ABSTRACT

WENG, QIFENG. Three Essays on Financial Econometrics and Patents. (Under the direction of Denis Pelletier and Zachary Brown.)

The dissertation contains three chapters. First two chapters are about financial econometrics. The last chapter is on patents.

Chapter 1: The nonparametric theoretical work of Li et al. (2013) shows that the realized volatility estimator is asymptotically biased under the assumption that observation times are endogenous. Motivated by this finding, we develop a model for irregularly spaced returns where durations, the time span between two consecutive transactions, are endogenous. The model contains a bivariate Ornstein-Uhlenbeck (OU) process that jointly models equity latent volatility and trading intensity. Together with two other processes modeling trading prices and durations, the time endogeneity is captured by our model. The mode has a linear state space representation. We obtain an asymptotically unbiased volatility estimator via the Kalman filter. Estimates from Microsoft (MSFT) high-frequency trading data reveal a positive time endogenous effect between durations and logarithmic prices.

Chapter 2: Implied volatility derived from the Black-Scholes model often performs poorly in practice due to the assumptions of underlying asset volatility being constant over the option lifetime and normality of returns. To circumvent this issue, Dumas, Fleming, and Whaley (1998) developed a Deterministic Volatility Function (DVF) model. They found that by incorporating implied volatility with option moneyness and maturity significantly reduces option pricing errors. On the other hand, in Chapter 1 we developed the so-called endogenous time volatility estimator based on high-frequency equity trading data. Inspired by their work, we investigate the benefits of integrating the underlying asset endogenous time volatility to the DVF framework. By doing so, we find that DVF with endogenous time volatility factor helps a little regarding in-sample fit. However, the original DVF model still beats our proposed model in terms of out-of-sample forecast performance.

Chapter 3: Decision makers aiming to promote innovation often care about the connections between industrial innovation and economic development. Analysis of industry-level patent statistics can shed light on these connections. Concordance systems for assigning patents to industries are necessary for such analysis. This paper uses patent data matched at the firm level to investigate the accuracy and precision of prevailing technology-industry concordances and proposes new concordances. In contrast to previous concordances, which link patents to specific industries using textual analysis of patent titles and abstracts and official industry descriptions, we use firm-level 'microdata' from the OECD to match patents to industries via the identified economic sectors of patent applicants. Using this applicant sector matching, we examine previous concordances via binary probability regression models. We propose two new concordances using applicant sector matching, namely the Raw OECD Concordance (ROC) and the OECD Bayesian Averaged Concordance (OBAC). The ROC concordance is based on simple conditional frequencies, whereas the OBAC uses Bayesian model averaging to combine multiple regression models using technology classifications to predict industry membership. We find that text-based concordances poorly predict applicant sector matches, and that our OBAC concentrates more of its probability mass among the most likely industries, as compared to ROC. This research therefore yields a new concordance for policy analysis, as well as introduces a new method (BMA) that can improve existing concordances.

Three Essays on Financial Econometrics and Patents

by
Qifeng Weng

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Economics

Raleigh, North Carolina

2016

APPROVED BY:

Walter Thurman	Barry Goodwin
Denis Pelletier Co-chair of Advisory Committee	Zachary Brown Co-chair of Advisory Committee

DEDICATION

To my grandparents and parents.

BIOGRAPHY

The author was born in Wuxi, China. He received his bachelor's degree in mathematics from Shanghai University in 2010. Later, he spent six years in NC State University where he completed his Ph.D. in economics. His research interests are financial econometrics, patent research, and applied econometrics.

ACKNOWLEDGEMENTS

I am deeply indebted to several people. First and foremost, I would like to express my gratefulness to my advisors, Dr. Denis Pelletier and Dr. Zachary Brown. Without their generous help, these papers will not be possible. Your advice on both research as well as on my career have been priceless. I would also like to thank my committee members, Dr. Barry Goodwin and Dr. Walter Thurman for your direction and invaluable advice.

Second, I thank my grandparents, my parents, and my family for their financial support and guidance along this marathon.

I also thank all my friends for making the whole process simpler, meaningful and less struggling.

Thank you all, sincerely.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Returns, Durations and Time Endogeneity

1.1 Introduction

A huge body of literature on time-varying volatility estimation has been developed during the past few decades due to its crucial role in option pricing, efficient portfolio allocation, and risk management. As the availability of high-frequency, or tick-by-tick, trading data has increased, the literature has evolved from parametric discrete time models estimated with daily or lower frequency data, to continuous time parametric or nonparametric models.

Discrete time parametric models take two main forms. The autoregressive conditional heteroscedasticity (ARCH) model was developed by Engle (1982). Its generalized version, the GARCH model, was proposed by Bollerslev (1986). The competitive alternative to the ARCH and GARCH models is often referred as the stochastic volatility (SV) model (see Hull and White, 1987; Kim et al., 1998). However, this class of models requires regularly spaced data (fixed time intervals for transactions). The requirement of regularly-spaced data is a drawback when dealing with high-frequency data, since valuable information can be lost when temporally aggregating millisecond data to a coarser time grid (see Engle, 2000). Other challenging issues, like the

difficulty of capturing long-memory dependencies and handling large dimensional systems, make the discrete time model models ill-suited for accommodating high-frequency data (see Bollerslev, 2001).

Nonparametric methods dominate the recent continuous time volatility literature. A standard approach is to model the asset price in continuous time with an Itô process $d \ln S_t = \mu_t dt + \sigma_t dW_t$, where $\ln S_t$ is the logarithmic price, μ_t is a drift term, σ_t is instantaneous volatility, and dW_t is a stochastic Wiener term, whose increments over a period of time Δt follow a normal distribution, $N(0, \Delta t)$. The process is observed at times t_i, $i = 0, 1, \cdots, N$ in the time interval $[0, 1]$. Thus the increment of $\ln S_t$ at the observation time t_i is defined as $\Delta X_i \equiv \ln S_{t_i} - \ln S_{t_{i-1}}$. By design, we see that the difference of log-prices, ΔX_i, is also the continuously compounded rate of return of the i-th trade. Within the same framework, additional terms (or processes) can be added to the Itô process to encompass unique features of high-frequency data, including price jumps and microstructure noise.

To evaluate the volatility of a financial asset, a natural way of defining the cumulative variance of asset prices over a period of time is simply to integrate instantaneous volatility over the time interval. It is called integrated volatility (IV) in the literature, defined as: $\text{IV} \equiv \int_0^T \sigma_t^2 dt$. Unfortunately, the instantaneous volatility is a latent variable. Searching for better volatility proxies has been an ongoing challenge in financial econometrics research.

