework is given by

$$\begin{aligned} \Phi(B) r_t &= \Theta(B) \epsilon_t \\ \epsilon &\sim \mathcal{N}(0,1), \end{aligned}$$

where $\Phi(B) = 1 - \sum_{i=0}^{p} \phi_i B^i$, $\Theta(B) = 1 - \sum_{i=0}^{q} \theta_i B^i$, B is the lag operator, and $\{\epsilon_t\}$ is *an iid* Gaussian innovation series. The series, r_t, is said to be *unit-root stationary* if the zeros of $\Phi(z)$ are outside the unit circle. For non-stationary series, appropriate differencing needs to be applied first before we can estimate the parameters of the model (Box and Jenkins, 1976). The lag order (i.e. the value for p and q) selection can be done iteratively via the *Akaike Information Criterion* (AIC) value of the fitted result. Recall that

$$AIC = 2k - 2 \log L,$$

where k is the number of parameters in the model and L is the likelihood function. The model fitting algorithm iteratively increase the lag order and select the final complexity with the lowest AIC.

Example 28. ARMA with Wavelet Smoothing

Market microstructure noises, such as bid-ask bounce and price quantization, usually have the maximum impact for sampling frequency in the region of 5 to 20 minutes (Andersen et al., 2001c). To effectively filter out these noises, we can first decompose the raw return series, r_t, into different resolutions, or more precisely different levels of wavelet details, by an application of wavelet *multi-resolution analysis* (MRA) (Mallat, 1989), followed by reconstitution of the time series signal by using a subset of wavelet details so that details below a threshold are filtered out. Recall by definition of *discrete wavelet transform* (DWT)

$$W = \mathcal{W} r_t,$$

where $W \in \mathbb{R}^{N \times N}$ is the orthogonal real-valued matrix defining the DWT. Consider the decomposition of the vector W into $J+1$ sub-vectors, so that $W = (W_1, \ldots, W_J, V_J)^\top$, where W_j is a column vector with $N/2^j$ elements[9] and $2^J = N$. So we have

$$r_t = \mathcal{W}^\top W = \sum_{n=0}^{N-1} W_n \mathcal{W}_{n\bullet} = \sum_{j=1}^{J} \mathcal{W}_j^\top W_j + \mathcal{V}_J^\top V_J := \sum_{j=1}^{J} \mathcal{D}_j + S_J, \tag{3.1}$$

which defines a MRA of our return series r_t, consists of a constant vector S_J and J levels of wavelet detail $\mathcal{D}_j$, $j = 1, \ldots, J$, each of which contains a time series related to variations in r_t at a certain scale. $\mathcal{W}_{n\bullet}$ is the n-th row of the DWT matrix. $\mathcal{W}_j$ and $\mathcal{V}_j$ matrices are defined by partitioning the rows of $\mathcal{W}$ commensurate with the partitioning of W into $W_1, \ldots, W_J$ and V_J. By discarding the first $K-1$ levels of detail, we obtain a series that is a locally filtered, smoothed series, $\tilde{r}_t = \sum_{j=K}^{J} \mathcal{D}_j + S_J$, which retains features with time scales larger than a predetermined threshold. Finally, a simple ARMA model is fitted to this filtered series,

$$\begin{aligned} \Phi(B)\,\tilde{r}_t &= \Phi(B)\,\epsilon_t \\ \epsilon &\sim \mathcal{N}(0,1), \end{aligned}$$

where again, AIC can be used to select the optimal lag parameters.

Example 29. ARMA-GARCH

To account for heteroskedasticity of the return series, we can jointly fit an ARMA and GARCH model, for example an ARMA(1,1)-GARCH(1,1) framework on the raw return series, r_t, so that

$$\begin{aligned} \Phi(B)\,r_t &= \Theta(B)\,\epsilon_t \\ h_t &= \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 h_{t-1} \\ \epsilon_t &\sim \mathcal{N}(0, h_t). \end{aligned}$$

[9] For cases where N is not an integral power of 2, we can simply truncate and discard some part of the data series, or, refer to Mallat (1989) if it is necessary to treat it as a special case.

As in Example (27), the algorithm first uses AIC to determine the optimal lag parameters (p*,q*) for the ARMA framework then, holding the parameter values constant, it fits an ARMA(p*,q*)-GARCH(1,1) model by maximizing the likelihood a Gaussian density for the innovation, $\{\epsilon_t\}$.

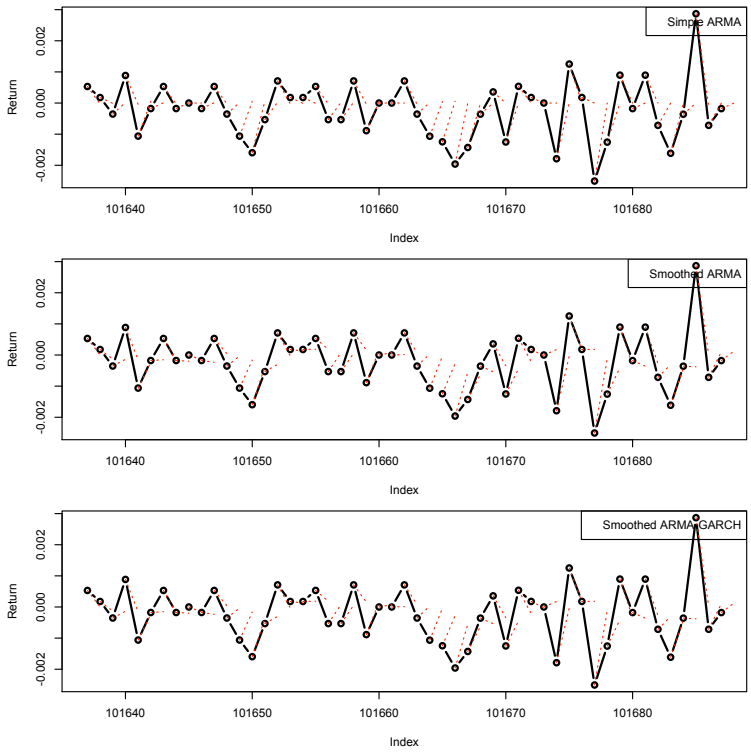

Figure 3.1: Output of three variations of the ARMA time series framework. Black circles are actual data-points and red dash-lines indicate the next period predictions. Top panel: simple ARMA; Center panel: ARMA with wavelet smoothing; Bottom panel: ARMA-GARCH. Dataset used is the continuously rolled near-expiry E-mini S&P 500 futures contract traded on the CME, sampled at 5 minute intervals. Sample period shown here is between 2008-05-06 13:40:00 and 2008-05-07 15:15:00. At each point, we fit a model using the most recent history of 1,000 data points, then make a 1-period ahead forecast using the fitting parameters.

See Figure 3.1 for predictions versus actual output for models given in the three examples above, all based on some variations of the ARMA framework.

Innovations State Space Model The innovations state space model can be written in terms of the underlying measurement and transition equations

$$\begin{aligned} y_t &= w\left(x_{t-1}\right) x_{t-1} + r\left(x_{t-1}\right) \epsilon_t \\ x_t &= f\left(x_{t-1}\right) x_{t-1} + g\left(x_{t-1}\right) \otimes \epsilon_t \end{aligned}$$

where y_t denotes the observation at time t, and x_t denotes the state vector containing unobserved components that describe the level, trend and seasonality of the underlying time series; $\{\epsilon_t\}$ is a white noise series with zero mean and constant variance σ^2. Hyndman et al. (2000) gives a comprehensive treatment of frameworks that belong to this class of models and propose a methodology of automatic

forecasting that takes into account trend, seasonality and other features of the data. They cover 24 state space models in their framework that include additive and multiplicative specifications for the seasonal component and an additional damped specification for the trend component. Model selection is done using the AIC of the fitted models.

The general formulation of innovative state space model also include *additive Holt-Winters* (see Holt, 1957 and Winters, 1969) as a specific instance. Recall the *Holt-Winters Additive Model* has the following specification

$$\begin{cases} l_t = \alpha\left(y_t - s_{t-m}\right) + (1-\alpha)\left(l_{t-1} + b_{t-1}\right) & \text{Level} \\ b_t = \beta\left(l_t - l_{t-1}\right) + (1-\beta)\, b_{t-1} & \text{Growth} \\ s_t = \gamma\left(y_t - l_{t-1} - b_{t-1}\right) + (1-\gamma)\, s_{t-m}. & \text{Seasonal} \end{cases}$$

The h-step ahead forecast is then given by

$$\hat{y}_{t+h|t} = l_t + b_t h + s_{t-m+h_m^+}$$

where $h_m^+ = (h+1) \bmod m + 1$. If we define $\epsilon_t \triangleq \hat{y}_{t|t-1} - y_t$, then the *Holt-Winters Additive Model* can be expressed in the linear innovation framework as

$$\begin{aligned} y_t &= \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} x_{t-1} + \epsilon_t \\ x_t &= \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x_{t-1} + \begin{bmatrix} \alpha \\ \alpha\beta \\ \gamma \end{bmatrix} \otimes \epsilon_t, \end{aligned}$$

where the state vector $x_t = \begin{pmatrix} l_t & b_t & s_{t-m} \end{pmatrix}^\top$. This is an example of a linear innovations state space model.

Figure 3.2 compares the 1-period predicted return with the actual realized return for the sample period between 2008-05-06 13:40:00 and 2008-05-07 15:15:00. At each point, we fit a model using the most recent history of 1,000 data points, then make a 1-period ahead forecast using the fitting parameters.

Neural Network Neural network is a popular subclass of nonlinear statistical models. See Hastie et al. (2003) for a comprehensive treatment of this class of models. The structural model for a *feed forward - back propagation* network, also known as *multilayer percetrons* in the neural network literature, is given by

$$F\left(\underline{x}\right) = \sum_{m=1}^{M} b_m S\left(a_{0m} + \sum_{j=1}^{m} a_{jm} x_j\right) + b_0$$

where $S : \mathbb{R} \to (0,1)$ is known as the activation function, and $\left\{b_m \left\{a_{jm}\right\}_0^M\right\}_0^M$ are the parameters specifying $F\left(\underline{x}\right)$. Figure 3.3 illustrates a simple feed-forward neural network with four inputs, three hidden activation units and one output. For illustration purpose, a (15,7,1) neural net is fitted to the same dataset, with tan-sigmoid, $S\left(x\right) = {}^{e^x - e^{-x}}/_{e^x + e^x}$, as the activation function for the hidden layer. The model inputs include lagged returns, lagged trading direction (+1 for positive return and -1 for negative return) indicators and logarithm of the ratio of market high to market low prices,

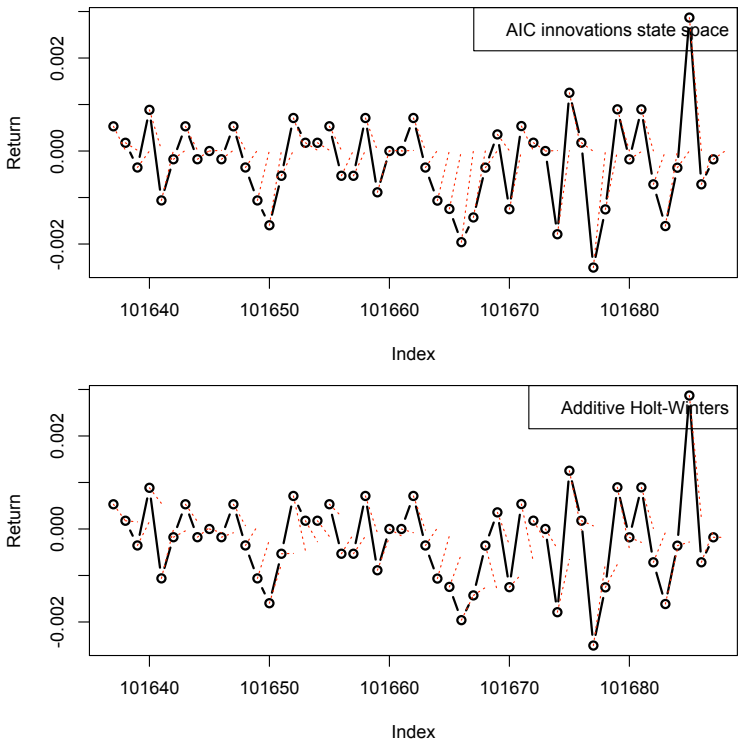

Figure 3.2: Output based on AIC general state space model following Hyndman et al. (2000) and the Holt-Winters Additive models. Black circles are actual data-points and red dash-lines indicate the next period predictions. Dataset used is the continuously rolled near-expiry E-mini S&P 500 futures contract traded on the CME, sampled at 5 minute intervals. Sample period shown here is between 2008-05-06 13:40:00 and 2008-05-07 15:15:00. At each point, we fit a model using the most recent history of 1,000 data points, then make a 1-period ahead forecast using the fitting parameters.

defined over a look-back period of 1,000 observations, as a proxy for volatility. Figure 3.4 shows the model output.

3.2 Self Excited Counting Process and its Extensions

Until recently, the majority of time series analyses related to financial data has been carried out using regularly spaced price data (see Section 3.1 for the outline of some commonly used models based on equally spaced data), with the goal of modeling and forecasting key distributional characteristics of future returns, such as expected mean and variance. These time series data mainly consist of daily closing prices, where comprehensive data are widely available for a large set of asset classes. With the recent rapid development of high-frequency finance, the focus has shifted to intra-day tick data, which record every transaction during market hours, and come with irregularly spaced time-stamps. We could resample the dataset and apply the same analyses as before, or we could try to explore additional information that the inter-arrival times may convey in terms of likely future trade direction.

In order to take into account of these irregular occurrences of transactions properly, we can adopt the framework of a point process. In a doubly stochastic framework (see Bartlett, 1963), both the counting process and the driving intensity are stochastic. A point process is called self-excited if the

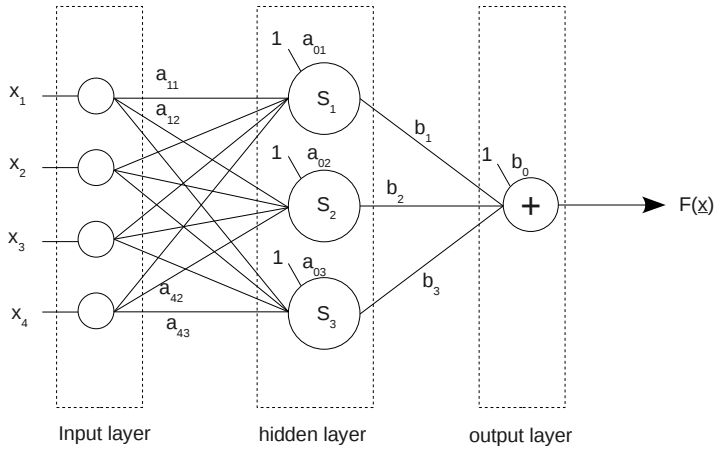

Figure 3.3: Diagrammatic illustration of a feed-forward (4,3,1) neural network.

current intensity of the events is determined by events in the past, see Hawkes (1971). It is widely accepted and observed that volatility of price returns tends to cluster. That is, a period of elevated volatility is likely to be followed by periods of similar levels of volatility. Trade arrivals also exhibit such clustering effect (Engle and Russell, 1995), for example a buy order is likely to be followed closely by another buy order. These orders tend to cluster in time.

There has been a growing amount of literature on the application of point process to model inter-arrival trade durations, see for example Engle and Russell (1997), and Bowsher (2003) for a comprehensive survey of the latest modeling frameworks.

This section of the thesis extends previous work on the application of self-excited process to model high frequency financial data. The extension comes in the form of a marked version of the process in order to take into account trade size influence on the underlying arrival intensity. In addition, by incorporating information from the *limit order book* (LOB), the proposed framework takes into account a measure of supply-demand imbalance of the market by parametrize the underlying base intensity as a function of this imbalance measure. Given that the main purpose of this section is to outline the methods to incorporate trade size and order-book information in a self-excited point process framework, empirical comparison of the proposed framework to other similar and related models are reserved for future research.

For a comprehensive introduction to the theory of point process, see Daley and Vere-Jones (2003) and Bremaud (1980). The following sub-sections give a brief outline of the self-excited point process framework in order to motivate the extensions that follow in Section 3.2.3.

3.2.1 Univariate Case

Consider a simple counting process for the number of trade events, N_t, characterized by arrival time of the trades, $\{t_i\}_{i\in\{0,1,2,\ldots T\}}$, a sequence of strictly position random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $t_0 = 0$ and $0 < t_i \leq t_{i+1}$ for $i \geq 1$. We allow the intensity of the process be itself stochastic, characterized by the following Stieltjes integral

$$\lambda_t = \mu + \int_{u<t} h\,(t-u)\,dN_u$$

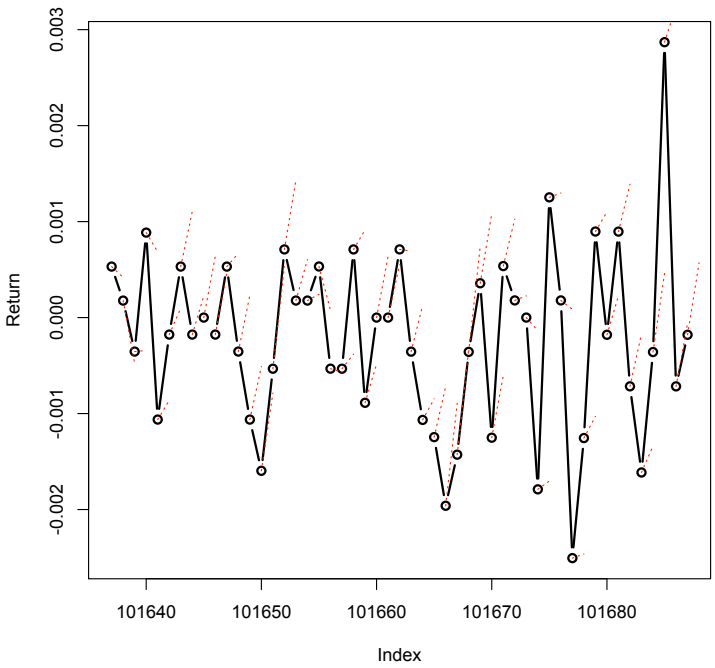

Figure 3.4: Output based on a feed-forward (15,7,1) neural network. Black circles are actual data-points and red dash-lines indicate the next period predictions. Dataset used is the continuously rolled near-expiry E-mini S&P 500 futures contract traded on the CME, sampled at 5 minute intervals. Sample period shown here is between 2008-05-06 13:40:00 and 2008-05-07 15:15:00. At each point, we fit a model using the most recent history of 1,000 data points, then make a 1-period ahead forecast using the fitting parameters.

where $(N_t : t \geq 0)$ is a non-explosive counting process with intensity λ_{t-}, μ is the *base intensity,* and $h : \mathbb{R}^+ \to \mathbb{R}^+$ is a non-negative function that parametrizes the self-excitation behavior.

Proposition 30. *Univariate Hawkes Process. Let λ_t be the intensity of the counting process of a particular form of the self-excited process that, under the usual assumptions, satisfies the following stochastic differential equation (SDE)*

$$d\lambda_t = \beta\left(\rho\left(t\right) - \lambda_t\right)dt + \alpha dN_t.$$

Assuming that $\rho\left(t\right) \equiv \mu$, then the solution for λ_t can be written as

$$\lambda_t = \mu + \alpha \int_0^t e^{-\beta(t-u)} dN_u \tag{3.2}$$

where μ is known as the long run or base intensity, i.e. the intensity if there has been no past arrival.

The linkage between the intensity and the underlying counting process N_t is via the *Doob-Meyer* decomposition and the two associated filtrations $\mathcal{H}_t \subset \mathcal{F}_t$, one for the intensity and the other for

the jump time, given by the following sigma-algebras

$$\mathcal{H}_t = \sigma\left\{\lambda_s : s \leq t\right\}$$

and

$$\mathcal{F}_t = \sigma\left\{N_s : s \leq t\right\}.$$

The characteristic function can be written as

$$\mathbb{E}\left[e^{i\upsilon(N_s - N_t)}\middle|\,\mathcal{F}_t\right] = e^{-\Psi(\upsilon)(\Lambda_s - \Lambda_t)}$$

where $\Psi(\upsilon) = 1 - e^{i\upsilon}$; $\Lambda_t = \int_0^t \lambda_u du$ is known as the *compensator* of the process, and $M_t = N_t - \Lambda_t$ is a $\mathcal{F}_t$-adapted martingale. Conditional on the realization of the compensator, the process is non-homogeneous Poisson with deterministic intensity

$$\begin{aligned}
\lim_{\delta t \to 0} \frac{1}{\delta t}\mathbb{E}\left[N_{t+\delta t} - N_t \middle|\, \mathcal{F}_t\right] &= \lim_{\delta t \to 0} \frac{1}{\delta t}\mathbb{E}\left[\mathbb{E}\left[\Lambda_{t+\delta t} - \Lambda_t \middle|\, \mathcal{H}_t \vee \mathcal{F}_t\right] \middle|\, \mathcal{F}_t\right] \\
&= \lim_{\delta t \to 0} \frac{1}{\delta t}\mathbb{E}\left[\int_t^{t+\delta t} \lambda_u du \middle|\, \mathcal{F}_t\right] \\
&= \lambda_t.
\end{aligned}$$

We can simulate this self-excited intensity process by the usual thinning method (Ogata, 1981). Figure 3.5 shows one particular realization of the simulated process. Note the clustering of intensity as a result of the self-excitation feature of the modeled process.

To obtain the compensator Λ_t, a simple piecewise integration of the intensity gives

$$\begin{aligned}
\int_0^t \lambda(u)\, du &= \int_0^t \mu du + \int_0^t \sum_{t_i < u} \alpha e^{-\beta(u - t_i)} du \\
&= \mu t - \frac{\alpha}{\beta} \sum_{t_i < t} \left(e^{-\beta(t - t_i)} - 1\right).
\end{aligned}$$

Theorem 31. *Time Change Theorem (see for exampleDaley and Vere-Jones (2003) for a more rigorous treatment). Given a point process with a conditional intensity function λ_t, define the time-change*

$$\Lambda_t = \int_0^t \lambda(u)\, du$$

where the filtration $\mathcal{H}_t = \sigma\left\{0 < t_1 < t_2, ..., t_i \leq t\right\}$. Assume that $\Lambda_t < \infty$ a.s. $\forall t \in (0, T]$, then $\left\{\Lambda(t_i)\right\}_{i=0,1,...,n}$ is a standard Poisson process.

By application of this *time change* theorem, we can transform an univariate Hawkes process back to a standard Poisson process. The transformed process can then be referenced against theoretical quantiles in order to assess goodness-of-fit, as it is done in Section 3.2.4.

3.2.2 Bivariate Case

In a multivariate setting, in addition to self-excitation, there is the possibility of *cross-excitation*, for which jumps of one process can elevate intensity and hence induce jump in other processes. A linear

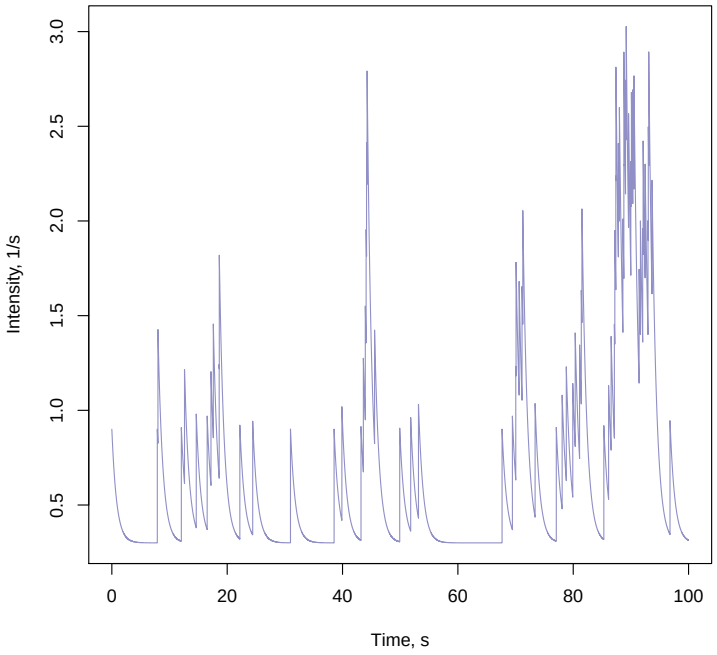

Figure 3.5: Simulated intensity of an univariate Hawkes process, with $\mu = 0.3$, $\alpha = 0.6$ and $\beta = 1.2$.

bivariate self-excited process with cross-excitation can be expressed, by modifying (3.2), to give

$$\begin{cases} \lambda_1(t) = \mu_1 + \int_0^t \upsilon_{11}(t-s)\, dN_1(s) + \int_0^t \upsilon_{12}(t-s)\, dN_2(s) \\ \\ \lambda_2(t) = \mu_2 + \int_0^t \upsilon_{22}(t-s)\, dN_2(s) + \int_0^t \upsilon_{21}(t-s)\, dN_1(s) \end{cases} \tag{3.3}$$

where we could consider λ_1 and λ_2 as the intensity of market orders traded on the bid and ask sides, respectively. Note that market orders traded on bid side are *sell orders* and those on the ask side are *buy order*s. Consider the following parametrization with exponential decay,

$$\upsilon_{ij}(s) = \alpha_{ij} e^{-\beta_{ij} s},\ \beta_{ij} \geq 0$$

then we can rewrite the Stieltjes integral in (3.3) as

$$\begin{cases} \lambda_1(t) = \mu_1 + \sum_{t_i<t} \alpha_{11} e^{-\beta_{11}(t-t_i)} + \sum_{t_j<t} \alpha_{12} e^{-\beta_{12}(t-t_j)} \\ \\ \lambda_2(t) = \mu_2 + \sum_{t_j<t} \alpha_{22} e^{-\beta_{22}(t-t_j)} + \sum_{t_i<t} \alpha_{21} e^{-\beta_{21}(t-t_i)} \end{cases}$$

where t_i and t_j are $\mathcal{F}_t$-adapted jump times for bid and ask side market orders, respectively. This exponential parametrization is in reasonable agreement with empirical findings, as illustrated in Figure 3.6 which shows the empirical conditional intensity.

Figure 3.7 shows a particular realization of a simulated bivariate intensity process. Observe the cross-excitation dynamics between the two processes and the exponential decay after each jump.

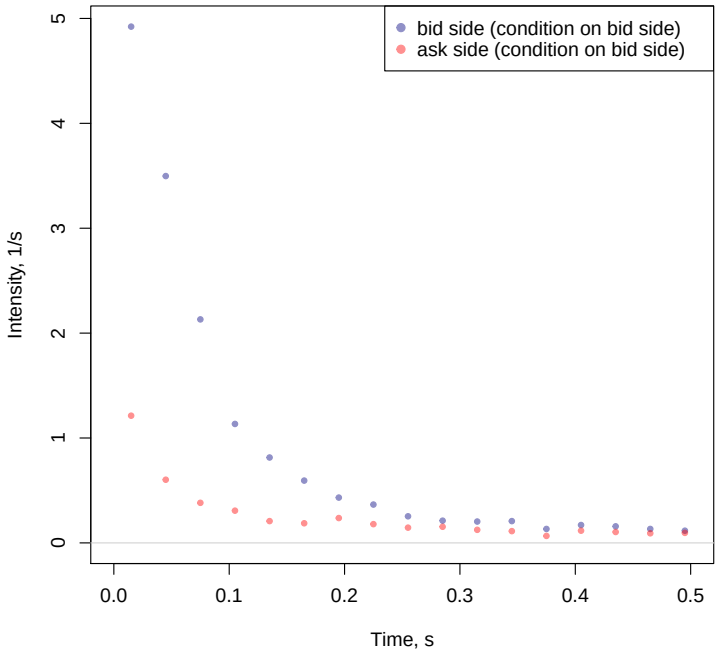

Figure 3.6: Conditional intensity of bid and ask side market orders following an order submitted on the bid side of the market, estimated with bin size ranging from 30 to 500 milliseconds, using BP tick data on 25 June 2010.

To obtain the compensator $\Lambda_1(t)$, we can integrate the intensity piecewise to give

$$\begin{aligned}\int_0^t \lambda_1(u)\,du &= \int_0^t \mu_1 du + \int_0^t \sum_{t_i<u} \alpha_{11} e^{-\beta_{11}(u-t_i)} du + \int_0^t \sum_{t_j<u} \alpha_{12} e^{-\beta_{12}(u-t_j)} du \\ &= \mu_1 t + \frac{\alpha_{11}}{\beta_{11}} \sum_{t_i<t} \left(1 - e^{-\beta_{11}(t-t_i)}\right) + \frac{\alpha_{12}}{\beta_{12}} \sum_{t_j<t} \left(1 - e^{-\beta_{12}(t-t_j)}\right)\end{aligned}$$

and similarly for $\Lambda_2(t)$.

Proposition 32. *The log-likelihood function for the bivariate process can be written as (Ogata, 1978)*

$$\begin{aligned}L_T(\mu_1,\mu_2,\beta_{11},\beta_{12},\beta_{21},\beta_{22},\alpha_{11},\alpha_{12},\alpha_{21},\alpha_{22}) &= L_T^{(1)}(\mu_1,\beta_{11},\beta_{12},\alpha_{11},\alpha_{12}) \quad (3.4)\\ &\quad + L_T^{(2)}(\mu_2,\beta_{21},\beta_{22},\alpha_{21},\alpha_{22}) \quad (3.5)\end{aligned}$$

where

$$\begin{aligned}L_T^{(1)}(\mu_1,\beta_{11},\beta_{12},\alpha_{11},\alpha_{12}) &= -\mu_1 T - \frac{\alpha_{11}}{\beta_{11}} \sum_{t_i<T} \left(1 - e^{-\beta_{11}(T-t_i)}\right) - \frac{\alpha_{12}}{\beta_{12}} \sum_{t_j<T} \left(1 - e^{-\beta_{12}(T-t_j)}\right) \\ &\quad + \sum_{\{i:t_i<T\}} \log\left(\mu_1 + \alpha_{11} R_{11}(i) + \alpha_{12} R_{12}(i)\right)\end{aligned}$$

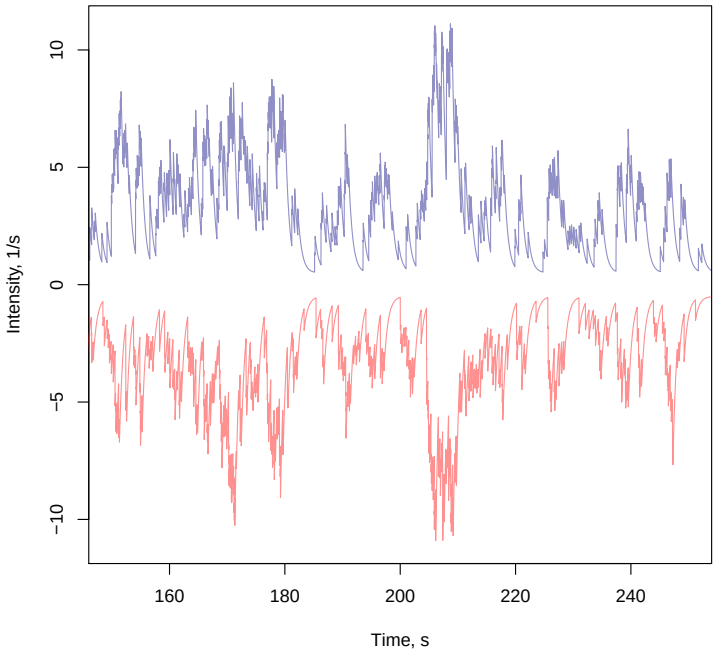

Figure 3.7: Simulated intensity of a bivariate Hawkes process. Path in blue (top) is a realization of the λ_1 process; path in red (bottom) is that of the λ_2 process, inverted to aid visualization. Parameters: $\mu_1 = \mu_2 = 0.5$, $\alpha_{11} = \alpha_{22} = 0.8$, $\alpha_{12} = \alpha_{21} = 0.5$, $\beta_{11} = \beta_{12} = 1.5$ and $\beta_{22} = \beta_{21} = 1.5$.

with the following recursion

$$R_{11}(i) = e^{-\beta_{11}(t_i - t_{i-1})}\left(1 + R_{11}(i-1)\right)$$

$$R_{12}(i) = e^{-\beta_{12}(t_i - t_{i-1})}\left(R_{12}(i-1)\right) + \sum_{\{j' : t_{i-1} \le t_{j'} < t_i\}} e^{-\beta_{12}\left(t_i - t_{j'}\right)}$$

and similarly for $L_T^{(2)}(\mu_2, \beta_{21}, \beta_{22}, \alpha_{21}, \alpha_{22})$, R_{22} *and* R_{21}. *Note by setting* α_{12} *and* α_{21} *to zero, we recover the log-likelihood functions for the univariate case.*

3.2.3 Taking volume and orderbook imbalance into account

For trading on the major US stock exchanges such as NYSE and NASDAQ, most transactions are in *round lots*, the normal size of trading for a security, which is generally 100 shares or more. *Odd lot* orders are orders with size less than the minimum round lot amount, and these orders have less favorable queue positioning and may incur additional clearing fees at the exchanges. See rules posted by the exchanges for a comprehensive treatment of the regulation and requirements related to odd lot execution and other transaction related rules (e.g. Rule 124(c) of the NYSE Rules). Since the trade size is decided by the order originator, it is a potential source of information. See for example Bouchaud et al. (2002) for a study of the statistical properties of market order size. Let w_{1i} and w_{2j} be the trade sizes for bid and ask side market orders at time t_i and t_j, respectively. Then the

resulting intensity for the marked point process can be written as

$$\begin{cases} \lambda_{1t} = \mu_1 + \frac{1}{\bar{w}_1}\sum_{t_i<t}\alpha_{11}w_{1i}e^{-\beta_{11}(t-t_i)} + \frac{1}{\bar{w}_2}\sum_{t_j<t}\alpha_{12}w_{2j}e^{-\beta_{12}(t-t_j)} \\ \\ \lambda_{2t} = \mu_2 + \frac{1}{\bar{w}_2}\sum_{t_j<t}\alpha_{22}w_{2j}e^{-\beta_{22}(t-t_j)} + \frac{1}{\bar{w}_1}\sum_{t_i<t}\alpha_{21}w_{1j}e^{-\beta_{21}(t-t_i)} \end{cases}$$

where $\bar{w}_1$ and $\bar{w}_2$ are the simple averages of trade size over the period,

$$\{t_i : 0 \leq t_i < t\} \cup \{t_j : 0 \leq t_j < t\}.$$

The intuition behind this functional form is straightforward - by giving more weight to trades accompanied by larger size, we are implicitly conjecturing that these trades convey more information than smaller, *noise* trades.

Proposition 33. *The log-likelihood function for the intensity process with trade size marks can be expressed as*

$$\begin{aligned} L_T^{(1)}(\mu_1, \beta_1, \alpha_{11}, \alpha_{12}) &= -\mu_1 T - \frac{\alpha_{11}}{\beta_{11}}\sum_{t_i<T}\frac{w_{1i}}{\bar{w}_1}\left(1 - e^{-\beta_{11}(T-t_i)}\right) - \frac{\alpha_{12}}{\beta_{12}}\sum_{t_j<T}\frac{w_{2j}}{\bar{w}_2}\left(1 - e^{-\beta_{12}(T-t_j)}\right) \\ &\quad + \sum_{\{i:t_i<T\}}\log\left(\mu_1 + \alpha_{11}R_{11}(i) + \alpha_{12}R_{12}(i)\right), \end{aligned}$$

with the following recursion

$$\begin{aligned} R_{11}(i) &= e^{-\beta_{11}(t_i - t_{i-1})}\left(\frac{w_{1i}}{\bar{w}_1} + R_{11}(i-1)\right) \\ R_{12}(i) &= e^{-\beta_{12}(t_i - t_{i-1})}\left(R_{12}(i-1)\right) + \sum_{\{j':t_{i-1}\leq t_{j'}<t_i\}}\frac{w_{2j'}}{\bar{w}_2}e^{-\beta_{12}\left(t_i - t_{j'}\right)}, \end{aligned}$$

and similarly for $L_T^{(2)}(\mu_2, \beta_{21}, \beta_{22}, \alpha_{21}, \alpha_{22})$, R_{22} *and* R_{21}.

The *limit order book* (LOB) is a trading method used by most electronic exchanges globally. It is an anonymous trading system that matches buyer and sellers by aggregating demands from both sides into a "trading book". At any time instance, the LOB contains multiple layers on the bid and ask sides of the book. Each layer corresponds to a different price level, normally separated by the minimum price increment. For most US exchanges, this minimum increment is \$0.01 for most stocks. Market agents have several options when it comes to placing an order to buy or sell securities. For example, *limit order* and *market order* are the two most common order types. A *limit orders* is an order to buy or sell a specific amount of shares of a stock at a specific price. When a limit order arrives into the exchange's order management system, it joins the bid or ask order queue at the price level specified by the order. The only change to the LOB that is visible to other market agents is an increase of queue size at that layer - no other information is disseminated. A *market order* is an order to buy or sell a stock at the current market price. For example, a market order to sell 1,000 IBM shares will take out 1,000 lots of liquidity at the top layer of the bid side of the order book. If the available liquidity is less than 1,000 at that level, the order will continue to execute at the next layer of the bid side order book with a lower price. This continues until 1,000 lots have been filled. The advantage of a market order is that it is almost always guaranteed to be executed. The

disadvantage is that the price one pays or gets depends on available liquidity of the market and the speed the order book changes over the short period of time[10] during which the market order seeks out liquidity.

Figure 3.8 shows snapshots of the evolution of a limit order book. For ease of visualization, the order queue at each time instance is scaled by the maximum size at that instance, across the whole book. Observe that the shape of the queue profile on both sides of the book varies as the price changes. The profile of the order book signals potential supply-demand imbalance of the market, and changes in this profile convey information of investor's reaction to price changes.

An appropriately defined buy-sell imbalance measure will help extract information of the likely change in trade direction and intensity over the short run. Imbalance, when defined simply as the difference between total aggregate buy and sell orders, ignores the important fact that orders at different layers of the book have significantly different probability of being executed.

Figure 3.9 shows the expected time, in seconds, it takes for limit orders submitted at specific *order-distance* (measured in units of median price increments) from the prevailing best bid and ask prices to get completely filled. For example, for a limit sell order, a distance of 0 corresponds to a sell order at the prevailing best bid (i.e. an effective market order), and a distance of 1 corresponds to a sell order at a price which is one price increment higher than the best bid (i.e. at the best ask price). From this, we can obtain the empirical cumulative distribution of order completion time, and hence deduce the probability of completion within a specific time period for a limit order submitted at a specific number of price increments away from the best bid and ask. Figure 3.10 shows the empirical probability of a complete limit order fill within 5 seconds after submission, as a function of order-distance, assuming that the underlying true price process has zero drift - a reasonable assumption given the short time frame.

One way to quantify market supply and demand is via a probability weighted volume, defined below.

$$\bar{v}\left(t,\tau,L;i\right)=\frac{1}{\sum_{i,l}v_{t,l;i}}\sum_{l=0}^{L}v_{t,l;i}p_{l,i,\tau},$$

where $p_{l,i,\tau}=\mathbb{P}\left(t_f<t+\tau|\,l,i\right)$ is the probability of an order of type $i\in\{1,2\}$ submitted at layer l getting completely filled at time t_f, which is within τ seconds from order submission at time t. $v_{t,l;i}$ is the queue size at time t, at the l-th layer and on side i of the limit order book. Figure 3.11 shows the time series of the difference between the bid and ask side probability weighted cumulative volumes, $\bar{v}\left(t,\tau,L;1\right)-\bar{v}\left(t,\tau,L;2\right)$, for $t=\{t_i:i=0,1,\ldots,n\}$.

The base intensity, μ, controls the mean arrival rate of market orders. It is intuitive to conjecture that when there are more buy limit orders on the bid side of the order book than there are sell orders on the ask side of the book, the likelihood of an uptick in price increases, and similarly when this imbalance is reversed. This order book "skew" is a signal of market imbalance, and one method to incorporate this in a point process framework is by using our probability weighted volume measure

[10] Note: with the current technological and algorithmic advances of computer driven market making, the order book can react in less than $50\mu s$.

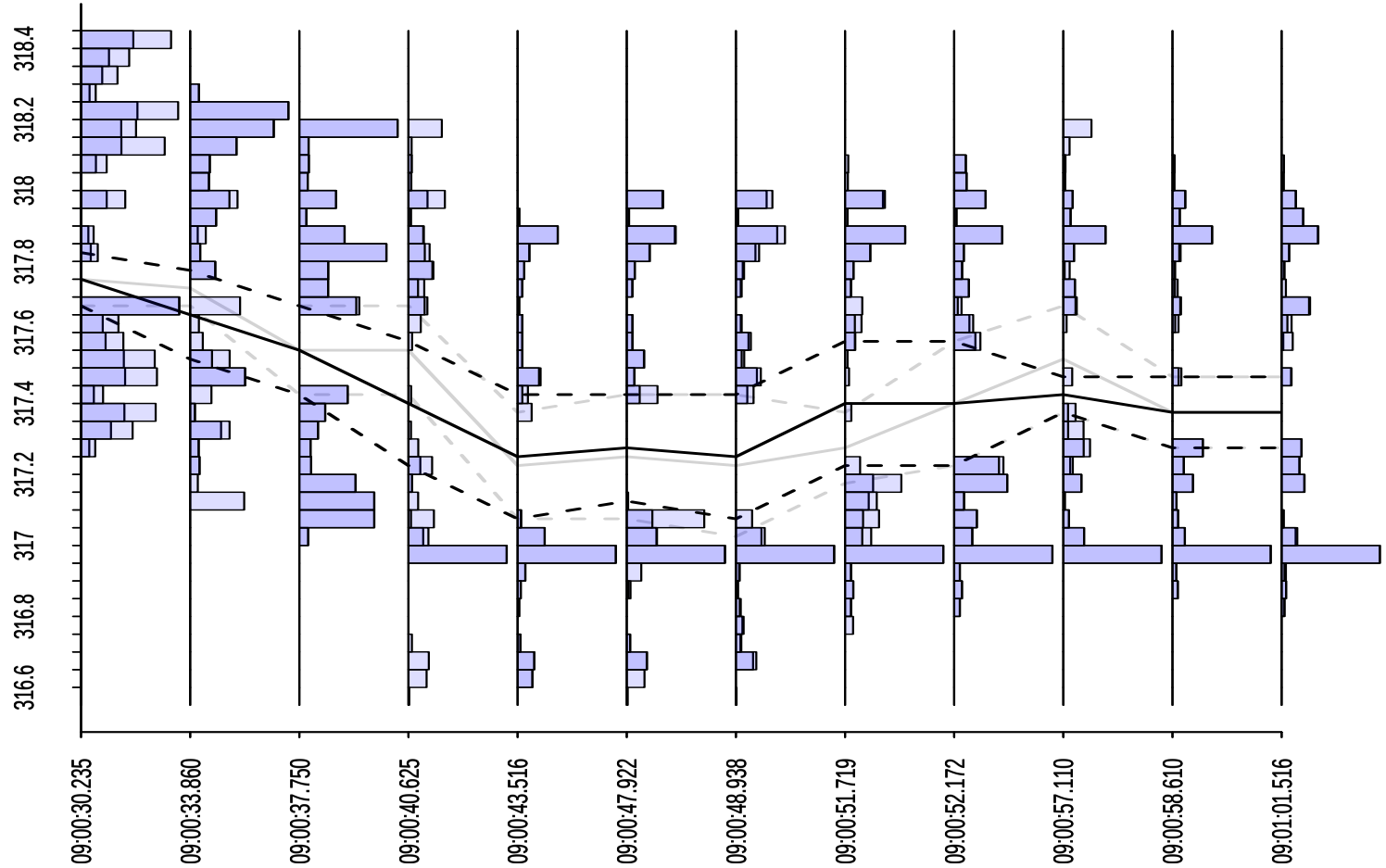

Figure 3.8: Snapshots showing the evolution of a ten layer deep limit order book just before a trade has taken place (gray lines) and just after (black lines) for BP. Dotted lines are for the best bid and ask prices. Solid line is the average or mid price. Bars are scaled by maximum queue size across the whole book and represented in two color tones to help identify changes in the order book just before and after a trade has taken place.

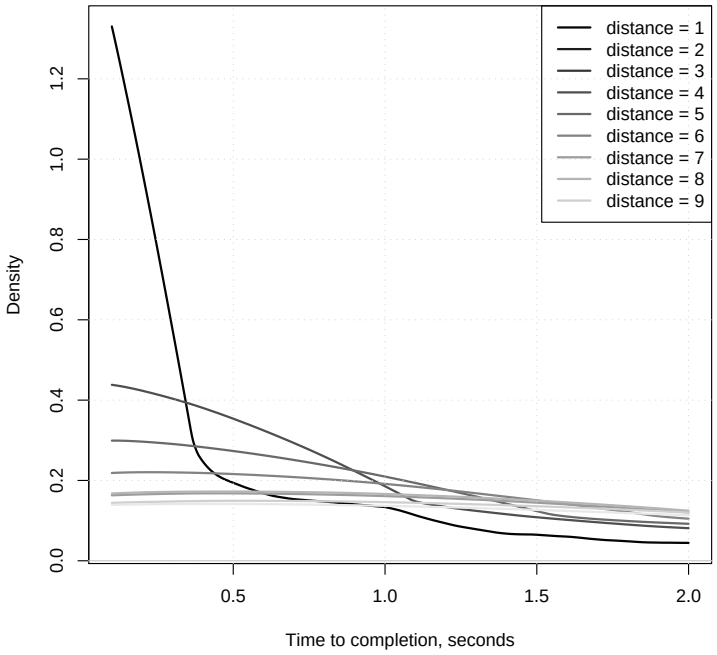

Figure 3.9: Time to order completion as a function of order submission distance from the best prevailing bid and ask prices. Distance is defined as the number of order book median price increments from the best top layer prices at the oppose side of the market. For example a distance of 1 corresponds to the top bid and ask prices and a distance of 2 corresponds to the second layer of the bid and ask sides.

to scale the base intensity of the underlying process, as shown below,

$$\begin{cases} \lambda_{1t} = \mu_1 \bar{v}_{2t} + \frac{1}{\bar{w}_1} \sum_{t_i<t} \alpha_{11} w_{1i} e^{-\beta_{11}(t-t_i)} + \frac{1}{\bar{w}_2} \sum_{t_j<t} \alpha_{12} w_{2j} e^{-\beta_{12}(t-t_j)} \\ \\ \lambda_{2t} = \mu_2 \bar{v}_{1t} + \frac{1}{\bar{w}_2} \sum_{t_j<t} \alpha_{22} w_{2j} e^{-\beta_{22}(t-t_j)} + \frac{1}{\bar{w}_1} \sum_{t_i<t} \alpha_{21} w_{1i} e^{-\beta_{21}(t-t_i)} \end{cases}$$

where $\bar{v}_{1t} = \bar{v}(t, \tau, L; 1)$ for bid side orders and similarly for $\bar{v}_{2t}$.

Proposition 34. *The log-likelihood function for the intensity process with both trade size mark and order book information can be expressed as*

$$\begin{aligned} L_T^{(1)}(\mu_1, \beta_1, \alpha_{11}, \alpha_{12}) &= -\mu_1 \sum_{t_i<T} \bar{v}_{2t_i}(t_i - t_{i-1}) - \frac{\alpha_{11}}{\beta_{11}} \sum_{t_i<T} \frac{w_{1i}}{\bar{w}_1} \left(1 - e^{-\beta_{11}(T-t_i)}\right) \\ &\quad - \frac{\alpha_{12}}{\beta_{12}} \sum_{t_j<T} \frac{w_{2j}}{\bar{w}_2} \left(1 - e^{-\beta_{12}(T-t_j)}\right) \\ &\quad + \sum_{\{i:t_i<T\}} \log\left(\mu_1 \bar{v}_{2t_i} + \alpha_{11} R_{11}(i) + \alpha_{12} R_{12}(i)\right) \end{aligned}$$

where the recursions R_{11} and R_{12} are the same as in Proposition 33, and similarly for $L_T^{(2)}(\mu_2, \beta_{21}, \beta_{22}, \alpha_{21}, \alpha_{22})$, R_{22} and R_{21}.

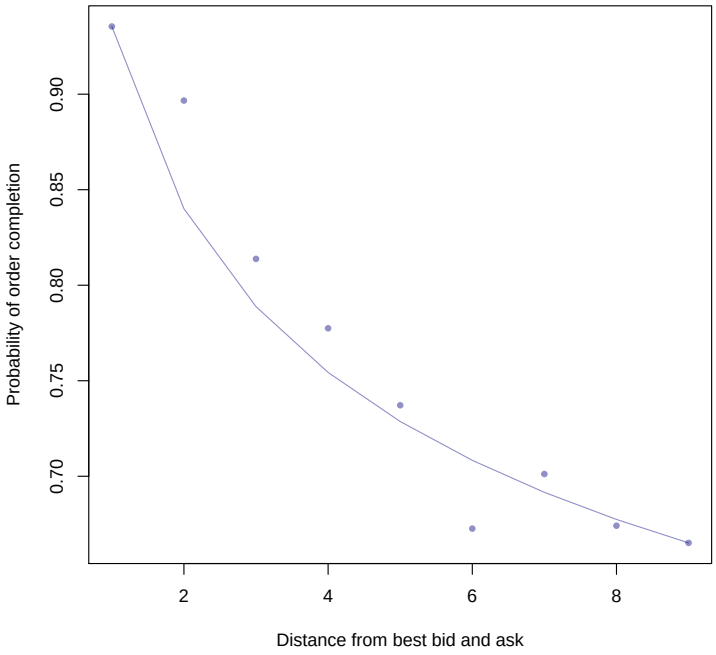

Figure 3.10: Probability of order completion within 5 seconds from submission. Dots are estimated based on empirical data and solid curve is based on the fitted power law function $0.935x^{-0.155}$.

3.2.4 Empirical Analysis

Data The order book data is obtained from the London Stock Exchange (LSE). This *Rebuild Order Book* (LSE, 2008) dataset provides full market depth (Level 3) intra-day order information and trading data, which allows us to reconstruct the complete order book system. The dataset contains order detail, order deletion and trade detail information. Records are time-stamped to the nearest millisecond.

British Petroleum PLC (BP) is used to illustrate the modeling framework and fitting process. Figure 3.12 shows the time series of price and bid-ask spread for 25 June 2010. The code that performs the order book reconstruction takes order details from the *order details record* file, then chronologically match the trade and deletion information in the *order history record* file. The result of this reconstruction procedure is a time series of snapshots of the order book, at every trade event. See Figure 3.8 for a 30-second picture of the evolution of the LOB for BP.

The sample period is chosen to be the entire trading day on 25 June 2010, from 08:05:00.000 to 16:25:00.000. The first and last 5 minutes near the opening and closing of the market is discarded, so as to stay clear of the periods near market auctions, where there is often a lot of noise (e.g. incorrectly recorded orders or transactions) in the data. There are a total of four key orders types: limit buy, limit sell, effective market buy and effective market sell. The effective market buy orders include market buy orders and limit orders that are submitted at or through the best ask price; and similarly for the effective market sell order which include market sell order and limit order that are submitted at or through the best bid price. These are orders submitted with the intention to induce

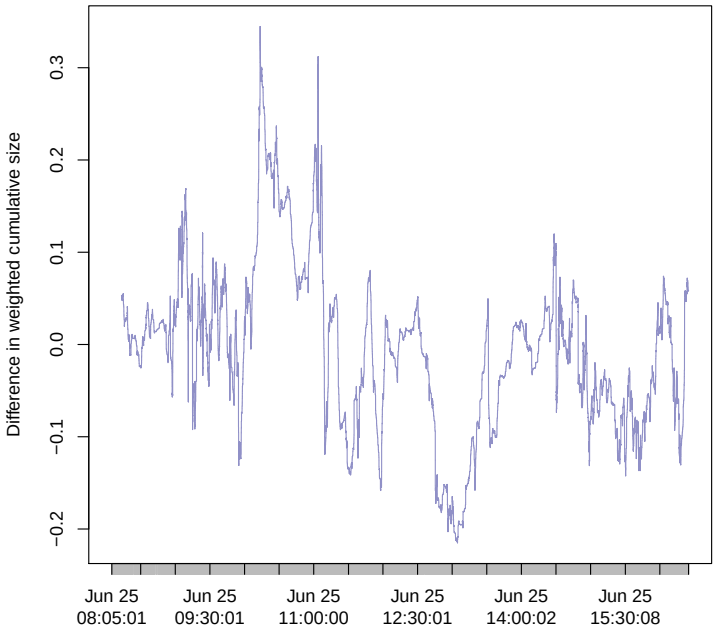

Figure 3.11: Time series for the difference in bid and ask side LOB probability weighted cumulative volume, $\bar{v}(t, \tau, L; 1) - \bar{v}(t, \tau, L; 2)$, for BP, on 25 June 2010.

immediate fill.

Definition of event For the analyses carried out in this paper, events are defined as the execution, both partial or complete, of buy and sell market orders. For the bid side order intensity, the inter-arrival times are defined to be the time, measured in milliseconds, between two effective market orders at the bid side; and similarly for the ask side orders. Even at this millisecond resolution, the dataset still contains trades that have taken place in quick successions such that they are stamped with the same time-stamp[11]. To be consistent with the definition of a simple point process, this thesis adopts the practice that, for trades stamped with the same time-stamp, it retains only the first trade and discards the rest. Although the dataset contains sequence identifiers that can potentially be used to sort trades in chronological order, there still remains the tasks of assigning unique time-stamps to the sorted trades. See Shek (2007) for analysis of other possible methods to deal with trades with identical time-stamps.

Figure 3.13 shows the empirical intensity of bid and ask side orders. Here the intensity is calculated based on the arrival rate of market orders within overlapping one minute windows. Note the widely observed *U-shape* activity profile, which indicates price discovery is concentrated near the opening and closing periods of the market. This suggests that we might consider introducing the extra complexity of modeling the base intensity with a periodic function that matches the observed activity profile, a point not explored in this thesis.

[11] Note that for liquid stocks, market interaction can and often happen at microsecond resolution.