Jacod and Protter (1998) laid the statistical theory foundation in probability theory by proving that for an Itô process, Realized Variance (RV), defined as $[X, X]_t = \sum_{t_i \leq t} (\Delta X_i)^2$, asymptotically converges to the IV of the process. Barndorff-Nielsen and Shephard (2001, 2002) introduced this theory to financial econometrics. Since then, more than 400 nonparametic volatility estimators of IV have been developed (Liu et al., 2015). Several of these estimators have proven superior to the rest from various perspectives. The first estimator, 5-minute Realized Volatility ($\text{RV}_{5\text{min}}$), is frequently used due to its ease of implementation (see Andersen and Benzoni, 2009, for details). Computing $\text{RV}_{5\text{min}}$ requires aggregating high-frequency data into 5 minute intervals and then summing up the squared returns at each 5-minute time interval: $\text{RV}_{5\text{min}} \equiv \sum (\Delta X_i^{5\text{min}})^2$,

where $\Delta X_i^{5\text{min}}$ is the log-return of the ith interval on a consecutive 5-minute time grid. However, in order to be a unbiased estimator, $\text{RV}_{5\text{min}}$ depends on the strong assumption that the market is frictionless and arbitrage-free. Additionally, the literature has dozens of jump-robust estimators of IV with many interesting insights behind their construction. Bipower Realized Volatility (BPV) (see Barndorff-Nielsen and Shephard, 2004) and Truncated Realized Volatility (TRV) (see Mancini, 2009) are representative choices. The intuition behind BPV is that jumps only occur infrequently. Instead of summing over the squared one-period price return, they sum over the product of absolute returns from two neighboring periods. By assumption, one jump would most likely occur in only one of the adjacent periods. At the limit, the cross-product efficiently kills the jumps. The idea behind TRV is to filter out jumps based on sophisticatedly selected thresholds. Another disturbance to the price process is market microstructure noise including bid-ask spread, discreteness of price changes, gradual response of prices to a block trade, strategic components of the order flow, inventory control effects, data recording errors, rounding effects, and others (see Aït-Sahalia and Yu, 2009). Realized Kernel (RK) (see Barndorff-Nielsen et al., 2011) serves as a robust estimator to both jumps and microstructure noises with substantially improved precision. Loosely speaking, RK is the summation of realized autocovariances weighted by a non-flat-top Parzen kernel. Barndorff-Nielsen et al. (2011) defined RK as: $\text{RK} \equiv \sum_{h=0}^{H} k(\frac{h}{H})\Gamma_h$, where $\Gamma_h = \sum_{j=h+1}^{n}(\Delta X_j \Delta X_{j-h})$ for $h \geq 0$, which is the autocovariance of high-frequency price returns. $k(\cdot)$ is the Parzen kernel, H is the bandwidth.

High-frequency data usually contains the following information: a time stamp indicating when the transaction occured with up to millisecond accuracy, transaction price, ask price, bid price, and the transaction volume. One important advantage of high-frequency data compared to lower frequency data is the inclusion of irregularly spaced transaction time. Intuitively, informed market participants would have clustered trading behavior when news hits the market and flat trading behavior when no news hits the market. Both clustered and flat trades break the demand-supply balance in the market and affect the volatility of the financial products (see Easley and O'Hara, 1992). Hence, by including transaction duration, the time span between two

consecutive transactions, one can generate a better volatility estimator. Following this idea, single factor duration volatility models and joint models between duration and price returns have been developed during the past couple of decades. The autoregressive conditional duration (ACD) model was proposed by Engle and Russell (1998). Another single factor model, the Stochastic Conditional Duration (SCD) model, was put forth by Bauwens and Veredas (2004). Ghysels et al. (2004) developed a joint model for duration and risk. Renault et al. (2013) built a passage hitting time model based on Abbring (2012) mixed hitting-time model that serves as a generalized ACD model and SCD model. Moreover, Pelletier and Zheng (2013) and Wei and Pelletier (2015) constructed a bivariate Ornstein-Uhlenbeck (OU) process for duration and price return that lays the foundation for this paper.

Li et al. (2014) made an interesting discovery that under the assumption that there exists an instantaneous correlation between duration and price return, the realized volatility is an asymptotically biased estimator of integrated volatility. They documented that such time endogeneity between the transaction durations and the price process exists in financial data. Furthermore, they brought tricity, defined as $[X, X, X]_t = \sum_{t_i \leq t} (\Delta X_i)^3$, into the picture for the first time in the nonparametric continuous time volatility literature. Before their paper, tricity was considered a non-contributor to volatility estimation since it converges to zero asymptotically. However, under the time endogeneity assumption, non-zero tricity becomes the source of RV's estimation bias. They established the central limit theory for the realized volatility in a general endogenous time setting.

Inspired by the discovery of Li et al. (2014), we modify Pelletier and Zheng (2013) joint model between duration and price return to capture the endogenous time effect. First, the log-price dynamics follow a Wiener process. The second layer of the model is a bivariate Ornstein-Uhlenbeck (OU) process that simultaneously captures two latent variables: transaction intensity and volatility. Afterward, the transaction duration process follows an exponential distribution conditional on trade intensity. Finally, we model the instantaneous dependence between duration and price return with a Gaussian copula function. Our model strictly differs from examples listed

in Li et al. (2014). Compared to nonparametric methods for estimation of integrated volatility, our model has the advantage of producing both estimation and prediction of latent volatility and duration. The flexibility of our model also allows for the future incorporation of market microstructure noise and price jumps.

The rest of this paper is organized as follows. In Section 2, we specify the joint model of price return and duration with an endogenous time setup. In Section 3, we give a detailed discussion of the implementation of QMLE via the Kalman filter for model estimation. Section 4 contains a summary of an empirical study on MSFT high-frequency trading data. Section 5 presents the simulation experiment design to investigate our model's ability to capture time endogeneity. Section 6 concludes.

1.2 Model Specification

In this section, we will define the stochastic process of durations and returns within an endogenous time setting. Our model consists of four layers. Loosely speaking, the first layer is a logarithmic price process linking observable prices with latent instantaneous volatilities. The second layer is a dynamic process for the durations. Durations are generated by an exponential distribution with the conditional mean equal to the inverse of the latent trade intensity. The third layer of our model is a bivariate Ornstein-Uhlenbeck (OU) process for two latent variables: the logarithmic trade intensity and the logarithmic instantaneous volatility. The last layer allows for time endogeneity via a Gaussian copula. The details of the model are presented below.

1.2.1 A Wiener Process for Logarithmic Price

The dynamics of logarithmic price S_t at time t, are defined as a simple Wiener process $W_{0,t}$ times a latent instantaneous volatility σ_t.

$$d \ln S_t = \sigma_t dW_{0,t}. \tag{1.1}$$

As suggested by Engle (2000) and Renault and Werker (2004), a Wiener process itself is powerful enough to accommodate high-frequency data. Therefore, we ignore the drift term for simplicity in our model. Regarding other ingredients common for a price return process, *e.g.*, price jumps or market microstructure noise, we will include them in future research.

1.2.2 A Poisson Process for Durations

Here, we introduce some general point process theory, which leads directly to the exponential distribution on which the duration dynamic is built (see Chen et al., 2013; Kiefer, 1988). A point process on $(0, \infty)$ is a sequence of nonnegative random variables $\{T_i\}_{i \in 1,2,\cdots}$ defined on a probability space (Ω, F, P), satisfying $0 < T_1 < T_2 < \cdots$, where T_i is the instant of the i-th occurrence of an event. More specifically, in our model, the arrival time for the i-th transaction or trade event is modeled as T_i.

λ_{t_i} is the stochastic intensity characterizing a point process. Intuitively, λ_{t_i} can be interpreted as the instantaneous probability that the i-th trade occurs at time t. It is also called the Hazard function. It is defined as:

$$\lambda_{t_i} = \lim_{\Delta t \to 0} \left(\frac{1}{\Delta t} Prob[N(t + \Delta t) - N(t) = 1 | F_{t-}] \right),$$

where $N(t) = \sum_{i \geq 1} 1(T_i \leq t)$ is a counting process, summing up the total number of events up to and including time t. F_{t-} is a sub-sigma field of F, which can be interpreted as all available information up to time t.

Naturally, we define the time span between the $(i-1)$-th trade and the i-th trade as duration d_i:

$$d_i \equiv T_i - T_{i-1}, \tag{1.2}$$

Chen et al. (2013) concluded that, requiring only weak regularity conditions, one will have a general result $di \cdot \lambda_{t_i} \sim$ i.i.d. Exp(1). We assume d_i follows a conditional exponential distribution

with mean $\lambda_{t_{i-1}}^{-1}$:

$$f(d_i|\lambda_{t_{i-1}}) = \frac{1}{\lambda_{t_{i-1}}} \cdot \nu_i \quad \text{with} \quad \nu_i \sim \text{Exp}(1). \tag{1.3}$$

1.2.3 OU Process for Intensity and Volatility

Following Pelletier and Zheng (2013), we consider the third layer of our model for two latent variables: the logarithmic trade intensity λ_t and the logarithmic instantaneous variance σ_t^2 following a bivariate Ornstein-Uhlenbeck (OU) process. Appealing features of an OU process are: 1) it is mean-reverting; 2) it has a closed form solution; 3) it can be interpreted as the continuous time version of an AR(1) process. The bivariate dynamic process is denoted as:

$$dX_t = A(\mu - X_t)dt + SdW_{-0,t} \tag{1.4}$$

where:

$$X_{t_i} = \begin{bmatrix} \ln \lambda_t \\ \ln \sigma_t^2 \end{bmatrix}, \quad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad S = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \quad W_{-0,t} = \begin{bmatrix} W_{1,t} \\ W_{2,t} \end{bmatrix}.$$

The vector X_t contains two latent variables: the logarithmic trade intensity λ_t and the logarithmic instantaneous volatility σ_t. The coefficient matrix A is called the transition matrix, which controls the persistence of X_t during the evolution of the process. The vector μ is the unconditional mean of X_t. The matrix S serves as the variance matrix for a Wiener process. $W_{-0,t}$ refers to the Wiener terms in this OU process. Its subscript -0 indicates that $W_{-0,t}$ differs from $W_{0,t}$.

1.2.4 Time Endogeneity Setting

For the purpose of introducing the endogenous time effect to our model and linking aforementioned stochastic processes together, we impose contemporaneous correlations, denoted as Γ, in our model.

Sklar's theorem (1959) states that the joint distribution function can be written as a unique

copulas function of random variables' marginal distributions under the condition that all marginal distributions are continuous. We choose the multivariate Gaussian copulas to implement these correlations due to its handy properties (Žežula, 2009): (i) Gaussian copulas allow for any marginal distribution and any positive definite correlation matrix. (ii) Gaussian copulas consider only pairwise dependence between the individual components of a random variable. Although the second property seems to be a drawback, it accommodates time endogeneity in our model perfectly.

The Gaussian copulas density function in our case is defined as:

$$c(x) = \frac{1}{|\Gamma|^{1/2}} \exp\left[-\frac{1}{2}u'(\Gamma^{-1} - I_4)u\right] \tag{1.5}$$

where:

$$x = \begin{bmatrix} \Delta W_{0,t_i}/\sqrt{d_i} \\ \Delta W_{1,t_i}/\sqrt{d_i} \\ \Delta W_{2,t_i}/\sqrt{d_i} \\ \nu_i \end{bmatrix}, \quad \Gamma = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_4 \\ \rho_1 & 1 & \rho_3 & 0 \\ \rho_2 & \rho_3 & 1 & 0 \\ \rho_4 & 0 & 0 & 1 \end{bmatrix}, \quad u = \begin{bmatrix} \Delta W_{0,t_i}/\sqrt{d_i} \\ \Delta W_{1,t_i}/\sqrt{d_i} \\ \Delta W_{2,t_i}/\sqrt{d_i} \\ \Phi^{-1}(F_{\exp}\nu_i) \end{bmatrix}.$$

$\Delta W_{0,t_i}/\sqrt{d_i}$, $\Delta W_{1,t_i}/\sqrt{d_i}$, and $\Delta W_{2,t_i}/\sqrt{d_i}$ are discretized per trade Wiener increments of process (1.1) and (1.4) respectively. They all follow a standard normal distribution. ν_i is the innovation term of equation (1.3). The normal quantile function u_i is given by: $u_i = \Phi^{-1}(F_i(x_i))$, where $F_i(\cdot)$ is the corresponding cumulative density function (cdf) of x_i. In particular, $F_{\exp}$ is the cumulative function of d_i's conditional exponential distribution. I_4 denotes a 4×4 identity matrix. To generate discretized innovation terms of equation (1.1), (1.3) and (1.4), we first use the Gaussian copulas density function with the imposed correlation matrix Γ to draw a 4×1 random variable vector, all ranging from $[0, 1]$, then take their related inverse cdf functions.

The specifications of ρ's in Γ are defined as follows: ρ_1 is the instantaneous correlation between Wiener terms in $\ln S_{t_i}$ and $\ln \lambda_{t_i}$ dynamics. It is defined as:

$$\rho_1 \equiv < W_{0,t}, W_{1,t} > . \tag{1.6}$$

ρ_2 is the parameter controlling the leverage effect. It is defined as the instantaneous correlation between Wiener terms in $\ln S_{t_i}$ and $\ln \sigma^2_{t_i}$ dynamics, denoted as:

$$\rho_2 \equiv < W_{0,t}, W_{2,t} > . \tag{1.7}$$

ρ_3 is the instantaneous correlation between Wiener terms in $\ln \lambda_{t_i}$ and $\ln \sigma^2_{t_i}$ dynamics. It is defined as:

$$\rho_3 \equiv < W_{1,t}, W_{2,t} > . \tag{1.8}$$

ρ_4 brings the time endogeneity effect to the model. If it is set to be zero then the model is counted for a exogenous time case. It is denoted as:

$$\rho_4 \equiv \mathrm{Corr}(\Phi^{-1}(F_{\exp}(\nu_i)), \Delta W_{0,t_i}/\sqrt{d_i}). \tag{1.9}$$

1.3 Estimation

In this section, we discretize our continuous time model, rewrite the model into state space representation, and employ the Kalman filter to do Quasi-Maximum Likelihood Estimation (QMLE) using similar techniques to those in Harvey et al. (1994); Pelletier and Zheng (2013). The log-squared return transformation that we take to derive the state space representation renders correlation parameters unidentified. We use the sign of the return to recover these parameters.