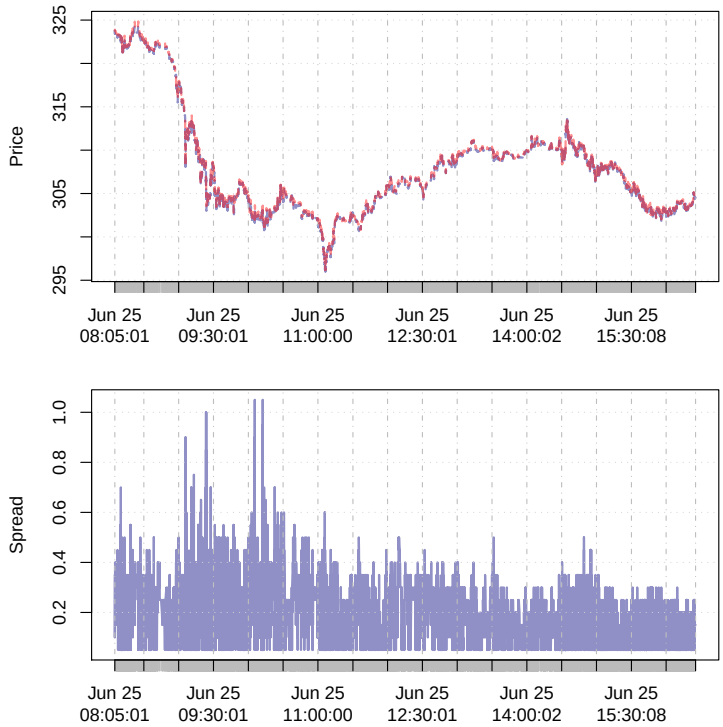

Figure 3.12: Top panel: time series for best prevailing bid and ask prices of the limit order book. Bottom panel: bid-ask spread. Both are for BP, on 25 June 2010.

Maximum Likelihood Estimation result Table 6 gives the parameters fitted by a MLE routine using the log-likelihood functions derived in the earlier sections. Results for four model specifications are presented here: bivariate, bivariate with trade size mark, bivariate with order book imbalance mark and bivariate with both trade size and order book imbalance marks. All parameters are significant (except for the two marked with †) at the 95% level. Some remarks on the fitted result:

- ▷ for both the bivariate and the bivariate with LOB marks models, two that best fit the data, μ_1 is larger than μ_2, which indicates that the mean intensity of market orders traded on the bid side is higher than those trade on the ask side; This reconciles with the overall downward drift of the market on that day, see Figure 3.12;
- ▷ self-excitation parameters α_{11} and α_{22} are both degrees of magnitude larger than their cross-excitation counterparts α_{12} and α_{21}, which suggests that although submitted orders on both sides of the LOB would induce an overall increase in trading activity, they are more likely to induce more orders of the same type;
- ▷ exponential decay rates β_{11} and β_{22} are also significantly larger in magnitude than their cross-excitation counterparts β_{12} and β_{21}, which suggests that persistency of intensity due to self-excitations is higher;
- ▷ β_{12} is higher than β_{21}, which suggests that market orders traded on the ask side is more likely to induce orders traded on the bid side than do bid side on ask side. This again reconciles with the overall downward drift of the market on the day.

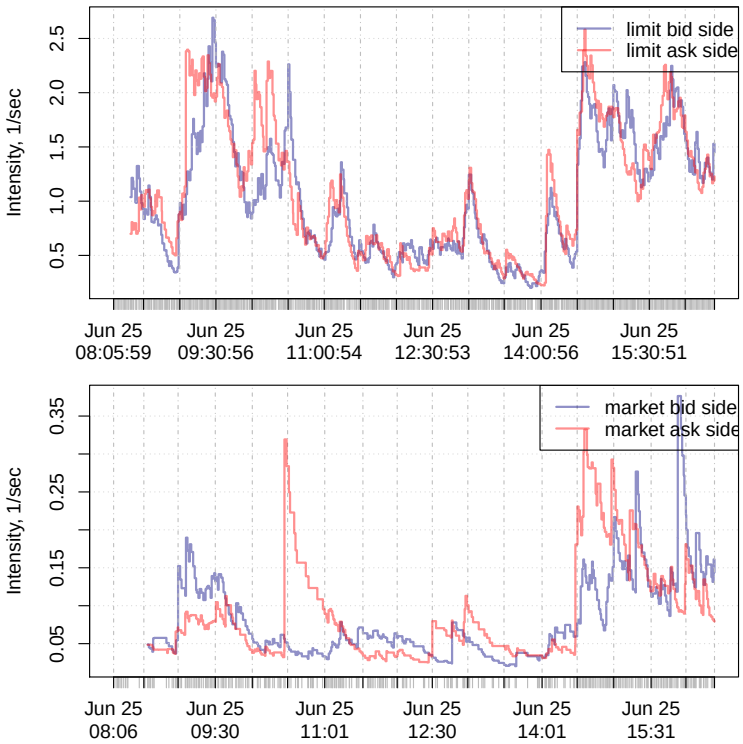

Figure 3.13: Unconditional arrival intensity of market and limit order on bid and ask sides of the order book, estimated using overlapping windows of one minute period, for BP on 25 June 2010.

Goodness of fit To assess the goodness of fit of the proposed models, the *Kolmogorov-Smirnov plot* (KS-plot) is used to visualize the relationship between empirical quantiles of the data and the theoretical quantiles of a reference distribution. Since the *time-changed* inter-arrival times lead to a standard, homogeneous, Poisson process, the reference distribution is therefore a standard exponential distribution. To construct the KS-plot, we first transform the time-changed inter-arrival times to uniform random variables on the interval $(0, 1)$ via the *cumulative distribution function* (CDF) of a standard exponential distribution. Then we order the transformed variables from smallest to largest, and plot these values against the CDF of a uniform distribution. If the model is correctly specified, then the points should align along the diagonal (Johnson and Kotz, 1972). The confidence bounds for the degree of fit can be constructed using the distribution of the *Kolmogorov-Smirnov* statistic, which for moderate to large sample sizes, the 99% confidence bounds are approximately $\pm 1.63/\sqrt{n}$ off the diagonal, where n is the sample size (Johnson and Kotz, 1972).

Figure 3.14 shows the KS-plot for the empirical inter-arrival times for bid and ask side market orders, together with the 99% confidence band and the *Kolmogorov-Smirnov* statistic. It clearly shows that a standard Poisson process is not adequate in capturing the dynamics of the underlying counting process. Figure 3.15 shows the KS-plot for models fitted to data on 25 June 2010, based on four variations of the framework discussed: bivariate, bivariate with trade size mark, bivariate with order book imbalance mark and bivariate with both trade size and order book imbalance marks. The KS-plots indicate that the self-excited point process framework is able to capture a significant amount of the underlying trading dynamics of market orders, in-sample. Also, it is quite clear

	Bivariate	Bivariate with size mark	Bivariate with LOB mark	Bivariate with size & LOB marks
μ_1	0.068	$0.000^{\dagger}$	0.193	$0.000^{\dagger}$
	(0.002)	(0.000)	(0.005)	(0.000)
μ_2	0.005	0.005	0.015	0.013
	(0.001)	(0.001)	(0.002)	(0.002)
α_{11}	2.726	2.901	2.803	2.900
	(0.114)	(0.129)	(0.116)	(0.129)
α_{22}	2.624	2.424	2.631	2.430
	(0.126)	(0.126)	(0.126)	(0.126)
α_{12}	0.575	0.004	0.563	0.004
	(0.068)	(0.000)	(0.068)	(0.003)
α_{21}	0.002	0.001	0.002	0.001
	(0.000)	(0.000)	(0.000)	(0.000)
β_{11}	5.211	6.740	5.418	6.740
	(0.217)	(0.295)	(0.224)	(0.295)
β_{22}	6.990	7.659	7.016	7.684
	(0.347)	(0.400)	(0.349)	(0.400)
β_{12}	8.422	0.007	8.204	0.007
	(1.076)	(0.001)	(1.072)	(0.001)
β_{21}	0.004	0.002	0.004	0.002
	(0.001)	(0.000)	(0.000)	(0.000)
$l(\theta)$	-16,319	-18.123	-16,208	-18,120

Table 6: MLE fitted parameters for the proposed models; standard errors are given in parenthesis. Sample date is 25 June 2010. † indicates that the value is not significant at 95% level.

from the plot that, for BP on 25 June 2010, including order size information does not help improve fit. This apparent lack of information from trade size could be a result of frequently used *slicing algorithms* that many trading systems adopt, in which large market orders are broken down into sizes comparable to the median trade size so as to minimize signaling effect. Note that the models that incorporate size and LOB marks do not nest the plain bivariate unmarked case, so there is no guarantee that the fitted result for those models will dominate that for the unmarked case. This can be seen in the KS-plot where for the in-sample period, the unmarked model seems to offer marginally better fit statistic. In-sample goodness of fit is only one part of model adequacy verification, we also need to verify how robust the fitted model is when applied to out-of-sample data.

For in-sample versus out-of-sample assessment, we have used two consecutive days of data to illustrate the robustness of our fitted models. We first estimate the parameters for the models using data from the in-sample period on 06 July 2009, with result shown in Figure 3.16. Then we apply the in-sample parameters to the out-of-sample period on 07 July 2009, in order to assess performance of the different models. From Figure 3.17, we see that all four models are reasonable robust, with only minor deterioration of performance in the out of sample period. Furthermore, we observe that the model which incorporates LOB information seems to have given a marginally more robust out-of-sample performance. Further research, when a more comprehensive dataset becomes available, is needed to further quantify this observation.

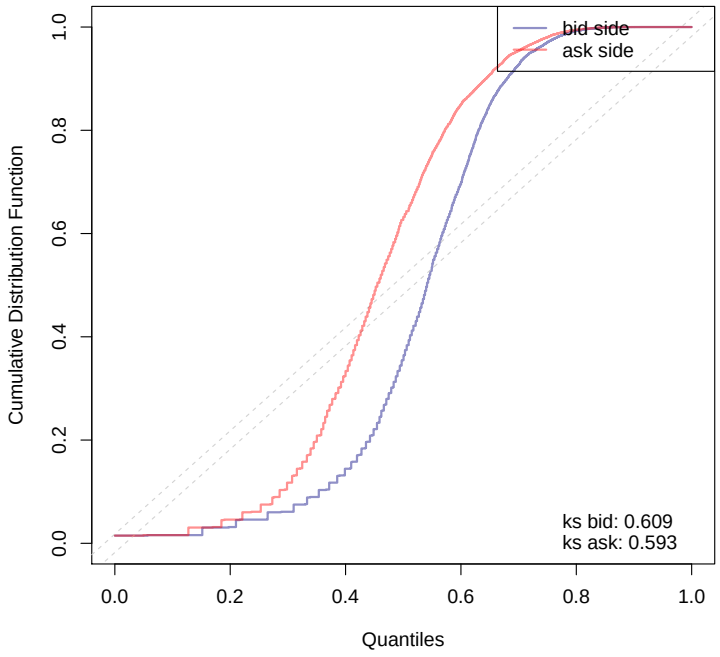

Figure 3.14: KS-plot for empirical inter-arrival times for market orders on the bid and ask sides of the market. Dash lines indicate the two sided 99% error bounds based on the distribution of the *Kolmogorov-Smirnov* statistic. Also shown is the value of the *Kolmogorov-Smirnov* test statistic for the two order types.

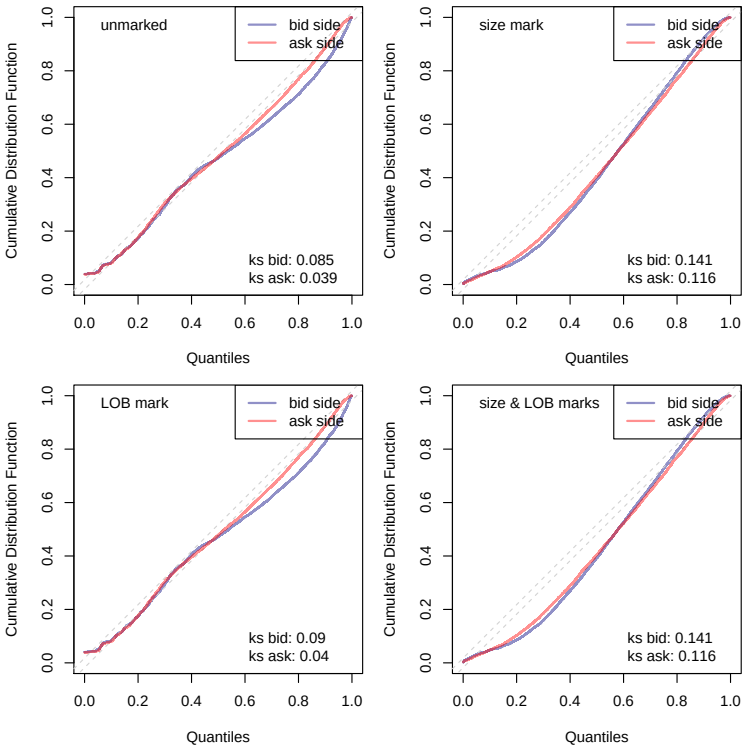

Figure 3.15: KS-plot based on four variations of the framework discussed: bivariate, bivariate with trade size mark, bivariate with order book imbalance mark and bivariate with trade size and order book imbalance marks. Fitted to sample data on 25 June 2010. Dash lines indicate the two sided 99% error bounds based on the distribution of the *Kolmogorov-Smirnov* statistic. Also shown is the value of the *Kolmogorov-Smirnov* test statistic for the two order types.

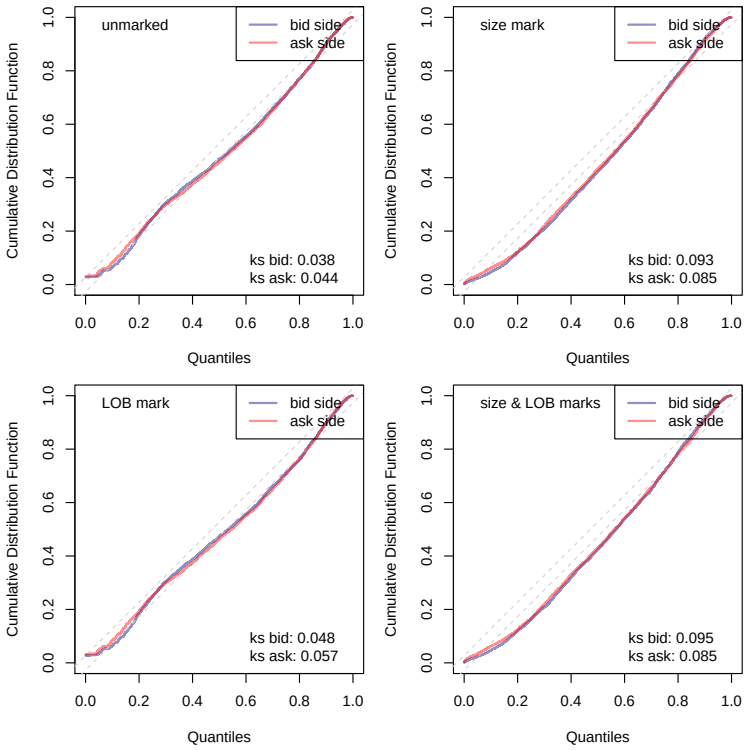

Figure 3.16: In sample KS-plot for empirical inter-arrival times for market order on the bid and ask sides of the LOB. Model parameters are fitted with data from in sample period on 06 July 2009 and applied to in sample period on 06 July 2009. Dash lines indicate the two sided 99% error bounds based on the distribution of the *Kolmogorov-Smirnov* statistic. Also shown is the value of the *Kolmogorov-Smirnov* test statistic for the two order types.

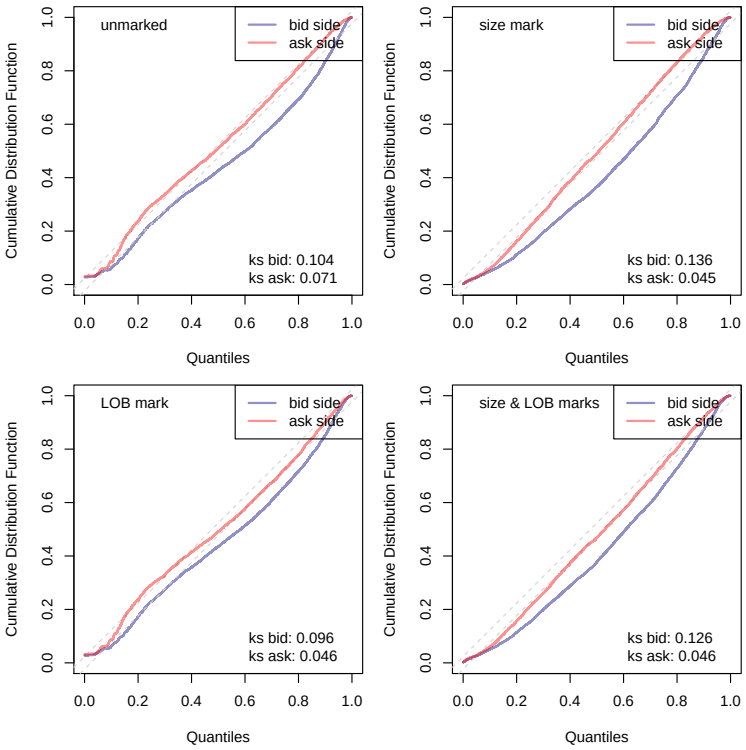

Figure 3.17: Out of sample KS-plot for empirical inter-arrival times for market order on the bid and ask sides of the LOB. Model parameters are fitted with data from in sample period on 06 July 2009 and applied to out of sample period on 07 July 2009. Dash lines indicate the two sided 99% error bounds based on the distribution of the *Kolmogorov-Smirnov* statistic. Also shown is the value of the *Kolmogorov-Smirnov* test statistic for the two order types.

4 Solving Cardinally Constrained Mean-Variance Optimal Portfolio

4.1 Literature Review

Portfolio optimization is a classic problem in financial economics. The underlying theory analyzes how wealth can be optimally invested in assets that differ in regard to their expected return and risk, and thereby also how risks can be reduced. In the classical one-period Markowitz mean-variance optimization framework (Markowitz, 1952), asset returns are modeled either under the assumption of a joint Gaussian distribution, or of a rational investor having a quadratic utility function. Given these assumptions, Markowitz has shown that the optimal portfolio for the investor is located on the *mean-variance efficient frontier*, in the sense that for all assets on this efficient frontier, for any given expected return, there is no other portfolio with lower variance; and for any given variance, there is no other portfolio with higher expected return. In the two dimensional mean-variance space, the efficient frontier is a parabola, whereas for mean-standard-deviation, it is a hyperbola.

Formally, an agent has mean-variance preference if her utility function has the following property

$$U\left(R_p\right) = f\left(E\left[R_p\right], var\left(R_p\right)\right),\ f_1 > 0,\ f_2 < 0,$$

where R_p is the portfolio return, and f_k is the partial derivative of the function f with respect to the k-th coordinate. It assumes that even when the underlying distribution of the portfolio return is not Gaussian, the investor still only cares about the first two moments.

With simple constraints of the original optimization problem and also in the case of many of its extensions, this problem is readily solvable using a standard quadratic programming (QP) solver, relying on algorithms such as those based on the null-space method, trust-region method or sequential quadratic-programming, see for example Markowitz (1987) and Perold (1984). However, computational issues can still arise if problems are very large and solutions are needed quickly. Figure 4.1 illustrates the exponential time complexity of a typical QP, where it can be seen that the time it takes to solve a QP as a function of problem size, N, scales approximately as $\mathcal{O}\left(N^3\right)$.

For practical reasons, such as in presence of transaction costs, fees and other administrative concerns, we often faces the problem of requiring a constraint that limits the number of assets in which we can invest as part of a portfolio, out of a large universe of potential candidates. One classic example is the problem of stock index tracking, where an index with a large number of constituents needs to be tracked by a portfolio using a much smaller subset of underlying assets. This leads to the introduction of cardinality constraints, which can increase the complexity of the problem significantly. In fact, the problem is known to be *NP*-hard (Shaw et al., 2008) and hence optimality is not guaranteed in polynomial time. For these type of problems, even at a modest size, computationally effective algorithms do not exist and, up until recently, there has been relatively little work presented in the literature.

General *cardinality constrained quadratic program* (CCQP) with linear constraints can be expressed as

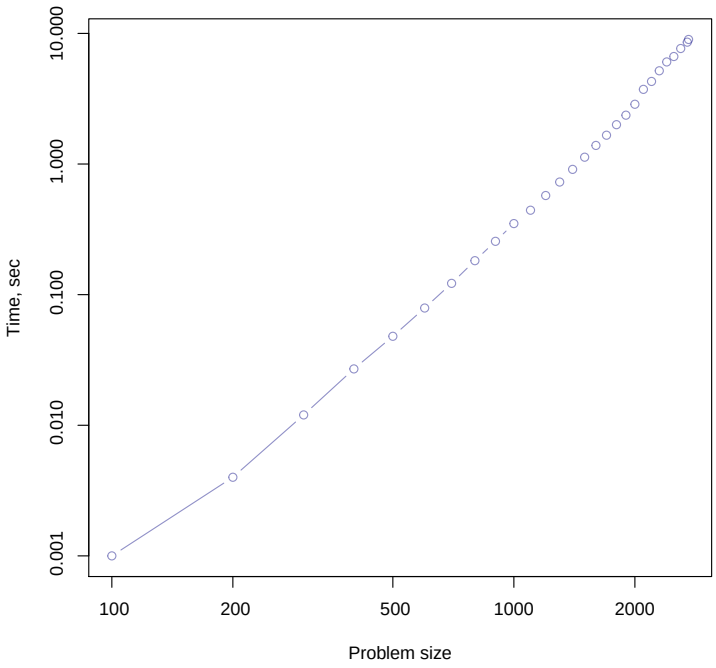

Figure 4.1: Time complexity of simple QP with only the budget constraint. Problem size is the number of names in the portfolio. Both axes are in log scale. Result produced using the *R* built-in QP solver based on a dual algorithm proposed by Goldfarb and Idnani (1982) that relies on Cholesky and QR factorization, both of which have cubic complexity.

$$\min_{x} \quad f(x) = -c^{\top}x + \lambda x^{\top}Hx \tag{4.1}$$

$$\text{s.t.} \qquad Ax \geq b \tag{4.2}$$

$$\textstyle\sum_i \mathbf{1}_{\{x_i \neq 0\}} = K, \tag{4.3}$$

where c, $x \in \mathbb{R}^{n\times 1}$, $H \in \mathbb{R}^{n\times n}$ is positive definite, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^{n\times 1}$ and $m \leq n$. K is the cardinality constraint and the scalar λ is known as the *relative risk aversion* parameter. Let x^* be our solution set, the indicator function in (4.3) is given by

$$\mathbf{1}_{\{x_i \neq 0\}} = \begin{cases} 1 & x_i \in x^* \\ 0 & \text{o.w.} \end{cases}$$

In other words, the constraint expressed in (4.3) forces the number of non-zero elements of the solution vector x be equal to a predetermined scalar, K. Note that there is no explicit non-negativity constraint of $x_i \geq 0 \quad \forall i$, hence short selling (i.e. selling what we do not own) is allowed. This together with the constraint $\sum_i x_i = 0$ gives what is known as a dollar neutral portfolio. It is called dollar neutral because the dollar exposure on the *long* part (i.e. the part of the portfolio that consists of the stocks bought) equals to the exposure on the *short* part (i.e. the part of the portfolio which

consists of the stocks that are sold short).

In the context of portfolio construction, x is the proportion of the capital allocation to each asset in the basket, c is the expected return of N candidate assets, i.e. a portfolio manager's subjective assessment of the one period ahead forecast of asset returns; H is the corresponding forecast for the variance-covariance matrix. The matrix A encapsulates M sets of linear constraints needed, for example, to impose holding limits, maximum leverage, etc. The scalar K limits the number of different assets the portfolio is allow to hold. c is often modeled using time series analysis based on historical asset returns together with other covariates that could enhance the forecasting signal, see Section (3). More sophisticated frameworks explore additional sources of information such as those embedded in the limit-order-book, see for example Shek (2010). H is often modeled based on historical returns, $H = \mathbb{E}\left[\left(R - \bar{R}\right)^\top \left(R - \bar{R}\right)\right]$, where $R \in \mathbb{R}^{T \times N}$ is the return matrix for T observations of N assets and $\bar{R}$ is the temporal mean. More sophisticated frameworks explicitly explore known dynamics, such as clustering effect of variances, and rely on using high-frequency intraday tick-data to enhance forecast power, see Section (2).

The cardinality constraint in (4.3) changes the complexity of the problem from that of an inequality constrained convex QP to that of a non-convex QP in which the feasible region is a mixed-integer set with potentially many local optima. Shaw et al. (2008) has reduced a 3-partitioning problem to a CCQP, hence establishing the *NP*-hardness of the problem.

Although by construction, the covariance matrix H is positive-definite, it is usually ill conditioned. One common method to rectify the problem is by factor analysis, where we decompose the return matrix into a rank k symmetric matrix, where $k \ll N$, and a diagonal matrix, known as the factor variance and specific variance matrix, respectively. In doing so, our implicit prior is that the aggregate market dynamics is spanned by a set of k orthogonal basis, with the remaining $N - k$ dimensions spanned by uncorrelated asset specific factors.

The type of problems that CCQP represents can be broadly categorized as a *mixed-integer programming* (MIP) problem with a quadratic objection function, resulting in a class of problem know as *mixed integer quadratic program* (MIQP). A typical CCQP can be expressed as

$$\begin{aligned}
\min_{x,y} \quad & f(x) = -c^\top x + \tfrac{1}{2} x^\top H x \\
\text{s.t.} \quad & Ax \geq b \\
& \textstyle\sum_i y_i = K \\
& y_i \geq x_i \geq -y_i \qquad i = 1, \ldots, N \\
& y_i \in \{0, 1\} \qquad i = 1, \ldots, N,
\end{aligned}$$

where we have introduced a binary variable $y \in \{0, 1\}$ to enforce cardinality of the solution set. Solvers with mixed-integer capability are less readily available than standard convex QP solvers. One obvious method is via successive truncation, see Algorithm 7, where a sequence of relaxed QP is solved and, at each iteration, a selection of assets with small or zero weights are truncated off. More sophisticated methods, such as branch-and-bound or branch-and-cut, are often used to solve MIP. Alternatively, the problem can be cast as a generic global optimization problem, then solved by heuristic based methods such as genetic or differential evolution algorithms and simulated annealing.

4.1.1 Branch-and-bound

Branch-and-bound (B-B) algorithm is one of the main tools used to solve the types of *NP*-hard discrete optimization problems to optimality, by essentially searching the complete space of solutions. Since explicit enumeration is normally impossible due to an exponentially increasing number of potential solutions, B-B algorithm uses bounds for the objective function combined with the value of the current best solution, which enable the algorithm to search parts of the solution space more efficiently. The algorithm was introduced by Land and Doig (1960), and generalized to non-linear functions by Dakin (1965). For a detailed overview of some typical B-B algorithms, see Clausen (2003). For application of B-B in solving CCQP, see for example Bienstock (1995), Leyffer (2001) and Shaw et al. (2008).

4.1.2 Heuristic methods

Although the objective function of CCQP is convex, given the non-convex nature of the constraints and hence the overall problem, a number of heuristic methods have been proposed to deal with such problems. These methods generally fall under the class of adaptive stochastic optimization algorithms which include Genetic Algorithm, Tabu Search and Simulated Annealing, and have been used extensively to solve global optimization problems with arbitrary objective function and constraints. Essentially, these methods cast the CCQP as a generic global optimization problem, oblivious to the underlying structure of the objective function.