1.3.1 Model Discretization and Linearization

The dynamics of logarithmic prices, latent volatility, and intensity are modeled in continuous time. However, even high-frequency data is not recorded frequently enough to fully satisfy the continuous time process assumptions. Our estimation strategy is to conduct QMLE via the Kalman filter on the discretized model. The bivariate OU process in our model has an analytical solution in the form of a discrete time AR(1) model (see Jacod and Protter, 2011; Pelletier and Zheng, 2013, for details). Using this solution will not cause any discretization error. However, the

continuous time process of logarithmic prices does not have a closed form solution. The Euler scheme is employed to discretize the process. The aforementioned point process of trade durations, equation (1.2), is used as the discretization time gird. Notice that utilizing any discretization schemes, *e.g.*, the Euler scheme, will introduce discretization errors. However, the discretization errors are negligible since the time intervals are very small. Details of the model discretization are presented below.

The logarithmic price process, equation (1.1) is discretized as:

$$\ln S_{t_{i+1}} - \ln S_{t_i} = \sigma_{t_i} \cdot \zeta_{t_{i+1}} \quad \text{with} \quad \zeta_{t_{i+1}} \sim N(0, d_{i+1}).$$

We define $\ln S_{t_{i+1}} - \ln S_{t_i}$ as the log-return $y_{t_{i+1}}$ of $(i+1)$-th trade:

$$y_{t_{i+1}} \equiv \ln S_{t_{i+1}} - \ln S_{t_i}. \tag{1.10}$$

Combining these two equation, we have:

$$y_{t_{i+1}} = \sigma_{t_i} \cdot \zeta_{t_{i+1}} \quad \text{with} \quad \zeta_{t_{i+1}} \sim N(0, d_{i+1}). \tag{1.11}$$

The solution of the OU process, equation (1.4), is given by:

$$X_{t_{i+1}} = [I_2 - \text{expm}(-Ad_{i+1})]\mu + \text{expm}(-Ad_{i+1})X_{t_i} + \Omega_{t_{i+1}}, \tag{1.12}$$

where the innovation term

$$\Omega_{t_{i+1}} = \int_0^{d_{i+1}} \text{expm}(A[\mu - t_{i+1}])SdW_{-0,\mu} \quad \text{follows} \quad \Omega_{t_{i+1}} \sim N(0, \Sigma^{\Omega}_{t_{i+1}}).$$

where $\text{vec}(\Sigma^{\Omega}_{t_{i+1}}) = (A \oplus A)^{-1}(I_4 - \text{expm}(-[A \oplus A]d_{i+1}))\text{vec}(\Sigma^{\Omega})$, with $\Sigma^{\Omega} = S \begin{bmatrix} 1 & \rho_3 \\ \rho_3 & 1 \end{bmatrix} S$.

Moreover, expm$(\cdot)$ is the matrix exponential function, defined as $\text{expm}(A) \equiv \sum_{k=0}^{\infty} \frac{A^k}{k!}$.

The covariance between the innovation term in the price process and the two in the OU process is:

$$\begin{aligned}
\Sigma_{t_{i+1}}^{\zeta,\Omega} &= \text{Cov}(\zeta_{t_{i+1}}, \Omega_{t_{i+1}}) \\
&= \text{Cov}(\int_{t_i}^{t_{i+1}} \text{expm}(A[u - t_{i+1}])SdW_{-0,u}, \int_{t_i}^{t_{i+1}} dW_{0,u}) \\
&= \int_{t_i}^{t_{i+1}} \text{expm}(A[u - t_{i+1}])S \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} du \\
&= A^{-1}[I_2 - \text{expm}(-Ad_{i+1})]S \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}. \qquad (1.13)
\end{aligned}$$

1.3.2 Kalman Filter

We follow the Kalman filter documented in De Jong (1991) and De Jong and Shephard (1995). This version of the Kalman filter allows the innovation terms from the state equation and the observation equation to be correlated. The state space representation is given by the following system of equations:

Observation Equation: $$y_t = X_t\beta + Z_t\alpha_t + G_t u_t;$$

State Equation: $$\alpha_{t+1} = W_t\beta + T_t\alpha_t + H_t u_t.$$

where u_t are independent $N(0, \sigma^2 I)$ variables. β, Z_t, G_t, T_t, and H_t are parameters of interest. To estimate the state space model, they implements the following Kalman filter to record e_t, D_t

and K_t:

Innovation: $$e_t = y_t - X_t\beta - Z_t a_t$$

Innovation Covariance: $$D_t = Z_t P_t Z_t' + G_t G_t'$$

Kalman Gain: $$K_t = (T_t P_t Z_t' + H_t G_t') D_t^{-1}$$

$$a_{t+1} = W_t\beta + T_t a_t + K_t e_t$$

$$P_{t+1} = T_t P_t L_t' + H_t J_t'$$

$$L_t = T_t - K_t Z_t$$

$$J_t = H_t - K_t G_t$$

where $a_1 = W_0\beta$ and $P_1 = H_0 H_0$. As the by-product of Kalman Filter, the log-likelihood function based on $Y = \{y_1, y_2, \cdots, y_t\}$ is:

$$\ln L(Y) = -\frac{1}{2}\{\sum_{t=1}^{T} \ln|D_t| + \sum_{t=1}^{T} e_t' D_t^{-1} e_t\}$$

1.3.3 State Space Representation

We follow Harvey et al. (1994) and Pelletier and Zheng (2013) to derive the state space representation of our model. First, we take the logarithm of squared returns in equation (1.11) to get:

$$\ln(y_{t_{i+1}}^2) = \ln(\sigma_{t_i}^2) + \ln(d_{i+1}) + \ln(\eta_{t_{i+1}}^2).$$

where $\eta_{t_{i+1}} \sim N(0,1)$. According to Abramowitz and Stegun (1965), we know that the mean of $\ln(\eta_{t_{i+1}}^2)$ is equal to -1.2704, and the variance is equal to 4.9348. We treat all innovation terms as though they were normally distributed to implement QMLE, thus assuming $\ln(\eta_{t_{i+1}}^2) \sim N(-1.2704, 4.9348)$. Hence, using the new notation ε_{t_i} to replace $\ln(\eta_{t_i}^2)$, we have

$$\ln(y_{t_{i+1}}^2) = \ln(\sigma_{t_i}^2) - 1.2704 + \ln(d_{i+1}) + \varepsilon_{t_{i+1}} \quad \text{with} \quad \varepsilon_{t_{i+1}} \sim N(0, 4.9348). \tag{1.14}$$

Taking the logarithm of equation(1.3), we have:

$$\ln(d_i) = -\ln(\lambda_{t_{i-1}}) + \ln \nu_i.$$

Likewise, the mean of $\ln \nu_{t_i}$ is equal to -0.5772 and the variance is equal to 1.6449. We approximate $\ln \nu_i \sim N(-0.5772, 1.6449)$. Rewriting it with the new notation ψ_{t_i}:

$$\ln(d_i) = -\ln(\lambda_{t_{i-1}}) - 0.5772 + \psi_{t_i} \quad \text{with} \quad \psi_{t_{i+1}} \sim N(0, 1.6449). \tag{1.15}$$