- ▷ Genetic Algorithms (GA) were introduced by Holland (1975). See, for example, Loraschi et al. (1995) for application of GA in portfolio optimization problems. An outline of the application of GA in solving CCQP consists of a four step process. Assuming the cardinality is set to K, the algorithm starts with a set of K asset portfolios, the *population* set, then loops over the following steps:
  - *Fitness function evaluation* - calculate the objective function values for portfolios in the current *population* set;
  - *Selection* - re-sample the population set with replacement and with probability being a function of values from the previous step;
  - *Crossover* - randomly mix assets from this newly created population set;
  - *Mutation* - randomly replace asset from previous step with assets not in the population.
- ▷ Tabu Search (TA) is a local search algorithm proposed by Glover (1986). It is similar to other greedy local search algorithms such as the *hill climbing* method (see for example Russell and Norvig (1995)) in that it uses a local search procedure to iteratively move from one solution to a better solution, until some stopping criterion has been satisfied. The main differentiating feature of TA is its maintaining a list of solutions that have been visited in the recent past and a list of prohibited moves that have certain attributes. See Chang et al. (2000) for application of TA in portfolio optimization.
- ▷ Simulated Annealing (SA) is another popular global optimization technique applied to optimization of non-convex problem in large discrete search space, originally proposed by Kirkpatrick (1984). Compared to greedy algorithms, where only *downhill* moves are allowed, SA

allows searches *uphill*. The probability of an uphill move is controlled by a global parameter, commonly referred to as the *temperature*, that is gradually decreased during the process. When *temperature* is high, moves can be in almost any random direction.

4.1.3 Linearization of the objective function

The *mean absolute deviation* (MAD) as opposed to the covariance matrix can be used to measure risk, as proposed by Konno and Yamazaki (1991), Speranza (1996) and Park et al. (1998), such that for cases where the returns have zero mean, we use $\mathbb{E}\left[|Rx|\right]$ instead of $x^\top \mathbb{E}\left[R^\top R\right] x$ for risk. Then the cardinality constrained problem can be formulated and solved via linear programming, with savings in computation. Depending on the structure of the problem, this simplification could potential lead to significant loss of information (Simaan, 1997).

Both *branch-and-bound* and heuristic methods are generic methods in the sense that the algorithm does not explicitly explore any specific characteristics of the underlying problem. As a result, these methods are often not the most efficient way to solve CCQP in a portfolio optimization setting. In the Sections (4.3) and (4.4), two new methods will be proposed to solve a CCQP. The *Global Smoothing* algorithm is an iterative method that first transforms CCQP to that of finding the global optimum of a problem in continuous variable, then solves a sequence of sub-problems. The *Local Relaxation* algorithm exploits the inherent structure of the objective function. It solves a sequence of small, local, quadratic-programs by first projecting asset returns onto a reduced metric space, followed by clustering in this space to identify sub-groups of assets that best accentuate a suitable measure of similarity amongst different assets. Since the *Global Smoothing* algorithm is closely related to the s*equential primal barrier QP method* (see for example Gill et al. (1981) for a more comprehensive treatment), Section (4.2) reviews this algorithm for solving a classic QP problem before moving on to the two proposed methods in dealing with the added cardinality constraint.

4.2 Sequential Primal Barrier QP Method

4.2.1 Framework

Null Space Reduction If we ignore cardinality constraint (4.3), then we are left with a simple linear inequality constrained QP (IQP), which can be solved by using a barrier method

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & F(x;\mu) \quad = c^\top x + \frac{1}{2}x^T H x - \mu \sum_i \log x_i \\ \text{s.t.} \quad & Ax = b. \end{aligned} \tag{4.4}$$

The gradient and Hessian are given by

$$\widetilde{g} = c + Hx - \mu X^{-1} e,\ \widetilde{H} = H + \mu X^{-2}$$

where $X^{-2} = \text{diag}\left(x_1^{-2}, x_2^{-2}, \ldots, x_n^{-2}\right)$. The KKT system is then given by

$$\begin{bmatrix} \widetilde{H} & A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ -\lambda^* \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix} \tag{4.5}$$

Note that we have assumed A has full row rank, since $m \leq n$, then the imposition of m linearly independent linear equality constraints on a problem with n variables can be viewed as reducing the dimensionality of the optimization to $n - m$. Let $Y \in \mathcal{R}\left(A^\top\right)$, i.e. Y is in the *range* of $A^\top$, and $Z \in \mathcal{N}(A)$, i.e. Z is in the *null space* of A, then (4.4) can be simplified to give

$$\min_{x_Z \in \mathbb{R}^{n-m}} x_Z^\top Z^\top (c + HYx_Y^*) + \frac{1}{2} x_Z^\top Z^\top HZx_Z - \mu \sum_{x_i \in x_Z} \log x_i \tag{4.6}$$

where the unique decomposition of the n-vector, x is given by

$$x = Yx_Y + Zx_Z.$$

Let $X_Z^{-2} = \operatorname{diag}\left(0, \ldots, 0, x_Z^{-2}\right)$. If $Z^\top \left(H + \mu X_Z^{*-2}\right) Z \succeq 0$, then the unique solution to (4.6), x_Z^*, must satisfy the equations

$$Z^\top HZx_Z^* = \underbrace{\mu X_Z^{-1} e}_{\text{term due to barrier}} - Z^\top (c + HYx_Y^*), \tag{4.7}$$

Note that due to the barrier term on the right hand side, the system in (4.7) is clearly nonlinear in x_z, so ruling out the possibility of using a simple *conjugate gradient* (CG) method.

Base on a similar analysis as the one above, an IQP can be posed as solving a sequence of unconstrained QP in a reduced space in $\mathbb{R}^{n-m}$, using a line research method. Let p denote the step to x^* from x, so that $p = x^* - x$. We can write (4.5) in terms of p to give

$$\widetilde{g} + \widetilde{H}p = A^\top \lambda^* \tag{4.8}$$

$$Ap = -v \tag{4.9}$$

where $-v = Ax - b$. Observe that p itself is the solution of an equality-constrained QP,

$$\begin{aligned} \min_{p \in \mathbb{R}^n} \quad & \widetilde{g}^\top p + \tfrac{1}{2} p^\top \widetilde{H} p \\ \text{s.t.} \quad & Ap = -v. \end{aligned}$$

Consider the partition $x = \begin{bmatrix} x_B & x_S \end{bmatrix}$ in m *basic* and $n-m$ *superbasic* variables, together with corresponding partition[12], $A = \begin{bmatrix} B & S \end{bmatrix}$ for some $B \in \mathbb{R}^{m \times m}$, $S \in \mathbb{R}^{m \times (n-m)}$, and $p = \begin{bmatrix} p_B & p_S \end{bmatrix}$ for some $p_B \in \mathbb{R}^{m \times 1}$, $p_S \in \mathbb{R}^{(n-m) \times 1}$, then (4.9) can be written as

$$Bp_B = v - Sp_S$$

[12] e.g. we could use the pivoting permutation, as part of the QP decomposition, to find the full rank square matrix B.

and we have

$$\begin{bmatrix} p_B \\ p_S \end{bmatrix} = \begin{bmatrix} -B^{-1}v - B^{-1}Sp_S \\ p_S \end{bmatrix} = -\begin{bmatrix} B^{-1} \\ 0 \end{bmatrix} v + \begin{bmatrix} -B^{-1}S \\ I \end{bmatrix} p_S.$$

Let

$$W = \begin{bmatrix} B^{-1} \\ 0 \end{bmatrix}, Z = \begin{bmatrix} -B^{-1}S \\ I \end{bmatrix}, Q^{-1} = \begin{bmatrix} B & S \\ 0 & I_{n-m} \end{bmatrix}.$$

Clearly we have the $n \times (n-m)$ matrix of *reduced-gradient basis*, $Z \in \mathcal{N}(A)$ and the product $AQ = \begin{bmatrix} I & 0 \end{bmatrix}$[13]. We have

$$p = Wp_B + Zp_S, \tag{4.10}$$

where the values of p_B and p_S can be readily determined by the following methods,

- ▷ p_B: by (4.9) and (4.10), we have $Ap = AWp_B = -v$ which simplifies to $p_B = -v$;
- ▷ p_S: multiply (4.8) by Z^T and using (4.10), we obtain

$$Z^\top \widetilde{H} Z p_S = -Z^\top \widetilde{g} - Z^\top \widetilde{H} Y p_B, \tag{4.11}$$

  which has a unique solution for p_S since $Z^\top \widetilde{H} Z \succ 0$.

If x is feasible (so that $v = 0$), then $p_B = 0$, $p \in \mathcal{N}(A)$ and (4.11) becomes

$$Z^\top \widetilde{H} Z p_S = -Z^\top \widetilde{g}. \tag{4.12}$$

In general, the reduced-gradient form of Z is not formed or stored explicitly. Instead, the search direction is computed using vectors of the form $Z\xi$, for some $\xi \in \mathbb{R}^{(n-m)\times 1}$, such that

$$Z\xi = \begin{bmatrix} -B^{-1}S \\ I \end{bmatrix} \xi = \begin{bmatrix} u_\xi \\ \xi \end{bmatrix}$$

where $Bu_\xi = -S\xi$, and can be solved using *LU-decomposition* of B, i.e. $u_\xi = -U^{-1}L^{-1}S\xi$. Similarly, $Z^\top \zeta$, for some n-vector $\zeta = \begin{bmatrix} \zeta_1 & \zeta_2 \end{bmatrix}^\top$, can be expressed as

$$Z^\top \xi = \begin{bmatrix} -B^{-1}S & I \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} = \begin{bmatrix} u_{\zeta_1} \\ \zeta_2 \end{bmatrix}$$

where $Bu_{\zeta_1} = -S\zeta_1$, and can be solved using *LU-decomposition* of B, i.e. $u_{\zeta_1} = -U^{-1}L^{-1}S\zeta_1$. Since these vectors are obtained by solving systems of equations that involve B and $B^\top$, thus Z may be represented using only a factorization (such as the *LU-decomposition*) of the $m \times m$ matrix B. The proposed algorithm is given in Algorithm 1.

[13] Note that $\widetilde{Y} \in \mathcal{N}(A)$ only if $Y^\top Z = 0$, e.g. when Q is orthogonal.

Algorithm 1 Null Space QP

[form permutation matrix] $P \leftarrow$ via pivoting scheme of the QR decomposition of matrix $\check{A}$, such that $\check{A}P = \widetilde{A} = \begin{bmatrix} \widetilde{B} & \widetilde{S} \end{bmatrix}$, where the full rank square matrix $\widetilde{B} \in \mathbb{R}^{m \times m}$ is simply given by the partition of $\widetilde{A}$.
partition of $\check{x}$ into m *basic* and $n - m$ *superbasic* variables
Given: μ_0, σ, max-outer-iteration, max-CG-iteration
start with feasible value x_0, which implies $p_B^{(0)} = 0$ and remain zero for all subsequent iterations
$k \leftarrow 1$
while $k \leq$ max-outer-iteration **do**
  $\widetilde{H}_k \leftarrow H + \mu^{(k)} X_k^{-2}$
  $\widetilde{g}_k \leftarrow c + Hx - \mu^{(k)} X_k^{-1} e$
  calculate $p_S^{(k)}$ based on (4.12), where $Z^\top \widetilde{H}_k Z p_S = -Z^\top \widetilde{g}_k$. We solve this linear system with Tuncated CG Algorithm 3 (setting maximum iteration to max-CG-iteration). Note that rather than forming Z's explicitly, we use stored *LU-decomposition* of B for the matrix-vector product Zp_S
  $p^{(k)} \leftarrow \begin{bmatrix} \mathbf{0} & p_S^{(k)} \end{bmatrix}$
  [calculate step-length]
  if $x^{(k)} + p^{(k)} > 0$ **then**
    $\alpha^{(k)} = 1$
  else
    $\alpha^{(k)} = 0.9$
  end if
  [new update]
  $x^{(k+1)} \leftarrow x^{(k)} + \alpha^{(k)} p^{(k)}$
  $k \leftarrow k + 1$
  if $\|\widetilde{g}_{k+1}\| - \|\widetilde{g}_k\| > -\epsilon$ **then**
    $\mu^{(k+1)} \leftarrow \sigma \mu^{(k)}$
  else
    $\mu^{(k+1)} \leftarrow \mu^{(k)}$;
  end if
end while
final result is given by $x^{(k^*)} P^\top$, where k^* is the terminating iteration

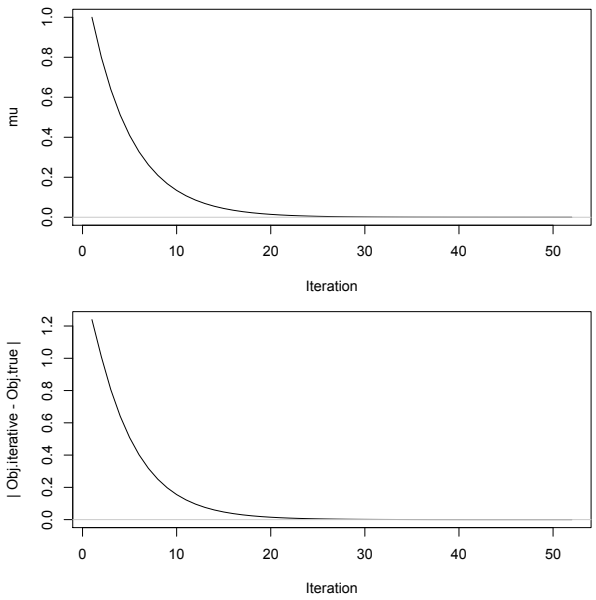

Figure 4.2: Null space line-search algorithm result. Top panel: value of barrier parameter, μ, as a function of iteration. Bottom panel: 1-norm of error. σ is set to 0.8 and initial feasible value, $x_0 = \begin{bmatrix} 0.3 & 0.5 & 0.2 & 0.7 & 0.3 \end{bmatrix}^\top$. Trajectory of the 1-norm error in the bottom panel illustrates that the algorithm stayed within the feasible region of the problem.

Small Scale Simulated Result Consider the case where

$$H = \begin{bmatrix} 6.97 & \bullet & \bullet & \bullet & \bullet \\ -0.34 & 8.04 & \bullet & \bullet & \bullet \\ -0.14 & -0.11 & 7.28 & \bullet & \bullet \\ -1.52 & -0.44 & -0.41 & 8.24 & \bullet \\ 1.70 & 0.17 & 0.30 & 0.07 & 5.11 \end{bmatrix}, \; c = \begin{bmatrix} 4.89 \\ 2.64 \\ 3.40 \\ 4.62 \\ 3.93 \end{bmatrix}$$

and

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \; b = \begin{bmatrix} 1.0 \\ 1.0 \\ 0.2 \end{bmatrix}.$$

By using the built-in QP solver in R, which implements the dual method of Goldfarb and Idnani (1982, 1983), we get the optimal solution $x^* = \begin{bmatrix} 0.688 & 0.11 & 0.20 & 0.31 & -0.00 \end{bmatrix}^\top$ and the corresponding objective function value of 7.61. Note the slight violation of the non-zero constraint in the last variable. Figure 4.2 shows the result of our line search algorithm, which converges to the same result strictly in the interior of the constrains.

4.2.2 Empirical Analysis

Data Empirical results in this section is based on the daily closing prices for 500 of the most liquid stocks, as measured by the median 21-day daily transacted volume, traded on the main stock exchanges in the US, spanning the period between 2008-01-22 to 2009-01-22. The setup of the mean-variance optimization problem is as follows,

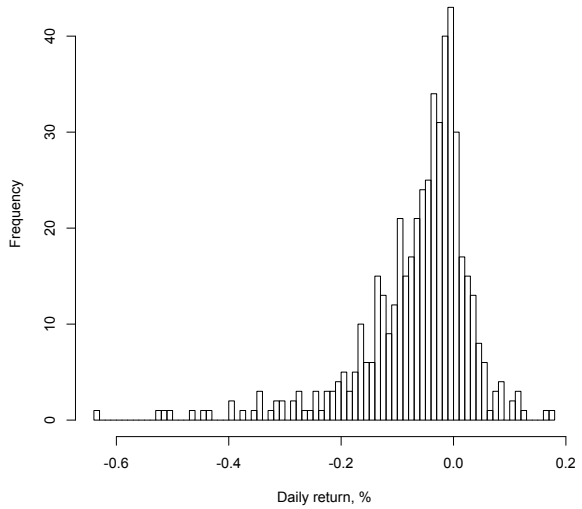

Figure 4.3: Histogram for mean daily return for 500 most actively traded stocks between 2008-01-22 to 2010-01-22

- ▷ H is the variance-covariance matrix for the log returns, based on open-to-close difference in log-prices.
- ▷ c is set to equal to the mean of log returns over the sample period, see Figure 4.3;
- ▷ a total of two *equality constraints* are set such that
  - weights for stocks with "A" in the first alphabet in their ticker name sum to $1/2$;
  - weights for stocks with "B" in the first alphabet in their ticker name sum to $1/2$;
- ▷ *inequality constraints* are set such that the vector of 500 weights are larger than -10, i.e. $x_i \geq -10 \, \forall i$.

Result Figure 4.4 shows the result using the built-in R routine that is based on the dual method of Goldfarb and Idnani (1982, 1983), for the cases with and without the inequality constraint. For the inequality constrained case, the number of iterations until convergence is set to 30, and the number of active constraints at solution is set to 29. Compare this with the result based on Algorithm 1, shown in Figure 4.5, we see that the two results are clearly not identical. This is an artifact of the iterative routines used in the underlying algorithms used by the two methods.

In terms of processing time, the built-in R function is faster than our algorithm. Figure 4.6 and Figure 4.7 shows the convergence diagnostics. Table 1 shows diagnostic output from Algorithm 1 at each outer iteration.

4.3 Solving CCQP with a Global Smoothing Algorithm

Murray and Ng (2008) proposes an algorithm for nonlinear optimization problems with discrete variables. This section extends that framework, by casting the problem in the portfolio optimization

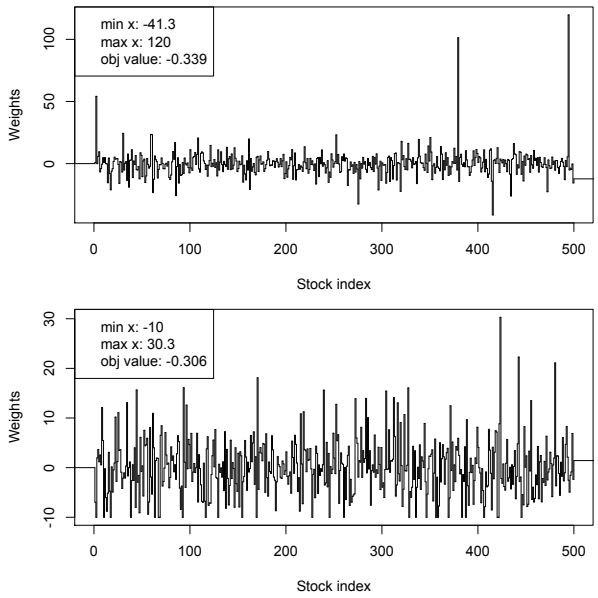

Figure 4.4: Optimization output using built-in R routine that is based on the dual method of Goldfarb and Idnani (1982, 1983). Top Panel: with equality constraint only. Bottom Panel: with equality and inequality constraints. $mix\ x$ and $max\ x$ are the minimum and maximum of the solution vector x, respectively.

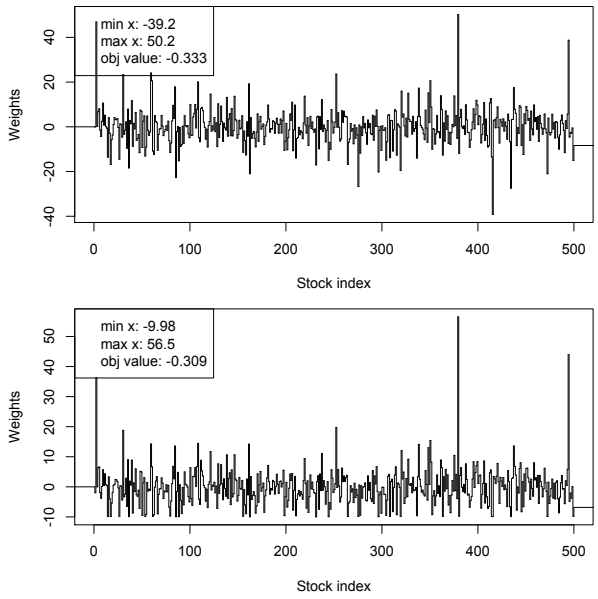

Figure 4.5: Optimization output using Algorithm 1. Top Panel: with equality constraint only. Bottom Panel: with equality and inequality constraints. $mix\ x$ and $max\ x$ are the minimum and maximum of the solution vector x, respectively.

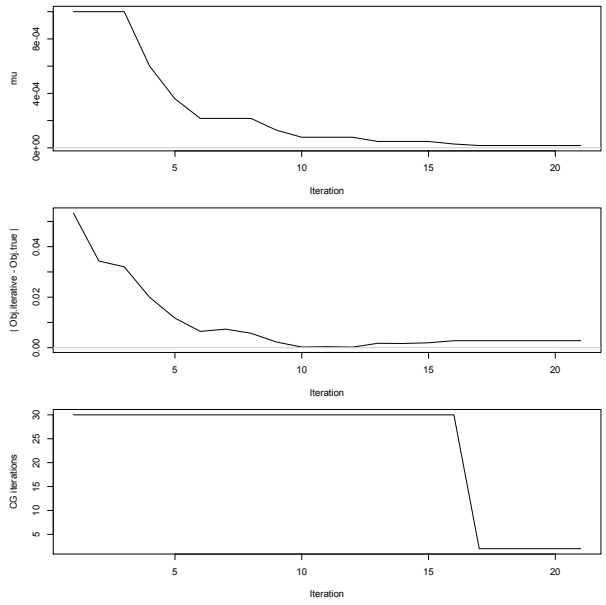

Figure 4.6: Algorithm 1 convergence diagnostics. Top panel: value of barrier parameter μ; Center panel: 1-norm of difference between objective value at each iteration and the true value (as defined to be the value calculated by the R function); Bottom panel: the number of CG iterations at each outer iteration.

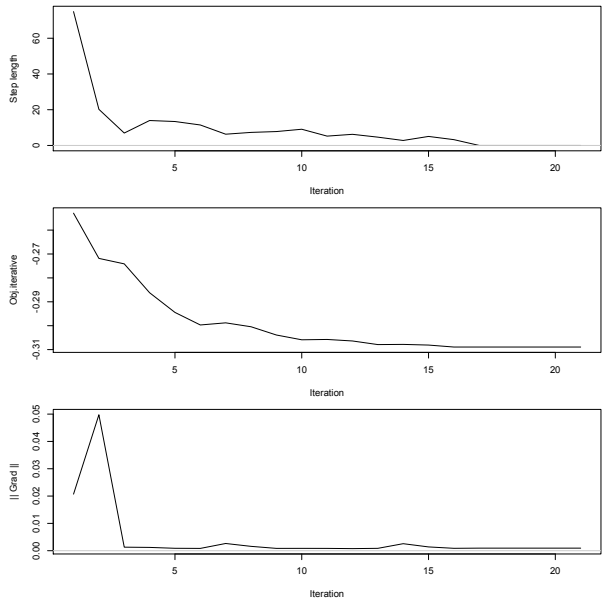

Figure 4.7: Additional Algorithm 1 convergence diagnostics. Top panel: step-length α; Center panel: objective function value; Bottom panel: 2-norm of the gradient.

	ngrad.r	ngrad	obj	cond	cond.r	mu	step	res.pre	res.post
1	2.12e-02	2.07e-02	-0.253	7.45e+03	7.81e+03	1.00e-03	74.90	6.21e-16	1.79e-15
2	4.75e-02	4.98e-02	-0.272	7.07e+03	9.92e+03	1.00e-03	20.22	1.79e-15	5.56e-15
3	1.26e-03	1.29e-03	-0.274	7.64e+03	1.45e+04	1.00e-03	6.95	5.56e-15	2.81e-15
4	1.15e-03	1.21e-03	-0.286	1.38e+04	2.40e+04	6.00e-04	13.96	2.81e-15	2.84e-15
5	8.32e-04	8.86e-04	-0.294	2.21e+04	7.06e+04	3.60e-04	13.40	2.84e-15	5.56e-15
6	7.82e-04	8.42e-04	-0.300	3.16e+04	1.61e+05	2.16e-04	11.46	5.56e-15	6.66e-15
7	2.59e-03	2.64e-03	-0.299	7.30e+04	3.29e+05	2.16e-04	6.30	6.66e-15	1.19e-14
8	1.49e-03	1.59e-03	-0.300	3.74e+04	1.92e+05	2.16e-04	7.31	1.19e-14	1.27e-14
9	7.28e-04	8.51e-04	-0.304	4.66e+04	2.96e+05	1.30e-04	7.78	1.27e-14	1.64e-14
10	6.71e-04	8.59e-04	-0.306	5.08e+04	3.49e+05	7.78e-05	9.08	1.64e-14	2.16e-14
11	5.42e-04	8.46e-04	-0.306	4.61e+04	3.92e+05	7.78e-05	5.23	2.16e-14	9.33e-15
12	4.03e-04	7.67e-04	-0.306	3.99e+04	3.99e+05	7.78e-05	6.23	9.33e-15	7.99e-15
13	6.17e-04	8.67e-04	-0.308	4.80e+04	4.75e+05	4.67e-05	4.65	7.99e-15	5.35e-15
14	2.45e-03	2.57e-03	-0.308	1.13e+05	1.62e+06	4.67e-05	2.79	5.35e-15	1.38e-14
15	1.17e-03	1.39e-03	-0.308	3.42e+04	5.16e+05	4.67e-05	5.05	1.38e-14	1.61e-14
16	5.85e-04	8.77e-04	-0.309	3.74e+04	5.22e+05	2.80e-05	3.25	1.61e-14	1.61e-14
17	6.72e-04	9.40e-04	-0.309	9.73e+04	1.48e+06	1.68e-05	0.00	1.61e-14	1.61e-14
18	6.72e-04	9.40e-04	-0.309	9.73e+04	1.48e+06	1.68e-05	0.00	1.61e-14	1.61e-14
19	6.72e-04	9.40e-04	-0.309	9.73e+04	1.48e+06	1.68e-05	0.00	1.61e-14	1.61e-14
20	6.72e-04	9.40e-04	-0.309	9.73e+04	1.48e+06	1.68e-05	0.00	1.61e-14	1.61e-14
21	6.72e-04	9.40e-04	-0.309	9.73e+04	1.48e+06	1.68e-05	0.00	1.61e-14	1.61e-14

Table 7: Nullspace CG algorithm output for the first 21 iterations. ngrad.r: norm of reduced gradient; ngrad: norm of gradient; obj: objective function value; cond: condition number of hessian; cond.r: condition number of reduced hessian; mu: barrier parameter; step: step length; res.pre: norm of residual pre CG; res.post: norm of residual post CG; maximum of outer iteration: 20; maximum of CG iteration: 30.

setting. There is an equivalence of the CCQP problem and that of finding the global minimizer of a problem with continuous variables (see for example Ge and Huang, 1989). However, it is insufficient to introduce only penalty function to enforce integrability as it risks introducing large numbers of local minimums which can significantly increase the complexity of the original problem. The idea of the global smoothing algorithm is to add a strictly convex function to the original objective, together with a suitable penalty function. The algorithm then iteratively increases the penalty on non-integrability and decreases the amount of smoothing introduced.

4.3.1 Framework

Here we add the cardinality constraint (4.3), and solve the following mixed integer optimization problem,

$$\min_{x} \quad f(x) = c^\top x + \tfrac{1}{2} x^\top H x \tag{4.13}$$

$$\text{s.t.} \qquad Ax = b \tag{4.14}$$

$$y_i \geq x_i \geq 0 \qquad i = 1, \ldots, n \tag{4.15}$$

$$\textstyle\sum_i y_i = K \qquad y_i \in \{0, 1\} \tag{4.16}$$

where $c, x \in \mathbb{R}^{n\times 1}$, $H \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^{m\times 1}$ and $m \leq n$. The transformed problem becomes:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & F(x,y) = f(x) + \underbrace{\frac{\gamma}{2}\sum_i y_i(1-y_i)}_{\text{penalty term}} - \underbrace{\mu\sum_i (\log s_i + \log x_i)}_{\text{smoothing term}} \\ \text{s.t.} \quad & \begin{bmatrix} A & 0 & 0 \\ 0 & e^\top & 0 \\ -I & I & -I \end{bmatrix} \begin{bmatrix} x \\ y \\ s \end{bmatrix} = \begin{bmatrix} b \\ K \\ 0 \end{bmatrix}, \end{aligned} \tag{4.17}$$

where $e \in \mathbb{R}^{n\times 1}$ is a column vector of one's, s is the slack variable, γ is the parameter that governs the degree of penalty on non-integrability, and μ is the parameter that governs the amount of global smoothing introduced into the problem. Let $\check{A} = \begin{bmatrix} A & 0 & 0 \\ 0 & e^\top & 0 \\ -I & I & -I \end{bmatrix}$, $\check{x} = \begin{bmatrix} x \\ y \\ s \end{bmatrix}$ and $\check{b} = \begin{bmatrix} b \\ K \\ 0 \end{bmatrix}$, then the gradient and Hessian can be expressed as,

$$\check{g} = \nabla F(x,y) = \begin{bmatrix} g_1 - \frac{\mu}{x_1} \\ \vdots \\ g_n - \frac{\mu}{x_n} \\ \frac{\gamma}{2}(1-2y_1) \\ \vdots \\ \frac{\gamma}{2}(1-2y_n) \\ -\frac{\mu}{s_1} \\ \vdots \\ -\frac{\mu}{s_n} \end{bmatrix}, \quad \check{H} = \nabla^2 F(x,y) = \begin{bmatrix} \check{H}_{11} & 0 & 0 \\ 0 & \check{H}_{22} & 0 \\ 0 & 0 & \check{H}_{33} \end{bmatrix}$$

where

$$g = c + Hx,\ \check{H}_{11} = H + \operatorname{diag}\left(\frac{\mu}{x_i^2}\right),\ \check{H}_{22} = \operatorname{diag}(-\gamma_i),\ \check{H}_{33} = \operatorname{diag}\left(\frac{\mu}{s_i^2}\right).$$

An outline of the global smoothing CCQP algorithm is given in Algorithm 2. Algorithm 3 is an adaptation of the conjugate gradient method that produces a descent search direction and guarantees that the fast convergence rate of the pure Newton method is preserved, provided that the step length is used whenever it satisfies the acceptance criteria. The CG iteration is terminated once a direction of negative curvature is obtained.

4.3.2 Empirical Analysis

Empirical results to illustrate an application of the *global smoothing algorithm* is based on the daily closing prices for 500 of the most liquid stocks, as measured by median 21-day daily transacted volume, traded on the main stock exchanges in the US, spanning the period between 2008-01-22 and 2009-01-22. Algorithm 2 illustrates an implementation of the framework to solve the following

Algorithm 2 Global Smoothing CCQP

[form permutation matrix] $P \leftarrow$ via pivoting scheme of the QR decomposition of matrix $\check{A}$, such that $\check{A}P = \widetilde{A} = \begin{bmatrix} \widetilde{B} & \widetilde{S} \end{bmatrix}$, where the full rank square matrix $\widetilde{B} \in \mathbb{R}^{m \times m}$ is simply given by the partition of $\widetilde{A}$.
partition of $\check{x}$ into m *basic* and $n - m$ *superbasic* variables
Given μ_0, σ, ν and γ_0
start with feasible value, $\check{x}_0$, which implies $p_B \equiv 0$ and remain zero for all subsequent iterations
$k \leftarrow 1$
while not-converged **do**
  $\widetilde{H}_k \leftarrow P^\top \check{H} P$ and $\widetilde{g}_k \leftarrow P^\top \check{g}$, where x_k is simply the first n elements of the vector $P\check{x}_k$ and y_k is the second n elements
  calculate p_S based on (4.12), where we solve the potentially indefinite Newton system $Z^\top \widetilde{H}_k Z p_S = -Z^\top \widetilde{g}_k$ based on either the Tuncated CG Algorithm 3, or the Modified Cholesky Factorization Algorithm 4
  $p_k \leftarrow \begin{bmatrix} \mathbf{0} & p_S \end{bmatrix}$
  calculate the maximum feasible step along α_M using the backtracking line search method in Algorithm 5
  $\check{x}_{k+1} \leftarrow \check{x}_k + \alpha_k p_k$
  $k \leftarrow k + 1$
  if $\|\widetilde{g}_{k+1}\| - \|\widetilde{g}_k\| > -\epsilon$ **then**
    $\mu_k \leftarrow \sigma \mu_k$
    $\gamma_{k+1} \leftarrow \gamma_k / \sigma^\nu$
  else
    $\mu_{k+1} \leftarrow \mu_k$ and $\gamma_{k+1} \leftarrow \gamma_k$
  end if
end while
[final result] $\check{x}_{k^*} P^\top$, where k^* is the terminating iteration

Algorithm 3 Truncated-CG

while not-reached-maximum-iteration **do**
  if $d_j^\top \widetilde{H}_k d_j \leq \xi$ **then**
    if if $j = 0$ **then**
      $p_k \leftarrow -\widetilde{g}_k$
    else
      $p_k \leftarrow z_j$
    end if
  end if
  $\alpha_j \leftarrow r_j^\top r_j / d_j^\top \widetilde{H}_k d_j$
  $z_{j+1} \leftarrow z_j + \alpha_j d_j$
  $r_{j+1} \leftarrow r_j + \alpha_j \widetilde{H}_k d_j$
  if $\|r_{j+1}\| < \epsilon_k$ **then**
    $p_k \leftarrow z_{j+1}$
  end if
  $\beta_{j+1} \leftarrow r_{j+1}^\top r_{j+1} / r_j^\top r_j$
  $d_{j+1} \leftarrow -r_{j+1} + \beta_{j+1} d_j$
end while

Algorithm 4 Modified Cholesky Factorization

Given: $\gamma \leftarrow \max(|a_{ii}| : i = 1, 2, \ldots, n)$, $\xi \leftarrow \min(a_{ii} : i = 1, 2, \ldots, n)$, ϵ_M the relative machine precision, where a_{ij} is the i, j element of the matrix A
$\beta \leftarrow \max(\gamma, \xi/\sqrt{n^2-1}, \epsilon_M)$, $\epsilon > 0$
$A \leftarrow \widetilde{H}_k$
for $l \leftarrow 1, 2, \ldots, n$ **do**
  $\mu_l \leftarrow \max\{|\hat{a}_{lj}| : j = l+1, l+2, \ldots, n\}$
  $r_{ll} \leftarrow \max\left\{\epsilon, \sqrt{|\hat{a}_{ll}|}, \mu_l/\beta\right\}$
  $e_{ll} \leftarrow r_{ll}^2 - \hat{a}_{ll}$
  for $j \leftarrow 1, 2, \ldots, n$ **do**
    $r_{lj} \leftarrow \hat{a}lj/r_{ll}$
  end for
  for $i \leftarrow 1, 2, \ldots, n$ **do**
    $\hat{a}_{ij} \leftarrow \hat{a}_{ij} - r_{lj} r_{li}$
  end for
end for

Algorithm 5 Backtracking Line Search (specialized to solve $1 \leq x \leq 0$

Given: $\bar{\alpha} > 0$, $\rho \in (0, 1)$, $c \in (0, 1)$, $\alpha \leftarrow \bar{\alpha}$
while $F(\check{x}_k + \alpha_k p_k) > F(\check{x}_k) + c\alpha \widetilde{g}_k^\top p_k$ **do**
  $\alpha \leftarrow \rho\alpha$
end while
$\alpha_k \leftarrow \alpha$

specific problem,

$$\begin{aligned} \min_{x,y} \quad & f(x) = c^\top x + \tfrac{1}{2} x^\top H x \\ \text{s.t.} \quad & y_i - x_i = s_i \qquad i = 1, \ldots, n \\ & \textstyle\sum_i y_i = K \qquad y_i \in \{0, 1\} \\ & s_i \geq 0 \\ & x \geq 0 \end{aligned}$$

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad F(x, y) \quad &= f(x) + \frac{\gamma}{2} \sum_i y_i (1 - y_i) \\ &\quad -\mu \sum_i (\log y_i + \log(1 - y_i) + \log s_i + \log x_i) \\ \text{s.t.} \quad & \begin{bmatrix} 0 & e^\top & 0 \\ -I & I & -I \end{bmatrix} \begin{bmatrix} x \\ y \\ s \end{bmatrix} = \begin{bmatrix} K \\ 0 \end{bmatrix}. \end{aligned}$$

Figure 4.8 shows the converged output for $K = 250$ and Figure 4.9 for $K = 100$.

While the proposed algorithm is able to solve small to medium scale problems, convergence is often sensitive to how γ and μ are modified at each iteration. Further research is needed to control these parameters to ensure consistent convergence performance and to generalize the algorithm to deal with more general constraints. Therefore, we treat this proposed algorithm as a prototype and a proof of concept that such a framework can potentially be adapted to solve a CCQP portfolio

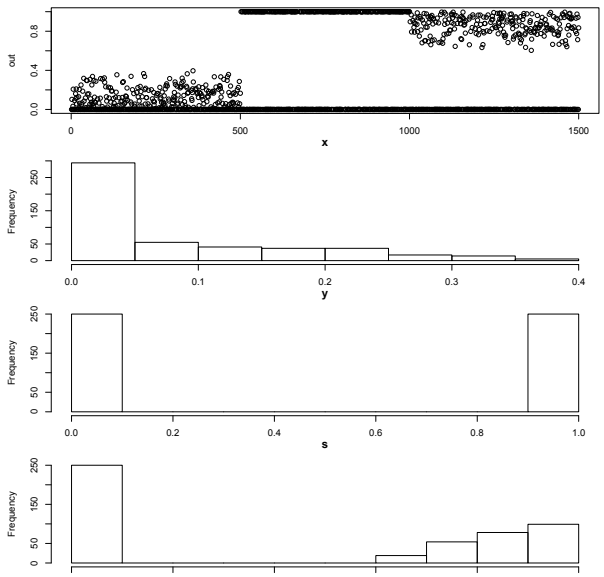

Figure 4.8: Converged output for $K = 250$. Panel-1: first 500 variables are x_i, second 500 are y_i and the last 500 are s_i; Panel-2: histogram of x_i; Panel-3: histogram of y_i; Panel-4: histogram of s_i.

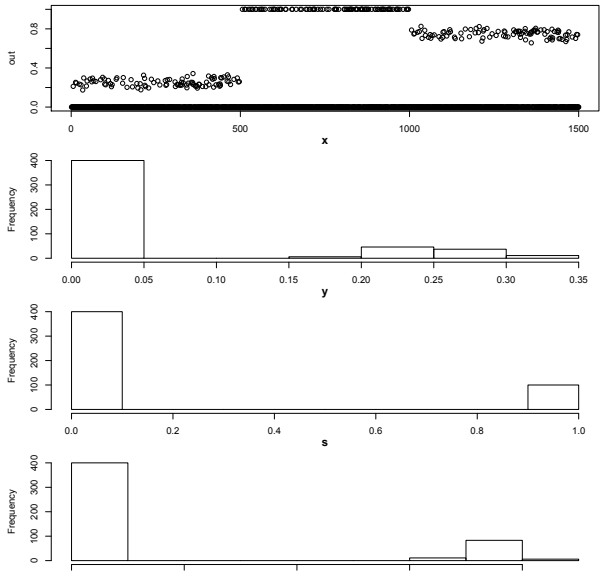

Figure 4.9: Converged output for $K = 100$. Panel-1: first 500 variables are x_i, second 500 are y_i and the last 500 are s_i; Panel-2: histogram of x_i; Panel-3: histogram of y_i; Panel-4: histogram of s_i.

problem. In the next section, we proposed another approach to solve the CCQP, which is both robust and scales well to the size of the underlying problem.

4.4 Solving CCQP with Local Relaxation Approach

Without loss of generality, we replace the generic constraint $Ax \geq b$ in (4.2) by an equality constraint $\sum_i x_i = 0$, so that it mimics the case of a *dollar neutral* portfolio. The resulting CCQP problem is given by

$$\begin{aligned} \min_x \quad & f(x) = -c^\top x + \lambda x^\top H x \\ \text{s.t.} \quad & \textstyle\sum_i x_i = 0 \\ & \textstyle\sum_i \mathbf{1}_{\{x_i \neq 0\}} = K. \end{aligned}$$

Murray and Shek (2011) proposes an algorithm that explores the inherent similarity of asset returns, and relies on solving a series of relaxed problems in a reduced dimensional space by first projecting and clustering returns in this space. This approach mimics line-search and trust-region methods for solving continuous problems, and it uses local approximation to the problem to determine an improved estimate of the solution at each iteration. Murray and Shanbhag (2007) and Murray and Shanbhag (2006) have used one variation of local relaxation approach to solve a grid-based electrical substation siting problem, and have demonstrated that the growth in effort with the number of integers is slow with their proposed algorithm vis-à-vis an exponential growth for a number of commercial solvers. The algorithm proposed here is different to their algorithm in a number of key areas. One, instead of a predefined 2-dimensional grid we project asset returns onto a multi-dimensional space based on exploring statistical properties of historic returns. Second, instead of using physical distance, we define the distance metric as a function of factor loading, risk aversion and expected return. Third, given the typical cardinality relative to the size of the search space and cluster group size, we cannot assume that the clusters do not overlap, hence we have proposed a necessary probabilistic procedure to assign cluster members to each centroid at each iteration.

4.4.1 Clustering

Recall that the goal of minimizing portfolio risk is closely related to asset variances and their correlation - our objective is to maximize expected return of the portfolio, while penalizing its variance. It is intuitive to think that assets from different sectors exhibit different correlation dynamics. However, empirical evidence seems to suggest this dynamic is weak relative to idiosyncratic dynamic of stocks both within and across sectors. Figure 4.10 shows the ordered covariance matrix for 50 assets, 5 from each of the 10 sectors defined by Bloomberg, such that the first group of stocks (JCP, CBS, ..., COH) belong to *Consumer Discretionary*, the second group of stocks (NHZ, KO, ..., GIS) belong to *Consumer Staples*, etc. We see that the covariance matrix does not seem to show any noticeable block structure that would suggest collective sectoral behavior. Covariance matrices calculated using different sample periods also confirm this observation.

Instead of relying on sector definition to cluster assets, we could use historic correlation of returns as a guide to help us identify assets that behave similarly. First we define a distance metric, d_{ij} for

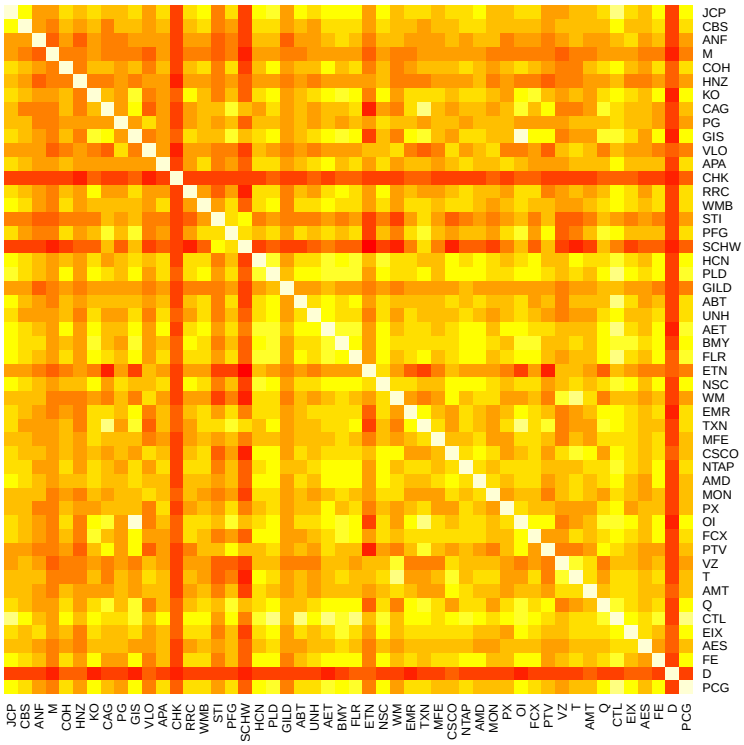

Figure 4.10: Heatmap of covariance matrix for 50 actively traded stocks, over the sample period 2008-01-10 to 2009-01-01. Darker colors indicate correlation closer to 1 and light colors closer to -1.

a pair of assets i and j,

$$d_{ij} = \sqrt{2\,(1 - \rho_{ij})}, \tag{4.18}$$

where ρ_{ij} is the Pearson's correlation coefficient based on the pairwise price return time series. This definition penalizes negatively correlated stocks more than stocks that are not correlated, which makes sense as our goal is to group stocks with similar characteristics as measured by the co-movement of their returns. Using this definition of distance metric, we can construct a *minimum spanning tree* (MST) (see for example Russell and Norvig (1995)) that spans our asset pairs with edge lengths given by (4.18). From Figure 4.11, we see that this method has identified some cluster structures, and verified that the dominant dimension is not along sector classification.

The MST is a two dimensional projection of the covariance matrix. An alternative, higher dimensional, method is to cluster the assets in a suitably chosen orthogonal factor space spanned by a more comprehensive set of spanning basis. Given a return matrix, $R \in \mathbb{R}^{T\times N}$, we aim to effectively project our universe of N asset returns onto an orthogonal k-dimensional space, where often $k \ll N$. We can write

$$R = P_k V_k^\top + U,$$

where $P_k \in \mathbb{R}^{T\times k}$ is the return of our k factors, $V_k \in \mathbb{R}^{N\times k}$ is the factor loadings and $U \in \mathbb{R}^{T\times N}$ is the matrix of specific returns. This reduces the dimension of our problem from N to k, such that the i-th column of the return matrix R, $r_i \in \mathbb{R}^T$, can be written as a linear combination of the k factors

$$r_i = v_{1,i}p_1 + v_{2,i}p_2 + \cdots + v_{k,i}p_k + u_i,$$

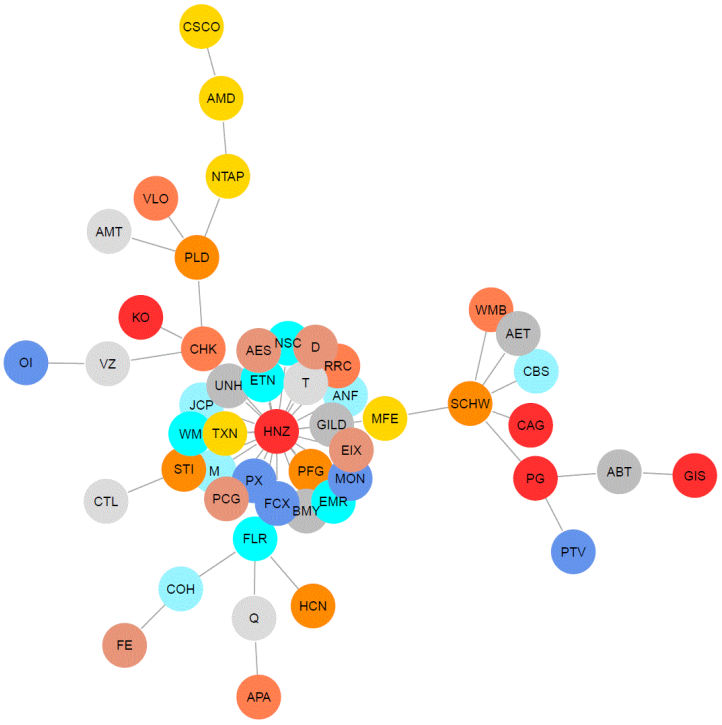

Figure 4.11: A *Minimum spanning tree* for 50 actively traded stocks, over the sample period 2008-01-10 to 2009-01-01. Using distance metric defined in (4.18). Each color identifies a different sector.

where $v_{i,j} \in \mathbb{R}$ is the (i,j) component of the matrix V_k, p_i and u_i are the i-th column of the matrix P and U respectively. One method to identify the k factors is by *principal component analysis* (PCA). An outline of the procedure can be expressed as follows,

- ▷ obtain spectral-decomposition[14]: $\frac{1}{T}R^\top R = V\Lambda V^\top$, where V and Λ are ordered by magnitude of the eigenvalues along the diagonal of Λ,
- ▷ form principal component matrix: $P = RV$, and
- ▷ finally take first k columns of P and V to give P_k and V_k , known as the principle component and factor loading matrix, respectively.

Once we have identified the k-dimensional space, then the proposed clustering method works as follows:

- ▷ define a cluster distance metric $d(r_i, r_j) = \|v_i - v_j\|_2$, where $v_i \in \mathbb{R}^k$ and $v_j \in \mathbb{R}^k$ are the i-th and j-th row of the factor loading matrix V_k.
- ▷ filter outliers in this space by dropping assets with distance significantly greater than the median distance from the *center-of-gravity*[15] of the k-dimensional space;

[14] Here we assume columns of R have zero mean.

[15] The *center-of-gravity* is defined in the usual sense by assigning a weight of one to each asset in the projected space, then the center is defined to be the position where a point mass, equal to the sum of all weights, that balances the levered weight of the assets.

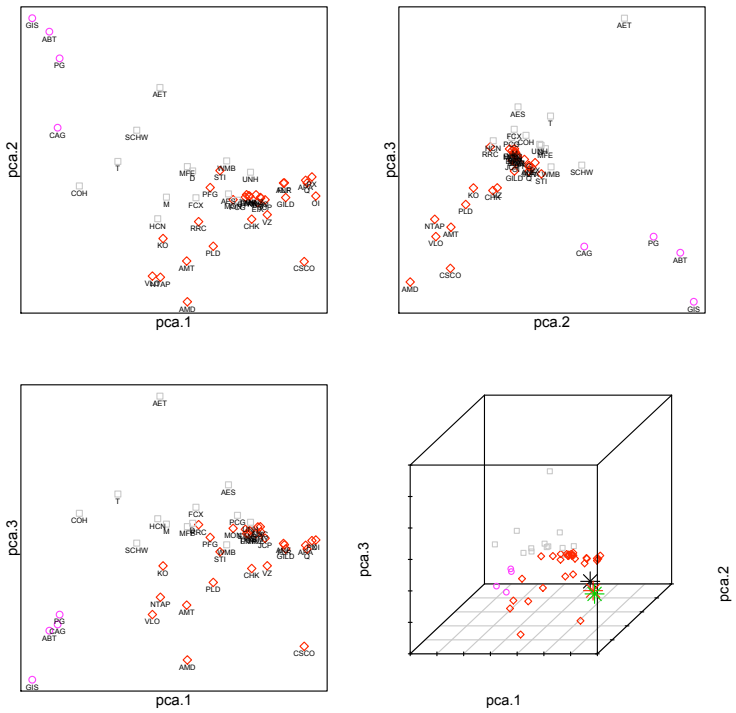

Figure 4.12: *k-means clustering* in a space spanned by the first three principal components based on log returns for 50 actively traded stocks, over the sample period 2008-01-10 to 2009-01-01.

- ▷ identify clusters (using for example *k-means clustering;* see Hastie et al. (2003) for details) in this metric space.

Figure 4.12 illustrates the proposed clustering method using the returns of 50 actively traded US stocks. Here we have set the number of clusters to three and have dropped four outliers stocks (FE, CBS, CTL and PTV) based on the criteria listed above, and then projected the remaining 46 returns onto a three-dimensional space spanned by the three dominant principal components.

Distance Measure We define the distance measures for an assets, s, in the projected space relative to an arbitrary reference point, a_0, by

$$d_m(s; a_o) = \sigma_d \|v_s - a_0\|_2 + (1 - \sigma_d)(\|c\|_\infty - |c_s|), \tag{4.19}$$

where c_s is the s-th element of the vector of expected return, c. The economic interpretation is that we use the tuning parameter σ_d to control the trade off between picking stocks highly correlated and stocks that have higher expected return. Note that $\sigma_d = 1$ corresponds to the simple Euclidean distance measure.

Identify Cluster Member Once the initial starting K centroid assets have been identified, the proposed algorithm identifies the member assets by the *center-of-gravity* method, defined for the

i-th cluster as

$$g_i = \frac{V_k^\top x_r^{*(i)}}{\left\| x_r^{*(i)} \right\|}, \tag{4.20}$$

where the vector $x_r^{*(i)}$ corresponds to optimal weights (see *relaxed-neighborhood QP* in Section 4.4.2) of assets in the i-th cluster. The member assets are picked according to the degree of closeness to centroid i, defined by the distance measure $d_m(s; g_i)$ for asset s.

Once the corresponding member assets have been identified for each cluster, the algorithm will propose a new set of K centroids at the end of each iteration. For the next iteration, new member assets needed to be calculated for each cluster. The composition of each new cluster clearly depends on the order that the centroids are picked. For example, clusters A and B could be sufficiently close in the projected space, such that there exists a group of assets which are equally close to both clusters. Since assets cannot belong to more than one group, so once cluster A has claimed an asset, the asset drops out of the feasible set of assets for cluster B.

Let p_i be the probability of centroid asset i being picked for cluster assignment, we propose the following logistic transform to parametrize this probability:

$$\log \frac{p_i}{1 - p_i} = \alpha_s \left(\frac{\left| x_{o,i}^* \right|}{\sum_i \left| x_{o,i}^* \right|} - \frac{1}{K} \right), \tag{4.21}$$

where $x_{o,i}^*$ corresponds to the optimal weight (see *centroid-asset QP* in Section 4.4.2) of the i-th centroid asset, such that the higher the centroid weight the greater the probability that it will have a higher priority over the other centroids in claiming its member assets.

4.4.2 Algorithms

Algorithm for identifying an initial set of starting centroids, $\mathcal{S}_o$ There are a number of possible ways to prescribe the initial set of centroid assets.

▷ *k-means* (Algorithm 6)

- form K clusters, where K is the cardinality of the problem, in the projected factor space spanned by $\mathcal{R}(P_k)$. Then in the projected space, identify the set of K *centroid-assets*, $\mathcal{S}_o \subset \mathcal{S}$, which are stocks closest to the cluster centers (note a cluster center is simply a point in the projected space, and does not necessarily coincide with any projected asset returns in this space);

▷ Successive truncation (Algorithm 7)

- at each iteration we essentially discards a portion of assets which have small weights, until we are left with the appropriate number of assets in our portfolio equal to the desired cardinality value.