Therefore the observation equation of the state space representation is:

$$Y_{t_i} = \begin{bmatrix} -0.5772 \\ -1.2704 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} X_{t_i} + \begin{bmatrix} \psi_{t_i} \\ \varepsilon_{t_i} \end{bmatrix}, \tag{1.16}$$

where

$$Y_{t_i} = \begin{bmatrix} \ln(d_i) \\ \ln(y_{t_i}^2) - \ln(d_i) \end{bmatrix}, \quad X_{t_i} = \begin{bmatrix} \ln(\lambda_{t_i}) \\ \ln(\sigma_{t_i}^2) \end{bmatrix}.$$

The state equation of the state space representation is the same as equation (1.12):

$$X_{t_i} = [I_2 - \text{expm}(-Ad_i)]\mu + \text{expm}(-Ad_i)X_{t_{i-1}} + \Omega_{t_i}. \tag{1.17}$$

1.3.4 State Space Representation Conditional on the Sign of the Return

After transforming the innovation term in the price process ζ_{t_i}, which follows a normal distribution, into $\ln \zeta_{t_i}^2$, the correlations between $\ln \zeta_{t_i}^2$ and Ω_{t_i} will be zero, irrespective of the value of ρ_1 and ρ_2. The reason for the information loss in ρ_1 and ρ_2 is bacause of the symmetry of the normal joint distributions of $f(\zeta_{t_i}, \Omega_{t_i})$. As for the correlation between $\ln \zeta_{t_i}^2$ and $\ln \nu_{t_i}$, the sign of the correlation will be unidentified because of the symmetric parabolic shape of the covariance function between $\ln \zeta_{t_i}^2$ and $\ln \nu_{t_i}$ centerring around $\rho_4 = 0$ (see panel (1,1) in Figure 1.1). To illustrate this point, moments of $\ln \nu_{t_i}$ are plotted in Figure 1.1. Unconditional moments of $\ln \nu_{t_i}$

are in the first column of Figure 1.1. Clearly, ρ_4 is completely unidentifiable in all cases of $\ln \nu_{t_i}$ moments without conditioning on the signs of the returns. Notice that to generate these plots we condition on the signs of ζ_{t_i}, whose signs are the same as the corresponding returns. Figures plotted in the second column of Figure 1.1 are conditioning on positive ζ_{t_i} draws. The third column is conditional on negative ζ_{t_i} draws.

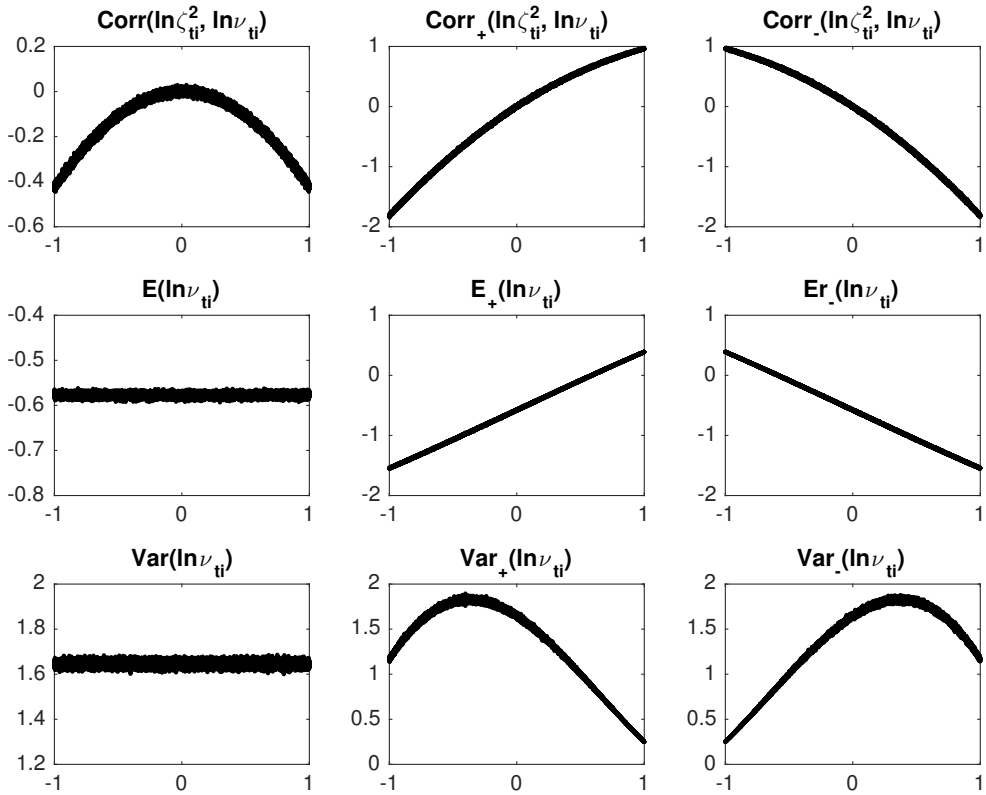


Figure 1.1: Moments of $\ln \nu_{t_i}$ Conditional on the Sign of the Returns

Harvey and Shephard (1996) solved a similar identification issue by conditioning on the signs of the returns. The following state space representation conditioning on return signs is similar to theirs. The subscript $*$ indicates the use of the sign. The observation equation conditioning on

the signs of the returns is:

$$Y_{t_i} = \begin{bmatrix} \mu^{\psi}_{*,t_i} \\ -1.2704 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} X_{t_{i-1}} + \begin{bmatrix} \psi_{*,t_i} \\ \varepsilon_{*,t_i} \end{bmatrix}, \tag{1.18}$$

where μ^{ψ}_{*} is the expected value of $\ln \nu_{t_i}$ conditional on the signs of returns. We do not have a closed-form solution for μ^{ψ}_{*}, but it can be pre-computed by simulation for all values of ρ_4. We also apply the same technique to the computation of $\sigma^2_{\psi,*}$. The state equation conditioning on the signs of the returns is:

$$X_{t_i} = [I_2 - \text{expm}(-Ad_i)]\mu + \mu^{\Omega}_{*,t_i} + \text{expm}(-Ad_i)X_{t_{i-1}} + \Omega_{*,t_i}, \tag{1.19}$$

where

$$\mu^{\Omega}_{*,t_i} = 0.7979 \cdot \Sigma^{\zeta,\Omega}_{t_i}(d_i)^{-0.5} \cdot \text{sign}(y_i) \quad \text{with} \quad \text{sign}(y_i) = \begin{cases} 1 & \text{if } y_i \text{ is positive} \\ -1 & \text{if } y_i \text{ is negative} \end{cases}.$$

The variance-covariance matrix of the innovation terms conditional on the signs of the returns is:

$$\text{Var}\begin{bmatrix} \psi_{*,t_i} \\ \varepsilon_{*,t_i} \\ \Omega_{*,t_i} \end{bmatrix} = \begin{bmatrix} \sigma^2_{\psi,*,t_i} & \Sigma^{\psi,\varepsilon}_{*,t_i} & \mathbf{0}_{1\times 2} \\ \Sigma^{\psi,\varepsilon}_{*,t_i} & \sigma^2_{\varepsilon,*} & \Sigma^{\varepsilon,\Omega\prime}_{*,t_i} \\ \mathbf{0}_{2\times 1} & \Sigma^{\varepsilon,\Omega}_{*,t_i} & \Sigma^{\Omega}_{*,t_i} \end{bmatrix}_{4\times 4}, \tag{1.20}$$

where

$$\begin{aligned} \Sigma^{\varepsilon,\Omega}_{*,t_i} &= 1.1061 \cdot \Sigma^{\zeta,\Omega}_{t_i}(d_i)^{-0.5} \cdot \text{sign}(y_i), \\ \Sigma^{\Omega}_{*,t_i} &= \Sigma^{\Omega}_{t_i} - \mu^{\Omega}_{*,t_i}\mu^{\Omega}_{*,t_i}{}'. \end{aligned}$$