Algorithm 6 K-means in factor space

Given: R, K
$\Omega \leftarrow \frac{1}{T} R^\top R$
[Singluar Value Decomposition] $\Omega = V \Lambda V^\top$
$V_k \leftarrow$ first k columns of V
[Random initialization] $m_1^{(0)}, m_2^{(0)}, ..., m_K^{(0)}$
while not-converged **do**
  for $i = 1$ to K **do**
    E-Step: $\mathcal{C}_i^{(n)} = \left\{ v_j : \left\| v_j - m_i^{(n)} \right\| \leq \left\| v_j - m_{\tilde{i}}^{(n)} \right\|, \forall \tilde{i} = 1, \ldots, K \right\}$
    M-Step: $m_i^{(n+1)} = \frac{1}{\left| \mathcal{C}_i^{(n)} \right|} \sum_{v_j \in \mathcal{C}_i^{(n)}} v_j$
  end for
end while

Algorithm 7 Successive truncation

Given: N, I
$j \leftarrow 0$
$\mathcal{S}_0 \leftarrow \mathcal{S}$; $n_0 \leftarrow N$; $\phi_0 \leftarrow 0$
repeat
  [Solve QP] for x^* where

$$\begin{aligned} x_j^* = \arg\min_x \quad & -c^\top x + \tfrac{1}{2} x^\top H x \\ \text{s.t.} \quad & e^\top x = 0 \\ & x \in \mathcal{S}_j \end{aligned}$$

  if arithmetic truncation **then**
    return $\phi_j \leftarrow \max\left(n_j - \lfloor N/I \rfloor, K \right)$
  else
    return $\phi_j \leftarrow \max\left(\lfloor (1 - K/N) \, n_j \rfloor, K \right)$
  end if
  $\mathcal{S}_{j+1} \leftarrow \left\{ s_{(n)}, s_{(n-1)}, \ldots, s_{(\phi_j)} \right\}$
  $n_{j+1} = |\mathcal{S}_{j+1}| - \phi_j$
until $\phi_j = K$

Algorithm for iterative local relaxation In the main Algorithm 8, at each iteration, we solve a number of small QPs for subsets of the original set of assets, $\mathcal{S}_o \subset \mathcal{S}_r \subseteq \mathcal{S}$, where $\mathcal{S}_r$ is the set of assets belonging to one of the K clusters identified. The QPs that we solve in the main algorithm include the following:

▷ *Globally-relaxed QP*

$$\begin{aligned} \min_{x} \quad & -c^\top x + \tfrac{1}{2} x^\top H x \\ \text{s.t.} \quad & e^\top x = 0 \\ & x \in \mathcal{S}, \end{aligned}$$

where c and H are the expected return and covariance matrix for the set of assets, $\mathcal{S}$; e is a column vector of 1's.

▷ *Centroid-asset QP*

$$\begin{aligned} \min_{x} \quad & -c_o^\top x_o + \tfrac{1}{2} x_o^\top H_o x_o \\ \text{s.t.} \quad & e^\top x_o = 0 \\ & x \in \mathcal{S}_o, \end{aligned}$$

where c_o and H_o are expected return and covariance matrix for the set of centroid assets, $\mathcal{S}_o$, only. This gives optimal weight vector x_o^*, where $x_{o,i}^*$ is the i-th element of this vector corresponding to i-th centroid asset, $s_{o,i}$.

▷ *Relaxed-neighborhood QP*

$$\begin{aligned} \min_{w} \quad & -c_r^\top x_r + \tfrac{1}{2} x_r^\top H_r x_r \\ \text{s.t.} \quad & e_i^\top x_r = x_{o,i}^* \qquad i = 1, \ldots, K, \\ & x_r \in \mathcal{S}_r, \end{aligned}$$

where e_i is a column vector with 1's at positions that correspond to the i-th centroid cluster assets and are zero everywhere else; c_r is the vector of expected returns and H_r is the corresponding variance-covariance matrix of our sub-universe of assets, $\mathcal{S}_r = \cup_{i=1}^{K} N(s_{o,i})$, where $N(s_{o,i})$ is the set of *neighbor* assets belonging to the cluster with $s_{o,i}$ as the centroid. We impose the constraint that the sum of weights of the neighbors of the i-th centroid add up to the centroid weight, $x_{o,i}^*$.

Algorithm for identifying new centroids Let $\{m_k\}_{k=1,\ldots,K}$ be the number of members in each of the K clusters. We generate $\|m\|_\infty$ number of centroid proposals based on assets within the same cluster to give $\Phi_1, \ldots, \Phi_{\|m\|_\infty}$. $\Phi_1 \in \mathbb{R}^K$ is a vector of the set of closest distances, defined by (4.19), to the centroid; Φ_2 is the vector of distances that are second closest, etc. For cluster group k with $m_k \leq \|m\|_\infty$, we pad the last $\|m\|_\infty - m_k$ entries with distance of the cluster member that is farthest away from the centroid in cluster k. Once we have generated the centroid proposals $\{\Phi_i\}_{i=1,\ldots,\|m\|_\infty}$, Algorithm 9 then loops over each of the proposals. During each loop, the algorithm replaces the

Algorithm 8 Local relaxation in projected orthogonal space

Given: α_s, σ_d, M_{min}, M_{max}, K, $\bar{\eta}$, ϵ_o, $\mathcal{S}$
$x^* \leftarrow$ solution of *globally-relaxed* QP
$\Xi \leftarrow$ ranked vector x^*
$\mathcal{S}_o \leftarrow$ initial centroids by Algorithm 7 or by Algorithm 6
while not-converged **do**
  $(x_o^*, f_o^*) \leftarrow$ solution of *centroid-asset* QP
  $\grave{\Xi} \leftarrow \Xi$
  while $I_\varepsilon = \left\{i : \left|x_{o,i}^*\right| < \epsilon_o\right\} \neq \emptyset$ **do**
    for each $i \in I_\varepsilon$ **do**
      $s \leftarrow \grave{s}_1$, where $s = \left\{s_{o,i} \in \mathcal{S}_o : \left|x_{o,i}^*\right| < \epsilon_o\right\}$, and $\grave{s} \in \grave{\Xi} \setminus \mathcal{S}_o$.
      $\grave{\Xi} \leftarrow \grave{\Xi} \setminus \{\grave{s}\}$
    end for
    $x_o^* \leftarrow$ solution of *centroid-asset* QP
  end while
  [Generate *random-centroid-asset-order*, $I_{\tilde{o}}$] Sampling, without replacement, from the index set $I_o = \{1, 2, \ldots, K\}$ with probability in (4.21)

$$p_i = \left(1 + \exp\left(\alpha_s \left(\frac{\left|x_{o,i}^*\right|}{\sum_i \left|x_{o,i}^*\right|} - \frac{1}{K}\right)\right)\right)^{-1}$$

  $\grave{\mathcal{S}} \leftarrow \mathcal{S} \setminus \mathcal{S}_o$
  for $\{s_{o,i} : s_{o,i} \in \mathcal{S}_o, i \in I_{\tilde{o}}\}$ **do**
    $m_i \leftarrow \max\left(\lfloor M_{max} p_i \rfloor, M_{min}\right)$
    [Identify m_i neighbors of $s_{o,i}$, $N\left(s_{o,i}\right) \subset \grave{\mathcal{S}}$], such that

$$\begin{aligned} \left\{s \in \grave{\mathcal{S}} : \left(\sigma_d \left\|v_s - v_{s_{o,i}}\right\|_2 + (1 - \sigma_d)\left(\|c\|_\infty - |c_s|\right)\right) < r\right\} &\subset N\left(s_{o,i}\right) \\ \left|N\left(s_{o,i}\right)\right| &= m_i - 1 \end{aligned}$$

    $\grave{\mathcal{S}} \leftarrow \grave{\mathcal{S}} \setminus N\left(s_{o,i}\right)$,
  end for
  $\mathcal{S}_r \leftarrow \bigcup_{i=1}^{K} N\left(s_{o,i}\right)$
  $x_r^* \leftarrow$ solution of *relaxed-neighborhood* QP
  $\mathcal{S}_0 \leftarrow$ output of Algorithm 9
  if [max-iteration-reached OR max-time-limit-reached] **then**
    break
  end if
end while

Algorithm 9 Identify new centroids

Given: Υ, $\{g_i\}_{i=1,\dots,K}$
[Generate centroid proposal list]
for $i = 1$ to $\|m\|_\infty$ **do**
  $\Phi_i \leftarrow$ proposal of K centroid assets with distance ranked i-th closeset based on (4.19)
end for
$j \leftarrow 1$; $\eta \leftarrow 1$; *exit-loop*←**false**
while *exit-loop* is **false do**
  for $i = 1$ to K **do**
    $s_{o,i} \leftarrow \Phi_{j,i}$
    $f \leftarrow$ objective function value of *centroid-asset* QP
    if $f < f_o^*$ **then**
      exit-loop←**true**
      break
    end if
  end for
  $j \leftarrow j + 1$
  if $j > \|m\|_\infty$ **then**
    if $\eta \geq \bar{\eta}$ **then**
      goto Algorithm 10
    else
      $M_{max} \leftarrow \lfloor \Upsilon^\eta M_{max} \rfloor$
      $\eta \leftarrow \eta + 1$
    end if
  end if
end while

current centroids with progressively less optimal proposals, in the sense of larger distance measure from the cluster centroid assets, and breaks out as soon as there has been an improvement to the objective function.

If none of the proposals gives a lower objective function value, the algorithm then attempts to swap centroids with assets from other clusters. The choice of substitution is governed by the expected returns of the assets, weighted by the magnitude of the relaxed solution from *relaxed-neighborhood* QP, to give a vector $\varpi \in \mathbb{R}^{\sum_i m_i}$,

$$\varpi = \left[\left| x_{o,1}^* \right| c_{s \in N(s_{o,1})}, \dots, \left| x_{o,2}^* \right| c_{s \in N(s_{o,2})}, \dots, \dots, \left| x_{o,K}^* \right| c_{s \in N(s_{o,K})} \right]^\top,$$

where $c_{s \in N(s_{o,k})}$ is a row vector consists of the expected returns for assets belonging to the k-th cluster.

4.4.3 Computational Result

Preparation of Dataset Empirical results in this section are based on two key datasets. The first dataset consists of daily closing prices for 500 of the most liquid stocks, as measure by mean daily transacted volume, traded on the main stock exchanges in the US, spanning the period between 2002-01-02 to 2010-01-22. The second dataset consists of 3,000 monthly closing prices for actively traded stocks in the US, spanning the period between 2005-05-31 to 2010-04-30. These two datasets test the performance of our algorithm under different market environments and for different scales

Algorithm 10 Swap centroids

```
k ← 1; q ← ‖S_r‖
while exit-loop is false do
  s_{o,k} ← ϖ_q
  f ← objective function value of centroid-asset QP
  if f < f*_o then
    η ← 0
    exit-loop←true
  end if
  k ← k + 1
  if k = K then
    q ← q − 1
    k ← 1
    if q ≤ 0 then
      exit-loop←true
    end if
  end if
end while
```

of the problem. The results based on the proposed algorithm are compared with those obtained by the successive truncation algorithm in Algorithm 7 and by the branch-and-cut MIQP solver used in CPLEX, a leading commercial optimization software.

Cleaning and Filtering We follow a three step procedure to clean, filter and prepare the datasets for analysis:

- ▷ *Recorded price validation.* We discard assets that do not have a complete price history over the whole sample period;
- ▷ *NAICS sector matching.* We purge any asset whose ticker cannot be matched to the latest *North American Industry Classification System* (NAICS) sector definition. This is in place to facilitate potentially imposing sectoral constraints.
- ▷ *PCA outliers detection.* We project the returns, defined as the difference in log-prices, onto a 4-dimensional space spanned by the dominant PCA factors. Any assets with projected return more than two standard deviations away from the projected median is considered an outlier asset and dropped.

Factorization of Covariance Matrix For the dataset with 3,000 assets, since $N \gg T$ so the covariance matrix from raw returns is rank deficient. We take the necessary step of factorizing the raw covariance matrix by using the first four dominant PCA factors. This ensures that the resulting covariance matrix is both full rank and well conditioned. The first four factors together capture 35.07% of the total variance for the 500 stock case, and 38.15% for the 3,000 stocks case.

4.4.4 Algorithm Benchmarking

Successive Truncation Figure 4.13 shows the result for the fully relaxed problem (i.e. without the cardinality constraint), ordered by value of the optimal weights. Positive weights mean that we

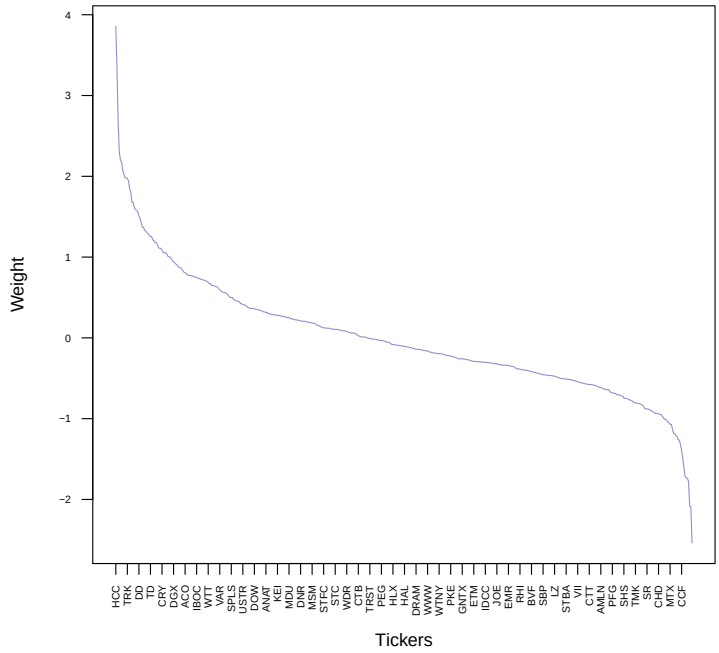

Figure 4.13: Fully relaxed (i.e. without cardinality constraint) QP solution for 500 actively traded US stocks (not all tickers are shown) sampled over the period from 2002-01-02 to 2010-01-22, with dollar neutral constraint.

go *long* those assets and negative weights correspond to *shorting* those assets. As we impose the constraint, $\sum_i x_i = 0$, so the net *dollar exposure* of the result portfolio is zero. This is one interpretation of a *market neutral* portfolio. An alternative definition of market neutrality requires neutralizing specific factors of the net portfolio exposure, by effectively taking into account correlation between different assets.

We apply the successive arithmetic truncation algorithm given in Algorithm 7 to CCQP. Table 8 gives the result for the 500-choose-15 (i.e. solving a 500 asset problem with cardinality equal to 15) case. It can be seen that if the number of iterations is set to one-tenth of the size of the universe, i.e. by discarding a large portion of stocks from the feasible set at each iteration, the method is able to obtain a result quickly. However, as we will show later, if we need a higher degree of optimality, this simple method often fails to deliver. Also, note that it is not guaranteed that the more iterations we use, the better the solution. Table 9 shows the result for the 3,000-choose-15 case, we see that the algorithm does not give monotone improvement of the objective function. For comparison, Table 10 shows the result for geometric truncation for the 3,000-choose-15 case. Again, we notice that increasing the number of iterations does not necessarily increase optimality.

Both of these truncation methods are widely used in practice. Although there is clearly no dispute on the speed of this heuristic algorithm, we later show that it can be significantly inferior than algorithms that are able to explore the solution space more effectively, at only a small cost of execution time.

Throughout the rest of this paper, we refer to successive arithmetic truncation as simple succes-

Number of Iterations	Objective Value	Time (sec)	Optimal Assets
10	-0.00615	0.246	AGII BIG CPO ETH FBP FICO HCC IGT LPNT PFG RYN SRT TAP UIS WPP
20	-0.00637	0.455	AGII BWA CDE CPO DYN ETH FICO HCC NWBI PFG RYN SRT TAP UIS WPP
50	-0.00629	0.832	AGII BWA CDE CPO EMC ETH FICO HCC IGT PFG RYN SRT TAP UIS WPP
100	-0.00635	1.677	AGII BWA CDE CPO DYN EMC ETH FICO HCC PFG RYN SRT TAP UIS WPP
500	-0.00638	8.073	AGII BWA CDE CPO ETH FICO HCC IGT PFG RYN SRT TAP TRK UIS WPP

Table 8: Successive arithmetic truncation method result for 500 of the most liquid US stocks, with cardinality, $K = 15$.

Number of Iterations	Objective Value	Time (sec)	Optimal Assets
10	-12.08450	33	AOL AONE ART CFL CFN CIE CIT CVE DGI DOLE EDMC MJN SEM TLCR VRSK
100	-12.13756	210	AOL AONE ART CFL CFN CIE CIT CVE DGI DOLE MJN SEM TLCR TMH VRSK
300	-12.23406	561	AOL AONE ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK SEM TLCR VRSK
500	-12.23406	929	AOL AONE ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK SEM TLCR VRSK
1,000	-11.72131	1,839	AOL ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK RA SEM TLCR VRSK
3,000	-11.72131	5,536	AOL ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK RA SEM TLCR VRSK

Table 9: Successive arithmetic truncation method result for 3,000 of the most liquid US stocks, with cardinality, $K = 15$.

Number of Iterations	Objective Value	Time (sec)	Optimal Assets
10	-12.37960	19	AOL AONE ART CFL CFN CIE CIT CVE DGI DOLE LOPE MJN SEM TLCR VRSK
100	-11.72131	63	AOL ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK RA SEM TLCR VRSK
300	-11.72131	161	AOL ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK RA SEM TLCR VRSK
500	-11.72131	290	AOL ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK RA SEM TLCR VRSK
1,000	-11.72131	507	AOL ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK RA SEM TLCR VRSK
3,000	-11.72131	1259	AOL ART CFL CFN CIE CIT CVE DGI DOLE MJN NFBK RA SEM TLCR VRSK

Table 10: Successive geometric truncation method result for 3,000 of the most liquid US stocks with cardinality constraint, $K = 15$.

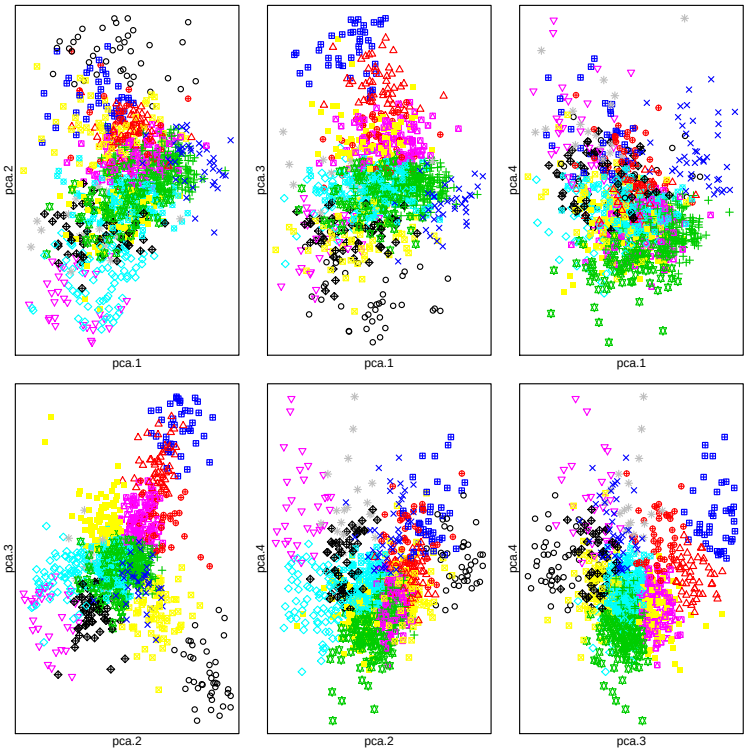

Figure 4.14: Projection onto a 4-dimensional space spanned by the first four dominant PCA factors, based on daily returns for the most liquid 500 US traded stocks, sampled over the period from 2002-01-02 to 2010-01-22. Symbols with different colors correspond to different cluster groups.

sive truncation, unless otherwise stated.

Local Relaxation and CPLEX MIQP Each of the six panels in Figure 4.14 represents a 2-dimensional view of the 4-dimensional PCA projection of the returns of our 500 assets. In the same figure, we have also shown the initial K clusters, indicated by different colors and symbols. The local relaxation method that solves a CCQP with cardinality equal to 15 may be initialized with the cluster centroids assets, iteratively solve for relaxed but smaller QPs by including a subset of the cluster members, then followed by moving to a new set of centroids with strictly increasing optimality.

A prototype of the local relaxation algorithm has been implemented in R running in a 64-Bit Linux kernel. The solution of the relaxed QP's required by the new algorithm and the truncation method are found using the QP solver in R, which is based on Goldfarb and Idnani (1982). To help assess the performance of our proposed algorithm, we present three solved problems, each with different universe sizes and cardinality values. The sets of parameters used in the algorithm for these three problems are given in Table 11.

For comparison, we have implemented CPLEX v12 in a 64-Bit Linux kernel using the CPLEX C++ API to access the built-in MIQP solver, which is based on the dynamic search branch-and-cut method (IBM, 2010). The code is set to run in parallel using up to eight threads on a Quad Intel Core i7 2.8GhZ CPUs with access to a total of 16GB of memory. Table 12 shows the parameter settings used. Table 13 shows the result, for the 500-choose-15 case, with a run-time cut-off set to

	(500, 15)	(3000, 15)	(3000, 100)
$M_{\min}$	5	5	3
$M_{\max}$	30	30	10
σ_d	0.2	0.2	0.2
α_s	80	80	80
ϵ_0	1×10^{-5}	1×10^{-5}	1×10^{-5}
$\bar{\eta}$	10	10	10
Υ	1.1	1.2	1.8

Table 11: Parameter setup for *local relaxation* method. (500, 15) means solving a 500 universe problem with cardinality equal to 15.

Parameter	Setting
Preprocessing	`true`
Heuristics	`RINS at root node only`
Algorithm	`branch-and-cut`
MIP emphasis	`balance optimality and feasibility`
MIP search method	`dynamic search`
Branch Variable Selection	`automatic`
Parallel mode	`deterministic, using up to 8 threads`

Table 12: CPLEX parameter settings (with the rest of the parameters set to their default values).

1,200 sec. Figure 4.15 shows the performance comparison between CPLEX and the new algorithm as a function of run-time. We see that the local relaxation method is able to obtain a better solution than CPLEX right from the beginning, and it strictly dominates CPLEX throughout the testing period.

Next, we assess the scalability of the new algorithm, by switching to the larger dataset which consists of 3,000 actively traded stocks. As before, we first project the factored covariance matrix onto a 4-dimensional PCA space and then identify the initial K clusters based on *k-means*, as shown in Figure 4.16. Table 14 shows the computational result. We see that the local relaxation method is able to reach a significantly better result than CPLEX quickly. Figure 4.17 shows the performance comparison of the local relaxation method verses CPLEX for the 3,000 choose 15 case, which illustrates dominance of the proposed algorithm over CPLEX for a prolonged period of time.

Local Relaxation with warm start One additional feature of the local relaxation algorithm is that it can be warm started using the result from another algorithm or from a similar problem

Method	Objective Value	Time (sec)	Optimal Assets
CPLEX	-0.00796	1200	AOI AYI CCF CPO DD DST DY FBP HCC LRCX MRCL NST TMK UIS WPP
Local Relaxation	-0.00870	1200	AOI AVT AYI BIG BPOP CCF CPO CSCO DST HCC HMA IBM MCRL TAP ZNT

Table 13: Computational result for the 500 asset case with cardinality, $K = 15$. For CPLEX and the *local relaxation* algorithm, we have imposed a maximum run time cutoff at 1,200 seconds.

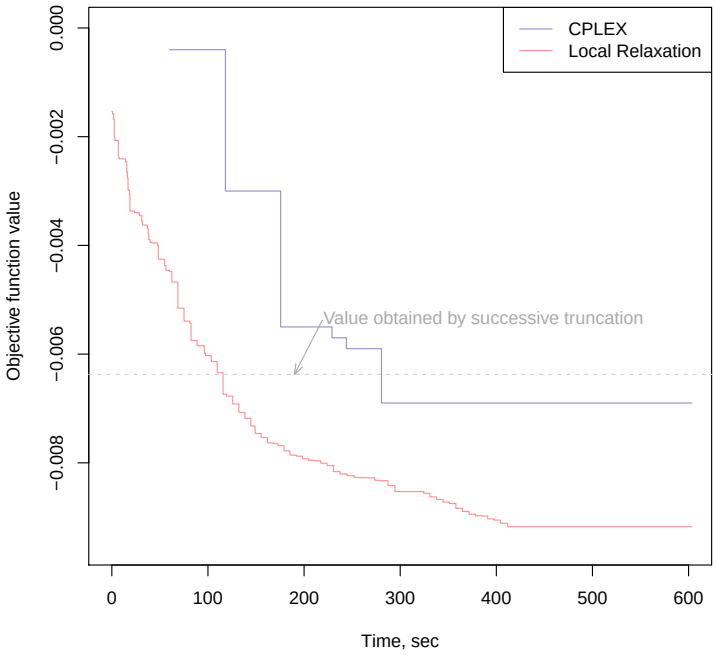

Figure 4.15: CPLEX versus *local relaxation* method performance comparison for the 500 asset case, with cardinality, $K = 15$, both are cold started.

Method	Objective Value	Time (sec)	Optimal Assets
Successive Truncation	-12.08425	36	AOL AONE ART CFL CFN CIE CIT CVE DGI DOLE EDMC MJN SEM TLCR VRSK
CPLEX	-12.30410	25200	AOL AONE ART CFL CFN CIE CIT CVE DGI DOLE MJN SEM TLCR VECO VRSK
Local Relaxation	-12.76492	7200	AOL AONE ARST ART BPI CFL CFN CIE CIT CVE DGI DOLE MJN SEM TLCR

Table 14: Computational result for the 3,000 asset case, with cardinality, $K = 15$. Successive truncation is done over 10 iterations.

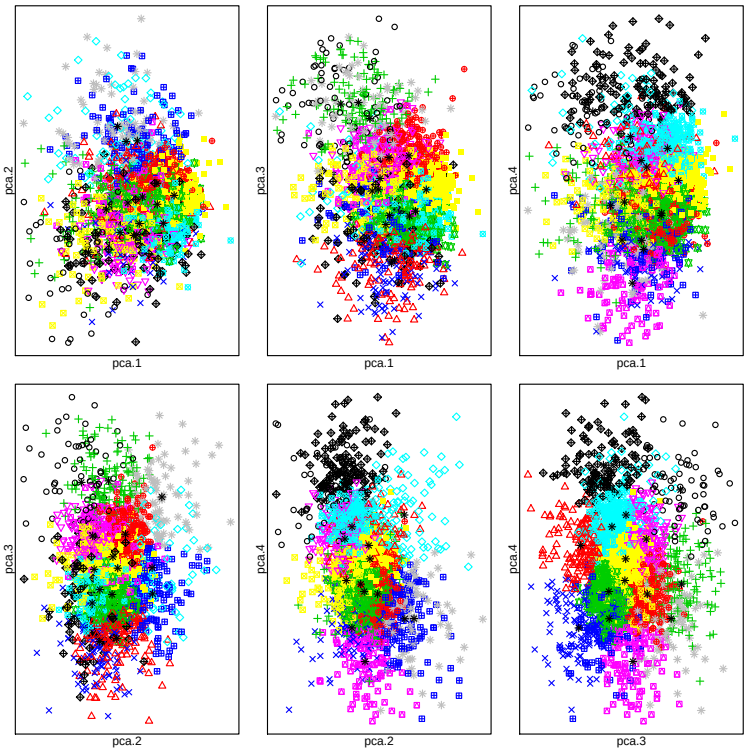

Figure 4.16: Projection onto a four dimensional space spanned by the first four dominant PCA factors, based on monthly returns for the most liquid 3,000 US traded stocks, sampled over the period from 2005-05-31 to 2010-04-30. Symbols with different colors correspond to different cluster groups.