Thus, the corresponding mappings to De Jong's Kalman filter algorithm are:

$$X_t\beta = \begin{bmatrix} \mu^{\psi}_{*,t_i} \\ -1.2704 \end{bmatrix}, \quad Z_t = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$W_t\beta = [I_2 - \text{expm}(-Ad_i)]\mu + \mu^{\Omega}_{*,t_i}, \quad T_t = \text{expm}(-Ad_i),$$

$$G_tG_t' = \begin{bmatrix} \sigma^2_{\psi,*,t_i} & \Sigma^{\psi,\varepsilon}_{*,t_i} \\ \\ \Sigma^{\psi,\varepsilon}_{*,t_i} & \sigma^2_{\varepsilon} \end{bmatrix}_{2\times 2}, \quad G_tH_t' = \begin{bmatrix} \mathbf{0}_{1\times 2} \\ \\ \Sigma^{\varepsilon,\Omega'}_{*,t_i} \end{bmatrix}_{2\times 2}, \quad H_tH_t' = \Sigma^{\Omega}_{*,t_i}.$$

1.3.5 Parameter Restrictions

Two parameter restrictions are imposed to ensure that the model is well defined. First, the persistence matrix A has to be positive definite (PD). This is required by the stationarity condition of the OU process. Second, the correlation matrix Γ has to be PD. A real square matrix A is PD if and only if the sum of a symmetric part of $(A + A')/2$ is PD.

1.3.6 Asymptotic Normality of QMLE Estimator

The QMLE estimator has a normal asymptotic distribution (see Hamilton, 1994, pp. 389):

$$\sqrt{T}(\hat{\theta} - \theta_0) \sim N(0, I^{-1}JI^{-1}), \tag{1.21}$$

where θ_0 is the true value of the parameter vector and $\hat{\theta}$ denotes its QMLE estimator. And

$$\begin{aligned} J &= \mathbf{E}_0\left[\left(\frac{\partial \ln L(\theta_0, y)}{\partial \theta_0}\right)\left(\frac{\partial \ln L(\theta_0, y)}{\partial \theta_0'}\right)\right], \\ I &= -\mathbf{E}_0\left[\frac{\partial^2 \ln L(\theta_0, y)}{\partial \theta_0 \partial \theta_0'}\right]. \end{aligned}$$

Their sample equivalents are consistent estimators of the Fisher information matrix J and the matrix I:

$$\begin{aligned}\hat{J} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial \ln L_t(\theta)}{\partial \theta}\right)\left(\frac{\partial \ln L_t(\theta)}{\partial \theta'}\right), \\ \hat{I} &= -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \ln L_t(\theta)}{\partial \theta \partial \theta'}.\end{aligned}$$

1.4 Empirical Results

1.4.1 Data Description

In this section, we apply our model to tick-by-tick transaction price data. We analyze a large liquid stock, Microsoft (ticker: MSFT), over the Jan. 16, 2012 to Feb. 15, 2012 period. We follow the data cleaning procedure described in Barndorff-Nielsen et al. (2009) to process the original data. We first remove all trading records outside the 9:30 am-4 pm EST window when the exchange is open. Second, we delete entries with corrected trades, whose correction indicator is different from zero, as well as those with an abnormal sale condition (where COND has a letter code other than 'E' and 'F'). Third, if multiple transactions have the same time stamp, we only keep one trade record with the volume-weighted price.

Soon after the standard data cleaning procedure, market microstructure noise remains a native feature of the high-frequency data. Examples of microstructure noise include bid-ask spread, discreteness of price changes, data recording error, and rounding effects. Hansen and Lunde (2006) point out that auto-correlated market microstructure noise contaminates observed returns, which biases the realized variance at ultra-high frequencies. A common treatment to mitigate microstructure noise in the realized volatility literature is to employ a calendar-time (every n minutes) sampling scheme under the assumption of observation times being exogenous. It is widely believed that the higher frequency of sampling schemes, the smaller magnitude of the noise, the lower noise-to-signal ratio. However, an obvious drawback of the calendar-time sampling scheme is the complete loss of temporal information in the raw data. For the purpose

of preserving temporal information and simultaneously diminishing market microsctructure noise, we employ the tick-time (every m trades) sampling scheme. Similar to the calendar-time sampling scheme, it combines returns and durations of all trades contained in each m-trade interval. Fukasawa (2010) proves that under the assumption of no rounding error and observation times being exogenous, the tick-time sampling scheme provides an appropriate RV estimator asymptotically converging to IV on price grids, including both transaction price grids and quote price grids.

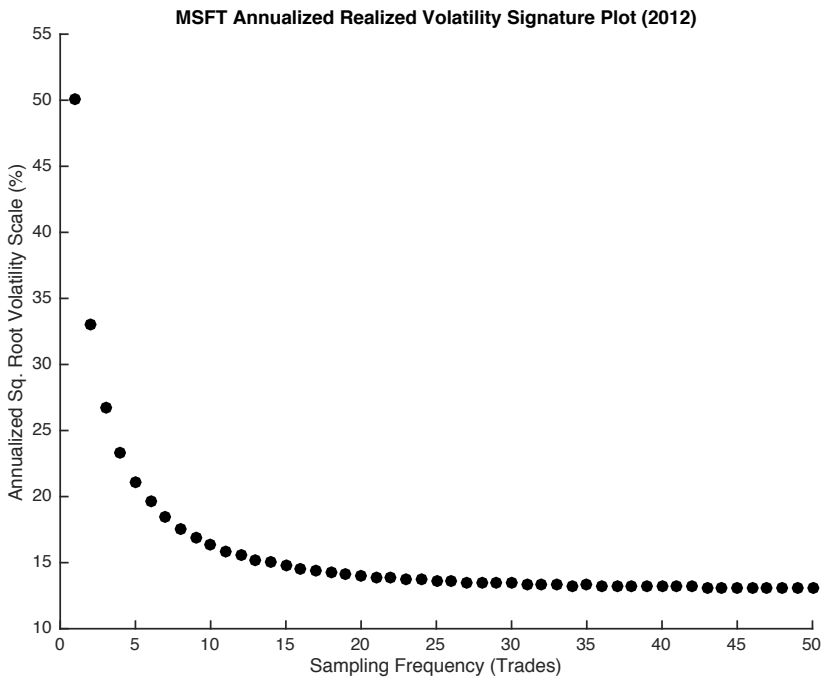


Figure 1.2: Annualized Realized Volatility Signature Plot

Figure 1.2 depicts the volatility signature for different tick-time sampling frequencies. The y-axis indicates the annualized square root of volatility. The x-axis refers to the sampling frequency by m-trade. One can observe that volatility starts to stabilize when sampling frequency is above the 15-trade mark, indicating that the microstructure noise is sufficiently diminished by

aggregating individual returns at a frequency of 15 trades or higher. From the estimation results in Table 1.2, we can see that the 'every 15 trades' sampling scheme leaves relatively too much microstructure noise, which results in our model being unable to capture persistence (elements of the persistence matrix A are not statistically significant). Taking all factors into account, we focus on the 'every 50 trades' tick-time sampling scheme for our empirical study.

1.4.2 Summary Statistics

Table **??** includes the descriptive statistics for the returns and durations over the 23 trading days of our sample. The number in the column name indicates the sampling frequency used for the analysis. For example, 'Return 50' represents the return series where an observation consists of the returns over 50 consecutive trades. The sample size of individual trades over the period is $T = 4,393,504$, roughly 190,000 transactions per trading day. In the case of the 50-trade frequency data, the sample size is $T = 87,753$, approximately 3,800 per day. The 50-trade sampling scheme still leaves a considerable sample size for estimation. We observe positive skewness and large excess kurtosis from the single-trade data. The 50-trade sampling scheme mitigates the issue. Both the return distribution and duration distribution have heavier tails than a normal distribution. It is worth noting that autocorrelation for the return series becomes negligible after order one in the single-trade frequency data set. However, for the 50-trade frequency data set, the return series appears uncorrelated. The duration series, on the other hand, shows increasing autocorrelation after being aggregated.

Table 1.1: MSFT Tick Data Summary Statistics (1/16/2012 - 2/15/2012)

	Return 1	Return 50	Duration 1	Duration 50
Obs.	4,393,504	87,753	4,393,504	87,753
Mean	0.000000	0.000002	0.436730	21.862079
Std.	0.00014	0.00032	1.03410	17.69725
Min	-0.01265	-0.00445	0.00300	0.16600
Max	0.01269	0.00332	30.88000	159.08000
Median	0.00000	0.00000	0.03000	17.40400
Skewness	0.01893	-0.02774	4.94251	1.47790
Kurtosis	94.446	7.741	41.910	6.014
ACF(1)	-0.37329	-0.04921	0.23906	0.48345
ACF(2)	-0.05921	0.01390	0.19626	0.38865
ACF(3)	-0.01881	0.00182	0.16909	0.36563
ACF(4)	-0.00624	0.00578	0.15169	0.35625
ACF(5)	-0.00300	-0.01215	0.13832	0.35701

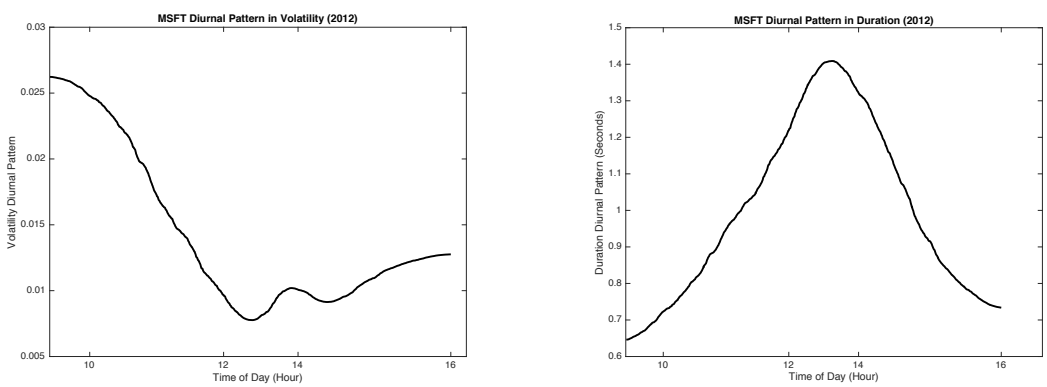
1.4.3 Diurnal Pattern

The diurnal pattern is a stylized feature of high-frequency financial data regarding volatility (see Andersen and Bollerslev, 1997) and trade intensity (see Engle and Russell, 1998). Traders have a tendency to trade more frequently near the opening of the market due to the overnight effect as well as near the closing time, and less frequently in the middle of the day. The more frequent trading at the beginning and end of the day generates higher volatility and a lower duration. Hence, it is important to remove diurnal patterns to have a correctly specified model. The simplest way to filter out this seasonality effect is to apply a deterministic function of time as the filter to the original return and duration data. Equation (1.22) denotes the diurnal-pattern-filtered

returns and durations:

$$
\begin{aligned}
y_i^* &= y_i/\sqrt{g_{d_i} g_{v_{i-1}}} \\
d_i^* &= d_i/g_{d_i}
\end{aligned}
\tag{1.22}
$$

where y_i^* and d_i^* are the adjusted duration and return respectively. g_{d_i} and g_{v_i} are the diurnal effects of duration and volatility at time t_i respectively. To obtain the function for g_{d_i}, we use the Nadaraya-Watson kernel estimator with a Gaussian kernel on the five-minute average durations in 2012 (up to the end of April in 2012) to generate the diurnal effect g_{d_i}. We have the same nonparametric kernel estimator applied to the average five-minute realized volatility in 2012 for g_{v_i} (see Campbell et al., 1997, p. 547-548). The diurnal pattern g_{d_i} (panel a) and g_{v_i} (panel b) are plotted in Figure 1.3.



(a) Volatility Diurnal Pattern

(b) Duration Diurnal Pattern

Figure 1.3: Diurnal Pattern Plots

1.4.4 Estimation Results

Our return and duration model is fitted to the deseasonalized return and duration data. To ensure that the logarithm of squared returns is well defined, we add a tiny number 10^{-4} to

those returns identically equal to zero. The aforementioned estimation procedure is developed for intraday estimation. To estimate the intra-day volatility for multiple days, we first calculate the likelihood of each individual day. We then maximize the sum of the daily log-likelihoods. Table 1.2 presents the estimation results from four different sampling frequency-by-trade schemes. The associated standard errors are shown in adjacent columns (see Hamilton, 1994, p.389).

Table 1.2: Estimation Results for MSFT Tick Data

	Individual Trade		15-Trade		50-Trade		100-Trade	
Parameters	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
A_{11}	0.53684	0.00001	0.00436	0.07959	0.04870	0.00945	0.02334	0.01364
A_{12}	-0.10675	0.00000	0.00594	0.01613	0.00146	0.00008	0.00057	0.00022
A_{21}	-0.99974	0.00001	0.07056	0.16153	0.08886	0.02794	0.02632	0.00033
A_{22}	18.23772	0.00002	0.33592	0.43609	0.04188	0.00199	0.00774	0.00000
σ_d	0.71824	0.00001	0.08642	0.06885	0.03791	0.01206	0.02983	0.00502
σ_v	56.51033	0.00001	4.79010	0.21194	1.35651	0.02616	0.49807	0.00925
μ_d	0.82357	0.00000	-2.34626	0.00788	-3.55998	0.03922	-4.28808	0.01232
μ_v	-15.01144	0.00002	-12.49792	0.33589	-12.09036	0.02097	-11.88744	0.00971
ρ_1	0.05720	0.00002	0.05296	0.38421	0.11060	0.02398	0.07236	0.01517
ρ_2	0.01046	0.00001	0.01360	0.20994	-0.04573	0.00545	-0.12913	0.01153
ρ_3	-0.43701	0.00001	0.97416	0.79751	0.98770	0.00814	0.97939	0.01232
ρ_4	-0.03252	0.00000	0.01513	0.16580	0.07309	0.01551	0.11891	0.02785