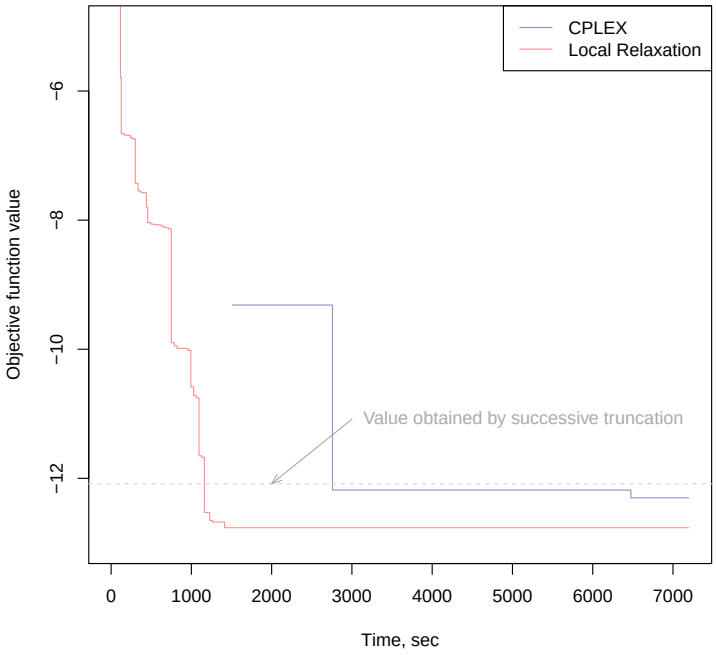

Figure 4.17: CPLEX versus *local relaxation* method performance comparison for the 3,000 asset case, with cardinality constraint $K = 15$, both cold started.

directly. We have seen that the simple successive truncation algorithm could often lead to a good local minimum with little computational cost. To take advantage of this, we propose warm starting the local relaxation algorithm using the result from successive truncation as the initial centroid assets. Since local relaxation method is strictly feasible and monotone improving, we can set a time limit and take the result at the end of the period. This way, we are guaranteed a solution no worse than the one obtained by successive truncation. This hybrid approach gives the best combination and can be shown to be significantly more efficient than a warmed started CPLEX solve.

Figure 4.18 shows the result, solving a 500 asset CCQP with $K = 15$, by warm starting the local relaxation method with output from successive truncation, based on 20 iterative truncations. It can be seen that the new algorithm dominates the CPLEX result.

Figure 4.19 shows the result, solving a 3,000 asset CCQP with $K = 100$, by warm starting the local relaxation method with output from successive truncation. Note that at least for this instance, CPLEX has never quite managed to find a minimum that comes close to the one obtained by the proposed algorithm, even after it has been left running for 35 hours.

Warm starting CPLEX with Local Relaxation It is reasonable to conjecture that the strength of the *local relaxation* algorithm is in its ability to explore structure of the problem, and hence in identifying the sub-group of assets within which the optimal solution set lies. To see whether the branch-and-cut method in CPLEX is able to improve on the solution from the new algorithm once the algorithm has identified a sub-asset group, we propose the following test:

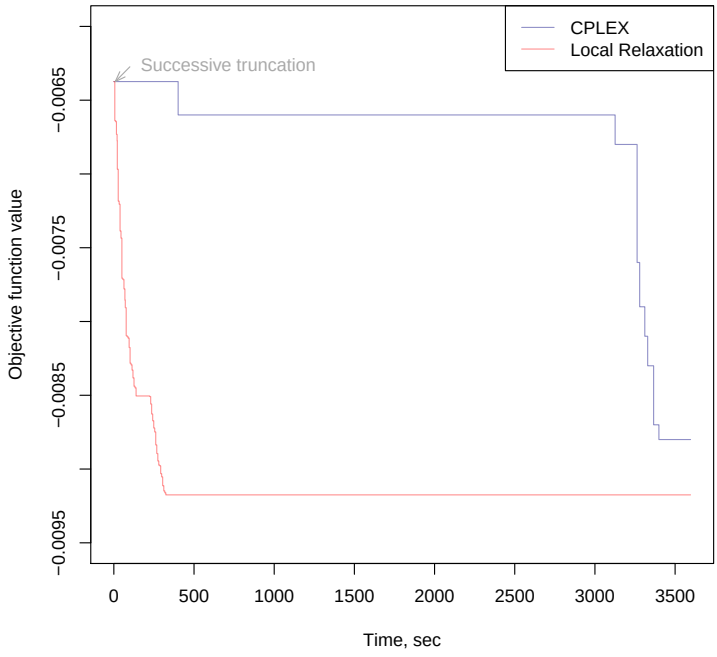

Figure 4.18: CPLEX versus *local relaxation* method performance comparison for the 500 assets universe, with cardinality constraint, $K = 15$; Both methods are warm started using the solution of successive truncation over 20 iterations. Maximum run-time is set to one hour.

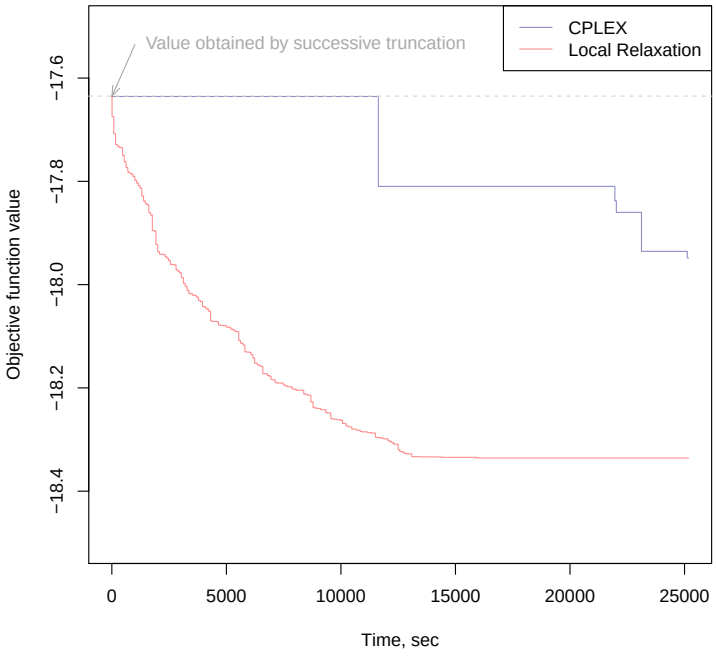

Figure 4.19: CPLEX versus *local relaxation* method performance comparison for the 3,000 assets universe, with cardinality constraint, $K = 100$; Both methods are warm started using the solution of arithmetic successive truncation over 20 iterations. Maximum run-time is set to seven hours.

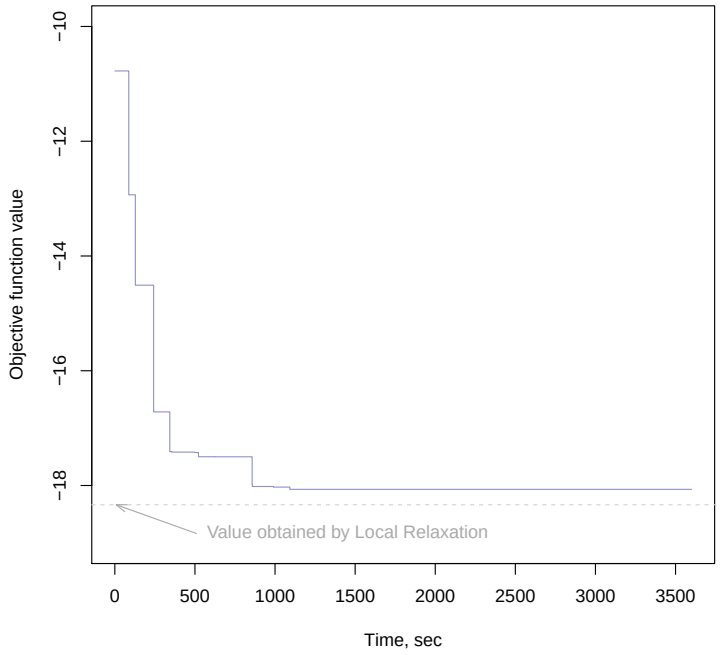

Figure 4.20: Apply CPLEX MIQP to the union of the cluster groups identified by *local relaxation* algorithm upon convergence (the beginning of the flat line in Figure 4.19). Maximum run-time is set to one hour.

- ▷ solve the CCQP with *local relaxation* algorithm;
- ▷ once the algorithm has reached a stationary group of assets (here stationary refers to the observation that the cluster groups do not change over certain number of iterations), stop the algorithm and record the union of assets over all identified cluster groups;
- ▷ solve the CCQP to optimality only for this group of assets using CPLEX.

To illustrate the idea, we apply the above procedure to the result for the 3,000-choose-100 case. In Figure 4.19 we see that the algorithm appears to have converged after 15,916 seconds. We take a union of the assets in the 100 clusters from the *local relaxation* algorithm, then feed these assets into CPLEX. Figure 4.20 shows the output from CPLEX at each iteration. It can be seen that the CPLEX branch-and-cut algorithm takes almost 20 minutes to reach a level close to the result given by the *local relaxation* method, and never quite reach it within the run-time limit of one hour.

Mean-variance efficient frontier The set of optimal portfolios formed based on different risk tolerance parameters is known as *frontier portfolios*, in the sense that all portfolios on the frontier are optimal, given the specific risk tolerance of an investor. Therefore, a mean-variance optimizing agent will only choose frontier portfolios. Figure 4.21 shows this efficient frontier for our universe of 3,000 US stocks. We observe that, in-sample, the CCQP solutions are strictly dominated by those based on QP - a consequence of the cardinality constraint which *cuts off* parts of the feasible space. We observe that this domination increases as we decrease the cardinality of the problem, as larger

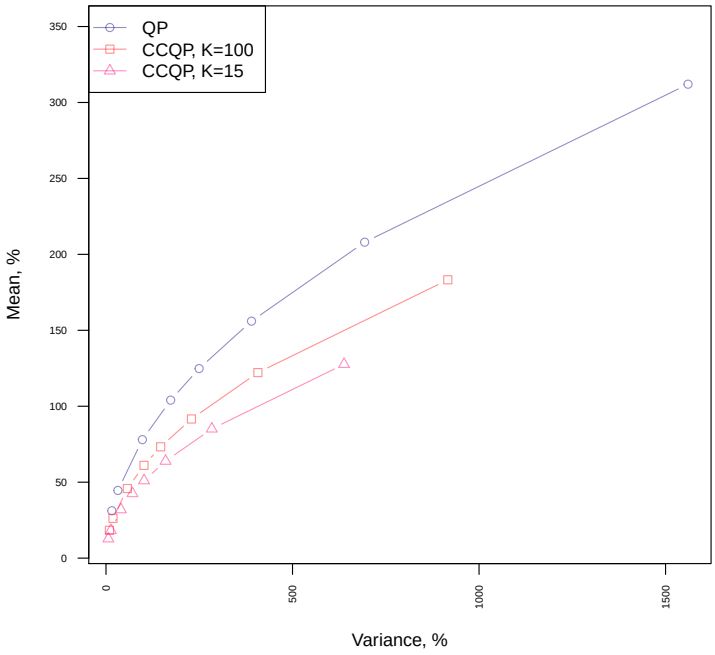

Figure 4.21: Mean-variance efficient frontiers for a 3,000 asset universe. QP is the frontier without the cardinality constraint. CCQP is the frontier in presence of cardinality. To produce the frontier for CCQP, we warm-start the local relaxation algorithm, based on the successive truncation solution, and set a maximum run time limit of 252,000 seconds for the case where $K = 100$ and 3,600 second for the case where $K = 15$.

parts of the feasible space becomes inaccessible to the CCQP. However, in practice, the robustness of forming a portfolio with smaller subset of assets with higher expected returns can often lead to better out of sample performance.

5 Conclusion

In this thesis, new approaches to model the three key statistical and algorithm aspects of optimal portfolio construction have been presented.

For the first aspect, we have proposed a complete model for returns and realized measures of volatility, x_t, where the latter is tied directly to the conditional volatility h_t. The motivation is to include high frequency data in a rigorous way in a classic GARCH framework that has so far been used predominately in dealing with less frequently sampled data, hence without having to be concerned with market microstructure noise. We have demonstrated that the model is straightforward to estimate and offers a substantial improvement in the empirical fit, relative to standard GARCH models. The model is informative about realized measurement, such as its accuracy. The proposed framework induces an interesting reduced-form model for $\{r_t, h_t\}$, that is similar to that of a stochastic volatility model with leverage effect. Our empirical analysis can be extended in a number of ways. For instance, including a jump robust realized measure of volatility would be an interesting extension, because Bollerslev, Kretschmer, Pigorsch, and Tauchen (2009) found that the leverage effect primarily acts through the continuous volatility component. Another possible extension is to introduce a bivariate model of open-to-close and close-to-open returns, as an alternative to modeling close-to-close returns, see Andersen et al. (2011). The *Realized GARCH* framework is naturally extended to a multi-factor structure. Say m realized measures and k latent volatility variables. The *Realized GARCH* model discussed in this thesis corresponds to the case $k = 1$, whereas the MEM framework corresponds to the case $m = k$. Such a hybrid framework would enable us to conduct inference about the number of latent factors, k. We could, in particular, test the one-factor structure, conjectured to be sufficient for the realized measure used in this paper, against the two-factor structure implied by MEM. For the extension of *Realized GARCH* to multivariate settings, we could explore more general frameworks by loosening the assumption that volatility and correlation operate at different time scales.

The increasing role that high frequency finance plays in the market has also motivated the part of the thesis on expected return forecast. The proposed model is based on self-excited point process augmented by trade size and limit order book information. The motivation for incorporate these two potential *marks* is based on underlying economic intuition. First, the trade size mark acts as a mechanism to differentiate large orders that is more likely to induce price movement than small noise trading of insignificant impact. Second, information from the limit order book adds an extra dimension to gauge supply-demand imbalance of the market. Game theoretic utility balancing argument aside, a more sell order loaded book will signal that more investors are looking for an opportunity to liquidate or sell short a position, and *vice versa* for a buy order loaded book. Nowadays, sophisticated execution algorithms contribute to over 60% of total trading volume on the NYSE, with similar dominance in other markets. These algorithms aim to minimize market impact of large orders, by slicing them into smaller units and then submitting to different layers of the limit order book at optimal times. The proposed frameworks in this paper are tested using the LSE order book data and result suggests that the inclusion of limit order book information can lead to a more robust estimation when compared with the basic bivariate model, as defined by goodness-of-fit of the model implied distribution to the empirical distribution of the inter-arrival times of the bid and ask side market orders. As for trade size, result suggests that it could have an adverse effect on model performance. This could be attributable to the fact that most orders, before they reach

the exchange, have been sliced by execution algorithms so that the size is close to the median daily average, so as to minimize signaling effect. Hence a large portion of the fluctuation around the median volume is noise, which lowers the overall signal to noise ratio.

The final key aspect of the thesis is portfolio optimization with cardinality constraints. The *NP*-hard nature of cardinality constrained mean-variance portfolio optimization problems has led to a variety of different algorithms with varying degrees of success in reaching optimality given limited computational resources and under the presence of strict time constraints in practice. The computational effort needed to be assured of solving problems of the size considered here is truly astronomical and way beyond the performance of machines envisaged. However, solving a problem exactly is of questionable valued over that of obtaining a "near" solution. The proposed *local relaxation* algorithm exploits the inherent structure of the objective function. It solves a sequence of small, local, quadratic-programs by first projecting asset returns onto a reduced metric space, followed by clustering in this space to identify sub-groups of assets that best accentuate a suitable measure of similarity amongst different assets. The algorithm can either be cold started using the centroids of initial clusters or be warm started based on the outputs of a previous result. Empirical result, using baskets of up to 3,000 stocks and with different cardinality constraints, indicates that the proposed algorithm is able to achieve significant performance gain over a sophisticated branch-and-cut method.

A Appendix of Proofs

Proof. [Proposition 19] The first result follows by substituting $\log x_t = \varphi \log h_t + \xi + w_t$ and $\log r_t^2 = \log h_t + \kappa + v_t$ into the GARCH equation and rearranging. Next, we substitute $\log h_t = (\log x_t - \xi - w_t)/\varphi$, $\log r_t^2 = (\log x_t - \xi - w_t)/\varphi + \kappa + v_t$, and multiply by φ, and find

$$\log x_t - \xi - w_t = \varphi\omega + \sum_{i=1}^{p\vee q}(\beta_i+\alpha_i)(\log x_{t-i} - \xi - w_{t-i}) + \varphi\sum_{j=1}^{q}\gamma_j \log x_{t-j} + \varphi\sum_{j=1}^{q}\alpha_j(\kappa + v_{t-j})$$

so with $\pi_i = \alpha_i + \beta_i + \gamma_i\varphi$, we have

$$\log x_t = \xi(1-\beta_\bullet - \alpha_\bullet) + \varphi\kappa\alpha_\bullet + \varphi\omega + \sum_{i=1}^{p\vee q}\pi_i \log x_{t-i} + w_t - \sum_{i=1}^{p}(\alpha_i+\beta_i)w_{t-i} + \varphi\sum_{j=1}^{q}\alpha_j v_{t-j}.$$

When $\varphi = 0$, the measurement equation shows that $\log x_t$ is an iid process. □

Proof. [Lemma 22] First note that

$$\frac{\partial g_t'}{\partial \lambda} = \left(0, \dot{h}_{t-1}, \ldots, \dot{h}_{t-p}, 0_{p+q+1\times q}\right) =: \dot{H}_{t-1},$$

Thus from the GARCH equation, $\tilde{h}_t = \lambda' g_t$, we have that

$$\dot{h}_t = \frac{\partial g_t'}{\partial \lambda}\lambda + g_t = \dot{H}_{t-1}\lambda + g_t = \sum_{i=1}^{p}\beta_i \dot{h}_{t-i} + g_t.$$

Similarly, the second order derivative, is given by

$$\begin{aligned}
\ddot{h}_t &= \frac{\partial(g_t + \dot{H}_{t-1}\lambda)}{\partial\lambda'} \\
&= \frac{\partial g_t}{\partial\lambda'} + \dot{H}_{t-1} + \frac{H_{t-1}}{\partial\lambda'}\lambda \\
&= \dot{H}_{t-1}' + \dot{H}_{t-1} + \sum_{i=1}^{p}\beta_i\frac{\partial \dot{h}_{t-i}}{\partial\lambda'} \\
&= \sum_{i=1}^{p}\beta_i \ddot{h}_{t-i} + \dot{H}_{t-1}' + \dot{H}_{t-1}.
\end{aligned}$$

For the starting values we observe that regardless of $(h_0, \ldots, h_{p-1})$ being treated as fixed or as a vector of unknown parameters, we have $\dot{h}_s = \ddot{h}_s = 0$. Given the structure of $\ddot{h}_t$ this implies $\ddot{h}_1 = 0$.

When $p = q = 1$ it follows immediately that $\dot{h}_t = \sum_{j=0}^{t-1}\beta^j g_{t-j}$. Similarly we have

$$\ddot{h}_t = \sum_{j=0}^{t-1}\beta^j(\dot{H}_{t-1-j} + \dot{H}_{t-1-j}) = \sum_{j=0}^{t-2}\beta^j(\dot{H}_{t-1-j} + \dot{H}_{t-1-j})$$

where $\dot{H}_t = (0_{3\times1}, \dot{h}_t, 0_{3\times1})$ and where the second equality follows by $\dot{H}_0 = 0$. The results now

follows by

$$\begin{aligned}
\sum_{i=0}^{t-2}\beta^i\dot{h}_{t-1-i} &= \sum_{i=0}^{t-2}\beta^i\sum_{j=0}^{t-1-i-1}\beta^j g_{t-1-i-j} \\
&= \sum_{i=0}^{t-2}\beta^i\sum_{k-i-1=0}^{t-i-2}\beta^{k-i-1}g_{t-k} \\
&= \sum_{i=0}^{t-2}\sum_{k=i+1}^{t-1}\beta^{k-1}g_{t-k} \\
&= \sum_{k=1}^{t-1}k\beta^{k-1}g_{t-k}.
\end{aligned}$$

□

Proof. [Proposition 23] Recall that $u_t = \tilde{x}_t - \psi' m_t$ and $\tilde{h}_t = g_t'\lambda$. So derivative with respect to $\tilde{h}_t$ are given by

$$\begin{aligned}
\frac{\partial z_t}{\partial \tilde{h}_t} &= \frac{\partial r_t \exp(-\frac{1}{2}\tilde{h}_t)}{\partial \tilde{h}_t} = -\frac{1}{2}z_t \qquad \text{so that} \qquad \frac{\partial z_t^2}{\partial \tilde{h}_t} = -z_t^2, \\
\dot{u}_t = \frac{\partial u_t}{\partial \tilde{h}_t} &= -\varphi + \frac{1}{2}z_t\tau'\dot{a}_t, \\
\ddot{u}_t = \frac{\partial \dot{u}_t}{\partial \tilde{h}_t} &= \frac{\partial\left(-\varphi + \frac{1}{2}z_t\dot{a}(z_t)'\tau\right)}{\partial \tilde{h}_t} = -\frac{1}{4}\tau'\left(z_t\dot{a}_t + z_t^2\ddot{a}_t\right).
\end{aligned}$$

So with $\ell_t = -\frac{1}{2}\{\tilde{h}_t + z_t^2 + \log(\sigma_u^2) + u_t^2/\sigma_u^2\}$ we have

$$\frac{\partial \ell_t}{\partial u_t} = 2\frac{u_t}{\sigma_u^2} \qquad \text{and} \qquad \frac{\partial \ell_t}{\partial \tilde{h}_t} = -\frac{1}{2}\left\{1 + \frac{\partial z_t^2}{\partial \tilde{h}_t} + \frac{\partial u_t^2/\partial \tilde{h}_t}{\sigma_u^2}\right\} = -\frac{1}{2}\left\{1 - z_t^2 + \frac{2u_t\dot{u}_t}{\sigma_u^2}\right\}.$$

Derivatives with respect to λ are

$$\begin{aligned}
\frac{\partial z_t}{\partial \lambda} &= \frac{\partial z_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial \lambda} = -\frac{1}{2}z_t\dot{h}_t \\
\frac{\partial u_t}{\partial \lambda} &= \frac{\partial u_t}{\partial \tilde{h}_t}\frac{\partial \tilde{h}_t}{\partial \lambda} = \dot{u}_t\dot{h}_t \\
\frac{\partial \dot{u}_t}{\partial \lambda'} &= \ddot{u}_t\dot{h}_t' \\
\frac{\partial \ell_t}{\partial \lambda} &= \frac{\partial \ell_t}{\partial \tilde{h}_t}\dot{h}_t = -\frac{1}{2}\left\{1 - z_t^2 + \frac{2u_t\dot{u}_t}{\sigma_u^2}\right\}\dot{h}_t.
\end{aligned}$$

Derivatives with respect to ψ are

$$\begin{aligned}
\frac{\partial u_t}{\partial \xi} &= -1, \qquad \frac{\partial \dot{u}_t}{\partial \xi} = 0, \qquad \text{and} \qquad \frac{\partial \ell_t}{\partial \xi} = \frac{\partial \ell_t}{\partial u_t}\frac{\partial u_t}{\partial \xi} = -2\frac{u_t}{\sigma_u^2}, \\
\frac{\partial u_t}{\partial \varphi} &= -\tilde{h}_t, \qquad \frac{\partial \dot{u}_t}{\partial \varphi} = -1, \qquad \text{and} \qquad \frac{\partial \ell_t}{\partial \varphi} = \frac{\partial \ell_t}{\partial u_t}\frac{\partial u_t}{\partial \varphi} = -2\frac{u_t}{\sigma_u^2}\tilde{h}_t, \\
\frac{\partial u_t}{\partial \tau} &= -a_t, \qquad \frac{\partial \dot{u}_t}{\partial \tau} = \frac{1}{2}z_t\dot{a}_t \qquad \text{and} \qquad \frac{\partial \ell_t}{\partial \tau} = \frac{\partial \ell_t}{\partial u_t}\frac{\partial u_t}{\partial \tau} = -2\frac{u_t}{\sigma_u^2}a_t.
\end{aligned}$$

Similarly, $\frac{\partial \ell_t}{\partial \sigma_u^2} = -\frac{1}{2}(\sigma_u^{-2} - u_t^2\sigma_u^{-4})$. Now we turn to the second order derivatives.

$$\begin{aligned}
-2\frac{\partial^2 \ell_t}{\partial \lambda \partial \lambda'} &= \dot{h}_t\left\{-\frac{\partial z_t^2}{\partial \lambda'} + \frac{2}{\sigma_u^2}\left(\dot{u}_t\frac{\partial u_t}{\partial \lambda'} + u_t\frac{\partial \dot{u}_t}{\partial \lambda'}\right)\right\} + (1 - z_t^2 + \frac{2u_t}{\sigma_u^2}\dot{u}_t)\frac{\partial \dot{h}_t}{\partial \lambda'} \\
&= \dot{h}_t\left\{z_t^2 + \frac{2}{\sigma_u^2}\left(\dot{u}_t^2 + u_t\ddot{u}_t\right)\dot{h}_t'\right\} + (1 - z_t^2 + \frac{2u_t}{\sigma_u^2}\dot{u}_t)\ddot{h}_t.
\end{aligned}$$

Similarly, since $\frac{\partial z_t}{\partial \psi} = 0$ we have

$$\begin{aligned}
-2\frac{\partial^2 \ell_t}{\partial \lambda \partial \xi} &= \frac{\partial (1 - z_t^2 + \frac{2u_t}{\sigma_u^2}\dot{u}_t)\dot{h}_t}{\partial \xi} = 2\dot{h}_t\left(\frac{\partial u_t}{\partial \psi'}\frac{\dot{u}_t}{\sigma_u^2} + \frac{u_t}{\sigma_u^2}\frac{\partial \dot{u}_t}{\partial \xi}\right) = 2\dot{h}_t\left(-\frac{\dot{u}_t}{\sigma_u^2} + 0\right) \\
-2\frac{\partial^2 \ell_t}{\partial \lambda \partial \varphi} &= \frac{\partial (1 - z_t^2 + \frac{2u_t}{\sigma_u^2}\dot{u}_t)\dot{h}_t}{\partial \varphi} = 2\dot{h}_t\left(\frac{\partial u_t}{\partial \varphi}\frac{\dot{u}_t}{\sigma_u^2} + \frac{u_t}{\sigma_u^2}\frac{\partial \dot{u}_t}{\partial \varphi}\right) = 2\dot{h}_t\left(-\tilde{h}_t\frac{\dot{u}_t}{\sigma_u^2} - \frac{u_t}{\sigma_u^2}\right) \\
-2\frac{\partial^2 \ell_t}{\partial \lambda \partial \tau'} &= 2\dot{h}_t\left(\frac{\partial u_t}{\partial \tau'}\frac{\dot{u}_t}{\sigma_u^2} + \frac{u_t}{\sigma_u^2}\frac{\partial \dot{u}_t}{\partial \tau'}\right) = 2\dot{h}_t\left(-a_t'\frac{\dot{u}_t}{\sigma_u^2} + \frac{u_t}{\sigma_u^2}\frac{1}{2}z_t\dot{a}_t\right),
\end{aligned}$$

so that

$$\frac{\partial^2 \ell_t}{\partial \lambda \partial \psi'} = \frac{\dot{u}_t}{\sigma_u^2}\dot{h}_t m_t' + \frac{u_t}{\sigma_u^2}\dot{h}_t b_t', \qquad \text{with} \quad b_t = (0, 1, -\frac{1}{2}z_t\dot{a}_t')'.$$

$$\frac{\partial^2 \ell_t}{\partial \lambda \partial \sigma_u^2} = -\frac{1}{2}\frac{\partial (1 - z_t^2 + \frac{2u_t}{\sigma_u^2}\dot{u}_t)\dot{h}_t}{\partial \sigma_u^2} = \frac{u_t\dot{u}_t\dot{h}_t}{\sigma_u^4}$$

$$\frac{\partial^2 \ell_t}{\partial \psi \partial \psi'} = -\frac{1}{\sigma_u^2}m_t m_t'$$

$$\frac{\partial^2 \ell_t}{\partial \psi \partial \sigma_u^2} = -\frac{1}{2}(-\frac{2u_t}{\sigma_u^4})m_t = \frac{u_t}{\sigma_u^4}m_t$$

$$\frac{\partial^2 \ell_t}{\partial \sigma_u^2 \partial \sigma_u^2} = -\frac{1}{2}\left(\frac{-1}{\sigma_u^4} + 2\frac{u_t^2}{\sigma_u^6}\right) = \frac{1}{2}\frac{\sigma_u^2 - 2u_t^2}{\sigma_u^6}.$$

□

Proof. [Proposition 26] We note that

$$\begin{aligned}
h_t &= \exp\left(\sum_{i=0}^{\infty}\pi^i(\mu + \gamma w_{t-1})\right) = e^{\frac{\mu}{1-\pi}}\prod_{i=0}^{\infty}\mathrm{E}\exp\left(\gamma\pi^i\tau(z_{t-i})\right)\mathrm{E}\left(\exp\left\{\pi^i\gamma u_{t-i}\right\}\right), \\
h_t^2 &= \exp\left(2\sum_{i=0}^{\infty}\pi^i(\mu + \gamma w_{t-1})\right) = e^{\frac{2\mu}{1-\pi}}\prod_{i=0}^{\infty}\mathrm{E}\exp\left(2\gamma\pi^i\tau(z_{t-i})\right)\mathrm{E}\left(\exp\left\{2\pi^i\gamma u_{t-i}\right\}\right),
\end{aligned}$$

and using results, such as

$$\mathrm{E}\left(\prod_{i=0}^{\infty}\exp\left\{\pi^i\gamma u_{t-i}\right\}\right)=\prod_{i=0}^{\infty}\mathrm{E}\left(\exp\left\{\pi^i\gamma u_{t-i}\right\}\right)=\prod_{i=0}^{\infty}e^{\frac{\pi^{2i}\gamma^2\sigma_u^2}{2}}=e^{\sum_{i=0}^{\infty}\frac{\pi^{2i}\gamma^2\sigma_u^2}{2}}=e^{\frac{\gamma^2\sigma_u^2/2}{1-\pi^2}},$$

we find that

$$\begin{aligned}
\frac{\mathrm{E}h_t^2}{(\mathrm{E}h_t)^2} &= \frac{e^{\frac{2\mu}{1-\pi}}\prod_{i=0}^{\infty}\mathrm{E}\exp\left(2\gamma\pi^i w_{t-1}\right)}{e^{\frac{2\mu}{1-\pi}}\prod_{i=0}^{\infty}\left\{\mathrm{E}\exp\left(\gamma\pi^i w_{t-1}\right)\right\}^2}\\
&= \left(\prod_{i=0}^{\infty}\frac{1-2\pi^i\gamma\tau_2}{\sqrt{1-4\pi^i\gamma\tau_2}}\right)\frac{e^{\sum_{i=0}^{\infty}\frac{4\pi^{2i}\gamma^2\tau_1^2}{2(1-4\pi^i\gamma\tau_2)}}}{e^{2\sum_{i=0}^{\infty}\frac{\pi^{2i}\gamma^2\tau_1^2}{2(1-2\pi^i\gamma\tau_2)}}}\frac{e^{-\frac{2\gamma\tau_2}{1-\pi}}}{e^{-2\frac{\gamma\tau_2}{1-\pi}}}\frac{e^{\frac{2\gamma^2\sigma_u^2}{1-\pi^2}}}{e^{\frac{\gamma^2\sigma_u^2}{1-\pi^2}}}\\
&= \left(\prod_{i=0}^{\infty}\frac{1-2\pi^i\gamma\tau_2}{\sqrt{1-4\pi^i\gamma\tau_2}}\right)e^{\sum_{i=0}^{\infty}\frac{2\pi^{2i}\gamma^2\tau_1^2}{(1-4\pi^i\gamma\tau_2)}-\frac{\pi^{2i}\gamma^2\tau_1^2}{(1-2\pi^i\gamma\tau_2)}}e^{\frac{\gamma^2\sigma_u^2}{1-\pi^2}}\\
&= \left(\prod_{i=0}^{\infty}\frac{1-2\pi^i\gamma\tau_2}{\sqrt{1-4\pi^i\gamma\tau_2}}\right)e^{\sum_{i=0}^{\infty}\frac{\pi^{2i}\gamma^2\tau_1^2}{(1-6\pi^i\gamma\tau_2+8\pi^{2i}\gamma^2\tau_2^2)}}e^{\frac{\gamma^2\sigma_u^2}{1-\pi^2}}
\end{aligned}$$

where the last equality uses

$$\begin{aligned}
\frac{2\pi^{2i}\gamma^2\tau_1^2}{(1-4\pi^i\gamma\tau_2)}-\frac{\pi^{2i}\gamma^2\tau_1^2}{(1-2\pi^i\gamma\tau_2)} &= \frac{2\pi^{2i}\gamma^2\tau_1^2(1-2\pi^i\gamma\tau_2)-\pi^{2i}\gamma^2\tau_1^2(1-4\pi^i\gamma\tau_2)}{(1-4\pi^i\gamma\tau_2)(1-2\pi^i\gamma\tau_2)}\\
&= \frac{\pi^{2i}\gamma^2\tau_1^2}{(1-4\pi^i\gamma\tau_2)(1-2\pi^i\gamma\tau_2)}=\frac{\pi^{2i}\gamma^2\tau_1^2}{(1-6\pi^i\gamma\tau_2+8\pi^{2i}\gamma^2\tau_2^2)}.
\end{aligned}$$

Recall that

$$\frac{\mathrm{E}(r_t^4)}{\mathrm{E}(r_t^2)^2}=3\left(\prod_{i=0}^{\infty}\frac{1-2\pi^i\gamma\tau_2}{\sqrt{1-4\pi^i\gamma\tau_2}}\right)\exp\left\{\sum_{i=0}^{\infty}\frac{\pi^{2i}\gamma^2\tau_1^2}{1-6\pi^i\gamma\tau_2+8\pi^{2i}\gamma^2\tau_2^2}\right\}\exp\left\{\frac{\gamma^2\sigma_u^2}{1-\pi^2}\right\}.$$

For the first term on the right hand side, we have

$$\begin{aligned}
\log\prod_{i=0}^{\infty}\frac{1-2\pi^i\gamma\tau_2}{\sqrt{1-4\pi^i\gamma\tau_2}} &\simeq \int_0^{\infty}\log\frac{1-2\pi^x\gamma\tau_2}{\sqrt{1-4\pi^x\gamma\tau_2}}\mathrm{d}x\\
&= \frac{1}{\log\pi}\left\{\sum_{k=1}^{\infty}\frac{(2\gamma\tau_2)^k}{k^2}-\frac{1}{2}\frac{(4\gamma\tau_2)^k}{k^2}\right\}(1-2^{k-1})\\
&= \frac{1}{\log\pi}\sum_{k=1}^{\infty}\frac{(2\gamma\tau_2)^k}{k^2}(1-2^{k-1})\\
&= \frac{\gamma^2\tau_2^2\left\{1+\frac{8}{3}\gamma\tau_2+7(\gamma\tau_2)^2+\frac{96}{5}(\gamma\tau_2)^3+\frac{496}{9}(\gamma\tau_2)^4+\cdots\right\}}{-\log\pi}.
\end{aligned}$$

The second term can be bounded by

$$\frac{\gamma^2\tau_1^2}{1-\pi^2}\leq\sum_{i=0}^{\infty}\frac{\pi^{2i}\gamma^2\tau_1^2}{1-6\pi^i\gamma\tau_2+8\pi^{2i}\gamma^2\tau_2^2}\leq\frac{\gamma^2\tau_1^2}{1-\pi^2}\frac{1}{1-6\pi\gamma\tau_2}.$$

So the approximation error is small when $\gamma\tau_2$ is small. □

Proof. [Proposition 30] Starting from the assumption that the following SDE is well defined

$$d\lambda_t = \beta\left(\rho\left(t\right) - \lambda_t\right)dt + \alpha dN_t,$$

then solution for λ_t takes the form

$$\lambda_t = c\left(t\right) + \int_0^t \alpha e^{-\beta(t-u)} dN_u$$

where

$$c\left(t\right) = c\left(0\right)e^{-\beta t} + \beta\int_0^t e^{-\beta(t-u)}\rho\left(u\right)du.$$

Verify by Ito's lemma on $e^{\beta t}\lambda_t$

$$\begin{aligned} e^{\beta t}\lambda_t &= c\left(0\right) + \beta\int_0^t e^{\beta u}\rho\left(u\right)du + \int_0^t \alpha e^{\beta u} dN_u \\ \beta e^{\beta t}\lambda_t dt + e^{\beta t} d\lambda_t &= \beta e^{\beta t}\rho\left(t\right)dt + \alpha e^{\beta t} dN_t \\ d\lambda_t &= \beta\left(\rho\left(t\right) - \lambda_t\right)dt + \alpha dN_t. \end{aligned}$$

Taking the limit we obtain

$$\begin{aligned} \lim_{t\to\infty} c\left(t\right) &= \lim_{t\to\infty}\left\{c\left(0\right)e^{-\beta t} + \beta\int_0^t e^{-\beta(t-u)}\rho\left(u\right)du\right\} \\ &= \lim_{t\to\infty}\beta\frac{\int_0^t e^{\beta u}\rho\left(u\right)du}{e^{\beta t}} \\ &= \lim_{t\to\infty}\rho\left(t\right) \end{aligned}$$

Treating $\rho\left(t\right)$ as a constant $\rho\left(t\right) \equiv \mu$, then we have

$$\begin{aligned} c\left(t\right) &= c\left(0\right)e^{-\beta t} + \beta\int_0^t e^{-\beta(t-u)}\rho\left(u\right)du \\ &= c\left(0\right)e^{-\beta t} + \mu e^{-\beta t}\left(e^{\beta t} - 1\right) \\ &= \mu + e^{-\beta t}\left(c\left(0\right) - \mu\right) \end{aligned}$$

Note that if we set $c\left(0\right) \equiv \mu$ then the process is simply

$$\lambda_t = \mu + \alpha\int_0^t e^{-\beta(t-u)} dN_u.$$

Therefore we can think of μ as the long run base intensity, i.e. the intensity if there has been no past arrival. □

Proof. [Proposition 32] Since the parameters are bounded, we have

$$\begin{aligned}
L_T^{(1)}\left(\mu_1,\beta_{11},\beta_{12},\alpha_{11},\alpha_{12}\right) &= -\left(\int_0^T \mu_1 dt + \sum_{t_i<t}\int_0^T \alpha_{11}e^{-\beta_{11}(t-t_i)}dt + \sum_{t_j<t}\int_0^T \alpha_{12}e^{-\beta_{12}(t-t_j)}dt\right) \\
&\quad + \int_0^T \log\left(\mu_1 + \sum_{t_i<t}\alpha_{11}e^{-\beta_{11}(t-t_i)} + \sum_{t_j<t}\alpha_{12}e^{-\beta_{12}(t-t_j)}\right) dN_1(t) \\
&= -\mu_1 T - \frac{\alpha_{11}}{\beta_{11}}\sum_{t_i<T}\left(1-e^{-\beta_{11}(T-t_i)}\right) - \frac{\alpha_{12}}{\beta_{12}}\sum_{t_j<T}\left(1-e^{-\beta_{12}(T-t_j)}\right) \\
&\quad + \sum_{t_i<T}\log\left(\mu_1 + \alpha_{11}\sum_{t_{i'}<t_i} e^{-\beta_{11}(t_i-t_{i'})} + \alpha_{12}\sum_{t_{j'}<t_i} e^{-\beta_{12}\left(t_i-t_{j'}\right)}\right).
\end{aligned}$$

We can recursively express

$$\begin{aligned}
R_{11}(i) &= \sum_{i'=1}^{i} e^{-\beta_{11}(t_i-t_{i'})} \\
&= e^{-\beta_{11}(t_i-t_{i-1})}\sum_{i'=1}^{i-1} e^{-\beta_{11}(t_{i-1}-t_{i'})} \\
&= e^{-\beta_{11}(t_i-t_{i-1})}\left(e^{-\beta_{11}(t_{i-1}-t_{i-1})} + \sum_{i'=1}^{i-2} e^{-\beta_{11}(t_{i-1}-t_{i'})}\right) \\
&= e^{-\beta_{11}(t_i-t_{i-1})}\left(1+R_{11}(i-1)\right).
\end{aligned}$$

Now let $j^* = \sup\{j : t_j \le t_i\}$ and $j^*_{-1} = \sup\{j : t_j \le t_{i-1}\}$, again we can recursively express

$$\begin{aligned}
R_{12}(i) &= \sum_{j'=1}^{i} e^{-\beta_{12}\left(t_i-t_{j'}\right)} \\
&= e^{-\beta_{12}\left(t_i-t_{j^*}\right)} + e^{-\beta_{12}\left(t_i-t_{j^*-1}\right)} + \cdots + e^{-\beta_{12}\left(t_i-t_{j^*_{-1}}\right)} + \sum_{j'=1}^{j^*_{-1}} e^{-\beta_{12}\left(t_i-t_{j'}\right)} \\
&= \sum_{\left\{j':t_{i-1}\le t_{j'}<t_i\right\}} e^{-\beta_{12}\left(t_i-t_{j'}\right)} + e^{-\beta_{12}(t_i-t_{i-1})}\sum_{j'=1}^{j^*_{-1}} e^{-\beta_{12}\left(t_{i-1}-t_{j'}\right)} \\
&= e^{-\beta_{12}(t_i-t_{i-1})}R_{12}(i-1) + \sum_{\left\{j':t_{i-1}\le t_{j'}<t_i\right\}} e^{-\beta_{12}\left(t_i-t_{j'}\right)}.
\end{aligned}$$

Similarly for $L_T^{(2)}$, R_{22} and R_{21}. □

Proof. [Proposition 33] It follows directly the proof above, by considering the following log-likelihood

function:

$$
\begin{aligned}
L_T^{(1)}\left(\mu_1, \beta_{11}, \beta_{12}, \alpha_{11}, \alpha_{12}\right) &= -\left(\int_0^T \mu_1 dt + \sum_{t_i<t}\int_0^T \alpha_{11}\frac{w_{1i}}{\bar{w}_1}e^{-\beta_{11}(t-t_i)}dt + \sum_{t_j<t}\int_0^T \alpha_{12}\frac{w_{2i}}{\bar{w}_2}e^{-\beta_{12}(t-t_j)}dt\right) \\
&\quad + \int_0^T \log\left(\mu_1 + \sum_{t_i<t}\alpha_{11}\frac{w_{1i}}{\bar{w}_1}e^{-\beta_{11}(t-t_i)} + \sum_{t_j<t}\alpha_{12}\frac{w_{2i}}{\bar{w}_2}e^{-\beta_{12}(t-t_j)}\right) dN_1(t) \\
&\quad + \sum_{t_i<T}\log\left(\mu_1 + \alpha_{11}\sum_{t_{i'}<t_i}\frac{w_{1i}}{\bar{w}_1}e^{-\beta_{11}(t_i-t_{i'})} + \alpha_{12}\sum_{t_{j'}<t_i}\frac{w_{2i}}{\bar{w}_2}e^{-\beta_{12}\left(t_i-t_{j'}\right)}\right)
\end{aligned}
$$

where w_{1i} and w_{2j} are the trade size for buy and sell orders at time t_i and t_j, respectively. □

References

Y. Aït-Sahalia, P. A. Mykland, and L. Zhang. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100 (472):1394, 2005.

Y. Aït-Sahalia, P. A. Mykland, and L. Zhang. Ultra high frequency volatilities estimation with dependent microstructure noise. September 2009.

T. G. Andersen and T. Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905, 1998.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and H. Ebens. The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76, 2001a.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. The distribution of exchange rate volatility. *Journal of the American Statistical Association*, 96(453):42–55, 2001b. Correction published in 2003, volume 98, page 501.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forcasting realized volatility: Appendix. *Unpublished appendix*, 2001c.

T. G. Andersen, T. Bollerslev, and X. Huang. A reduced form framework for modeling volatility of speculative prices based on realized variation measures. *Journal of Econometrics*, 160:176–189, 2011.

O. E. Barndorff-Nielsen and N. Shephard. *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress*, chapter Variation, jumps and high frequency data in financial econometrics, pages 328–372. Econometric Society Monographs. Cambridge University Press, 2007.

O. E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society* **B**, 64:253–280, 2002.

O. E. Barndorff-Nielsen and N. Shephard. Power and bipower variation with stochastic volatility and jumps (with discussion). *Journal of Financial Econometrics*, 2:1–48, 2004.

O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. Unpublished paper: Nuffield College, Oxford, 2008a.

O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, 76: 1481–1536, 2008b.

O. E. Barndorff-Nielsen, P. R. Hansen, A. Lunde, and N. Shephard. Realised kernels in practice: Trades and quotes. *Econometrics Journal*, 12:1–33, 2009.

M. Bartlett. The spectral analysis of point processes. 25:264–296, 1963.

D. Bienstock. Computational study of a family of mixed-integer quadratic programming problems. *Mathematical programming*, 74:121–140, 1995.

T. Bollerslev. Glossary to ARCH (GARCH). In T. Bollerslev, J. R. Russell, and M. Watson, editors, *Volatility and Time Series Econometrics: Essays in Honour of Robert F. Engle*. Oxford University Press, Oxford, UK, 2009.

T. Bollerslev. Generalized autoregressive heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

T. Bollerslev and J. M. Wooldridge. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariance. *Econometric Reviews*, 11:143–172, 1992.

T. Bollerslev, U. Kretschmer, C. Pigorsch, and G. Tauchen. A discrete-time model for daily s&p500 returns and realized variations: Jumps and leverage effects. *Journal of Econometrics*, 150:151–166, 2009.

J.-P. Bouchaud, M. Mézard, and M. Potters. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2:251–256, 2002.

C. G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. Nuffield College Economics Working Paper 2003-W3., 2003.

G. E. P. Box and G. M. Jenkins. *Time Series Analysis; Forecasting and Control*. Prentice Hall, 1976.

P. Bremaud, editor. *Point Process and Queues, Martingale Dynamics*. Springer-Verlag, 1980.

C. T. Brownless and G. M. Gallo. Comparison of volatility measures: A risk management perspective. *Forthcoming in Journal of Financial Econometrics*, 2010.

M. Carrasco and X. Chen. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18:17–39, 2002.

T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha. Heuristics for cardinality constrained portfolio optimisation. *Computers and Operations Research*, 27(13):1271–1302, 2000.

F. Cipollini, R. F. Engle, and G. M. Gallo. A model for multivariate non-negative valued processes in financial econometrics. *working paper*, 2009.

J. Clausen. Branch and bound algorithms - principles and examples, 2003.

D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* **B**, 34:187–220, 1972.

T. J. Dakin. A tree-search algorithm for mixed integer programming problems. *The Computer Journal*, 8(3):250–255, 1965.

D. Daley and D. Vere-Jones, editors. *An Introduction to the Theory of Point Processes*. Springer-Verlag, 2003.

D. W. Diamond and R. E. Verrecchia. Constraints on short-selling and asset price adjustments to private information. *Journal of Financial Economics*, 18:277–311, 1987.

D. Easley and M. O'Hara. Time and the process of security price adjustment. *Journal of Finance*, 47:905–927, 1992.

R. Engle. New frontiers for arch models. *Journal of Applied Econometrics*, 17:425–446, 2002a.

R. Engle. New frontiers for arch models. *Journal of Applied Financial Economics*, pages 425–446, 2002b.

R. Engle and G. Gallo. A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, pages 3–27, 2006.

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

R. F. Engle. Dynamic conditional correlation - a simple class of multivariate garch models. *Journal of Business & Economic Statistics*, 20:339ï¿œ–350, 2002c.

R. F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 45:987–1007, 1982.

R. F. Engle and V. Ng. Measuring and testing the impact of news on volatility. *Journal of Finance*, 48:1747–1778, 1993.

R. F. Engle and J. G. Rangel. The spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies*, 21:1187–1222, 2008.

R. F. Engle and J. R. Russell. Forecasting transaction rates: The autoregressive conditional duration model. UCSD working paper, 1995.

R. F. Engle and J. R. Russell. Forecasting the frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model. *Journal of Empirical Finance*, 4:187–212, 1997.

R. F. Engle and J. R. Russell. Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 66(5):1127–1163, 1998.

R. Ge and C. Huang. A continous approach to nonlinear integer programming. *Applied Mathematics and Computation*, 34:39–60, 1989.

E. Ghysels and X. Chen. News - good or bad - and its impact on volatility predictions over multiple horizons. *Review of Financial Studies*, 2010. Forthcoming.

P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, 1981.

F. Glover. Future paths for integer programming and links to artificial intelligence. *Comput. Oper. Res.*, 13(5):533–549, 1986.

D. Goldfarb and A. Idnani. *Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs. In Numerical Analysis*. Springer-Verlag, 1982.

D. Goldfarb and A. Idnani. numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programmin*, 27:1–33, 1983.

P. R. Hansen. In-sample fit and out-of-sample fit: Their joint distribution and its implications for model selection. Working paper, Stanford University, 2009.

P. R. Hansen and G. Horel. Quadratic variation by markov chains. 2010.

P. R. Hansen and A. Lunde. An optimal and unbiased measure of realized variance based on intermittent high-frequency data. 2003. CIRANO-CIREQ conference on Realized Volatility, Montreal, Canada, http://www.cireq.umontreal.ca/activites/031107/0311lunde.pdf.

P. R. Hansen, Z. Huang, and H. H. Shek. Realized garch: A joint model of returns and realized measures of volatility. *Journal of Applied Econometrics*, 2011. Forthcoming.

T. Hastie, R. Tibshirani, and J. Friedman, editors. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2003.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. 58:83–90, 1971.

J. H. Holland, editor. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, 1975.

C. C. Holt. Forecasting trends and seasonals by exponentially weighted moving averages. *ONR Research Memorandum*, 52, 1957.

R. Hyndman, A. Koehler, R. Snyder, and S. Grose. A state space framework for automatic forecasting using exponential smoothing methods. Monash Econometrics and Business Statistics Working Papers 9/2000, Monash University, Department of Econometrics and Business Statistics, Aug. 2000. URL `http://ideas.repec.org/p/msh/ebswps/2000-9.html`.

IBM. *IBM ILOG CPLEX Optimization Studio V12.2 Documentation*. IBM ILOG, 2010.

S. T. Jensen and A. Rahbek. Asymptotic normality of the QMLE estimator of ARCH in the nonstationary case. *Econometrica*, 72:641–646, 2004a.

S. T. Jensen and A. Rahbek. Asymptotic inference for non-stationary garch. *Econometric Theory*, 20:1203–1226, 2004b.

N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley & Sons, Inc., 1972.

S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34:975–986, 1984.

H. Konno and H. Yamazaki. Mean-absolute deviation portfolio optimization model and its applications to tokyo stock market. *Management Science*, 37(5):519–531, 1991.

A. H. Land and A. G. Doig. An automatic method of solving discrete programming problems. *Econometrica*, 28(3):497–520, 1960.

J. Large. Estimating quadratic variation when quoted prices jump by a constant increment. *Economics Series Working Papers*, (340), 2007.

S. Lee and B. E. Hansen. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, 10:29–52, 1994.

S. Leyffer. Integrating sqp and branch-and-bound for mixed integer nonlinear programming. *Computational Optimization and Applications*, 18(3):295–309, 2001.

A. Loraschi, A. Tettamanzi, M. Tomassini, and P. Verda. Distributed genetic algorithms with an application to portfolio selection. In *In Artificial neural nets and genetic*, pages 384–387. Springer-Verlag, 1995.

LSE. Rebuild order book service and technical description. Technical report, London Stock Exchange, 2008.

R. L. Lumsdaine. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica*, 10:29–52, 1996.

S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, 11(7):674–693, 1989.

H. M. Markowitz. Portfolio selection. 7:77–91, 1952.

H. M. Markowitz, editor. *Mean-Variance Analysis in Portfolio Choice and Captial Markets*. Blackwell Publishers, 1987.

W. Murray and K.-M. Ng. An algorithm for nonlinear optimization problems withÂ binary variables. *Computational Optimization and Applications*, pages 1–32, 2008. 10.1007/s10589-008-9218-1.

W. Murray and U. V. Shanbhag. A local relaxation approach for the siting of electrical substation. *Computational Optimization and Applications*, 38(3):299–303, 2007.

W. Murray and V. V. Shanbhag. A local relaxation method for nonlinear facility location problems. *Multiscale Optimization Methods and Applications*, 82:173–204, 2006.

W. Murray and H. H. Shek. Local relaxation method for the cardinality constrained portfolio optimization problem. *Journal of Computational Optimization and Applications*, 2011. Submitted.

D. B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59 (2):347–370, 1991.

Y. Ogata. The asymptotic behavour of maximum likelihood estimators for stationary point processes. 30:243–261, 1978.

Y. Ogata. On lewis' simulation method for point processes. IT-27(1), 1981.

J. S. Park, B. H. Lim, Y. Lee, and M. R. Young. A minimax portfolio selection rule with linear programming solution. *Management Science*, 44(5):673–683, 1998.

A. F. Perold. Large scale portfolio optimization. 30(10):1143–1160–91, 1984.

RiskMetrics. Riskmetrics technical document. Technical Report 4, JP Mogan/Reuters, 1996.

S. Russell and P. Norvig, editors. *Artificial Integlligence - A Modern Approach*. Prentice Hall Series in Artificial Intelligence, 1995.

F. W. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.

D. X. Shaw, S. Liu, and L. Kopman. Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. *Optimization Methods Software*, 23(3):411–420, 2008.

H. H. Shek. Modeling order arrivals with bivariate hawkes process. Working Paper, Stanford University, 2007.

H. H. Shek. Modeling high frequency market order dynamics using self-excited point process. 2010. Working Paper, Stanford University.

N. Shephard and K. K. Sheppard. Realising the future: Forecasting with high frequency based volatility (HEAVY) models. *Forthcoming in Journal of Applied Econometrics*, 2010.

Y. Simaan. Estimation risk in portfolio selection: the mean variance model versus the mean absolute deviation model. *Management Science*, 43(10):1437–1446, 1997.

M. G. Speranza. A heuristic algorithm for a portfolio optimization model applied to the milan stock market. *Computers and Operations Research*, 23(5):433–441, 1996.

M. Takahashi, Y. Omori, and T. Watanabe. Estimating stochastic volatility models using daily returns and realized volatility simultaneously. *Computational Statistics and Data Analysis*, 53: 2404–2406, 2009.

P. R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6:324–342, 1969.

Y. Zeng. A partially observed model for micromovement of asset prices with bayes estimation via filtering. *Journal of Mathematical Finance*, 13(3):441–444, 2003.

Y. Zeng. Estimating stochastic volatility via filtering for the micromovement of asset prices. *IEEE Transaction on Automatic Control*, 49(3):338–348, 2004.

L. Zhang. Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli*, 12:1019–1043, 2006.