First, the persistence of the OU process is measured by $\text{expm}(-Ad)$. Larger values of A indicates less persistence in logarithmic intensity and logarithmic latent volatility. As the sampling frequency increases, all elements in the transition matrix A get smaller, and one observes a slowly decaying, persistent time path of the process. It is consistent with the common belief that volatility or trading intensity does not fluctuate dramatically within one trading day. This also supports the expectation that a lower sampling frequency mitigates the microstructure noise

effect. Second, an interesting finding is that for the single trade data case, A_{22} is estimated above 18, indicating a highly impersistent path for $\ln \sigma_t^2$. Third, since ρ_2 stands for the correlation between price and trading intensity, it is expected to be negative due to the leverage effect. The parameter ρ_2 is estimated as positive by lower frequency sampling schemes. Fourth, when employing a 15-trade sampling scheme, many parameters are statistically insignificant. All parameters are statistically significant for the case of 50-trade and 100-trade sampling schemes. Most importantly, as expected, the time endogeneity effect, captured by ρ_4, becomes more and more prominent changing from -3.3% to 11%, as sampling frequency decreases. This indicates that the empirical evidence strongly supports the existence of the endogenous time effect. For the 50-trade case, ρ_4 is 0.07, large enough to cause substantial inconsistency for the Realized Volatility and minor bias for the Realized Kernel estimator based on the Monte Carlo experiment showed in the next section.

1.4.5 Forecast

To demonstrate the forecasting accuracy performance of our model, we compare our model with a parsimonious AR(1) model. We next generate five intra-day volatility predictors from both models. To evaluate the forecasts, we choose the loss function family proposed in Patton (2011) to implement Diebold and Mariano (1995) and West (1996) (henceforth DMW) tests. Four volatility proxies are selected as benchmarks for the tests: the intra-day RK estimator, the daily squared return, the $\mathrm{RV}_{5\mathrm{min}}$, and the $\mathrm{RV}_{15\mathrm{min}}$. The null hypothesis of the DMW test is that the two models have the equivalent predictive accuracy. The loss function family suggested in Patton (2011) is given by:

$$L(\hat{\sigma}^2, h; b) = \begin{cases} \frac{\hat{\sigma}^{2b+4} - h^{b+2}}{(b+1)(b+2)} - \frac{\hat{\sigma}^2 - h}{b+1} \cdot h^{b+1} & \text{for } b \notin \{-1, -2\} \\ h - \hat{\sigma}^2 + \hat{\sigma}^2 \log \frac{\hat{\sigma}^2}{h} & \text{for } b = -1 \\ \frac{\hat{\sigma}^2}{h} - \log \frac{\hat{\sigma}^2}{h} - 1 & \text{for } b = -2. \end{cases}, \tag{1.23}$$

where an individual benchmark is denoted as h. $\hat{\sigma}^2$ refers to the volatility proxy generated by a candidate model. The loss function family nests two of the most widely used loss functions in the volatility forecasting literature: the mean squared error (MSE) loss and the quasi-likelihood (QLIKE) loss in the volatility forecasting literature. Up to additive and multiplicative constants, when $b = 0$ the loss function becomes MSE loss: $L(\hat{\sigma}^2, h) = (\hat{\sigma}^2 - h)^2$. The loss function becomes the QLIKE loss when $b = -2$. The particular interest in MSE and QLIKE loss functions is due to their robustness to noise in the volatility proxy. Furthermore, Patton and Sheppard (2009) pointed out that using QLIKE functions provides higher power to reject inferior estimators than MSE. However, notice that to apply the theory, the volatility proxy must be conditionally unbiased, while the forecasts do not have to be unbiased. In the case of time endogeneity, all of the volatility proxies that are applied in this test are not conditionally unbiased. The task remains for us to find the conditionally unbiased volatility proxies to properly employ the Patton (2011) loss function in the case of time endogeneity. One natural nonparametric candidate can be developed from Li et al. (2014). We leave this for future work.

For the forecasting exercise, we calculate 82 intra-day RK, daily squared return, $\mathrm{RV}_{5\mathrm{min}}$, and $\mathrm{RV}_{15\mathrm{min}}$ as the benchmarks based on MSFT data from Jan. 3rd to April 30th, 2012. We use the parameter values in Table 1.2 to simulate out-of-sample 5-day horizon $\hat{\mathrm{IV}}$ as the volatility forecasts from our model. To produce forecasts of the alternative AR(1) model, $\hat{\mathrm{RV}}_{5\mathrm{min}}$, we use 77 in-sample $\mathrm{RV}_{5\mathrm{min}}$ as training data. A t-statistic greater than 1.96 (the critical value for the 5% significance level) in absolute value indicates the rejection of the null of equal predictive ability. A positive t-statistic indicates that the test favors our model.

Table 1.3 reveals several important findings. First, both MSE loss and QLIKE loss reach the same testing results. Second, the two models are statistically significantly different in three cases of benchmarks: daily squared return, $\mathrm{RV}_{5\mathrm{min}}$, and $\mathrm{RV}_{15\mathrm{min}}$. Only in the daily squared return case, our model significantly outperforms the AR(1) model. It falls within our expectation that our model is beaten by the AR(1) model in the latter two cases. Because the training data used by AR(1) is $\mathrm{RV}_{5\mathrm{min}}$. A third interesting finding is that, in the Realized Kernel case, although the

test fails to reject the null hypothesis, the test still favors our model. From the simulation study in Section 5 we know that RK has a smaller bias from IV than RV when time is endogenous. Hence the result preferring our model is in line with the expectation of our model having a superior forecasting accuracy.

Table 1.3: Comparison of proposed model and AR(1) model forecasts

Loss Function	Realized Kernel	Daily Squared Return	5-min RV	15-min RV
$b = 1$	1.12	2.84	-12.79	-12.79
$b = 0$ (MSE)	1.23	2.82	-12.18	-12.00
$b = -1$	1.27	2.76	-11.53	-11.19
$b = -2$ (QLIKE)	1.11	2.60	-9.85	-9.93
$b = -5$	-0.57	0.56	-2.01	-2.23

It is easy for our model to beat an AR(1) model, since it is the simplest model capturing the time-varying intra-day volatility. As future work, we will compare our model with more sophisticated models, *e.g.*, an univariate autoregressive fractionally integrated moving average (ARFIMA) model (see Andersen and Bollerslev, 2003) or a popular heterogeneous autoregressive (HAR) forecasting model proposed by Corsi (2009). The ARFIMA model is effective in empirical modeling long memory features are present in the log-realized volatility. It will be interesting to see if our modeling would beat ARFIMA model when taking the endogenous time effect into account. The HAR model is a AR-type model with the feature of capturing its long memory properties. HAR model has been proven successful in various volatility forecasting applications.

1.5 Simulation Study

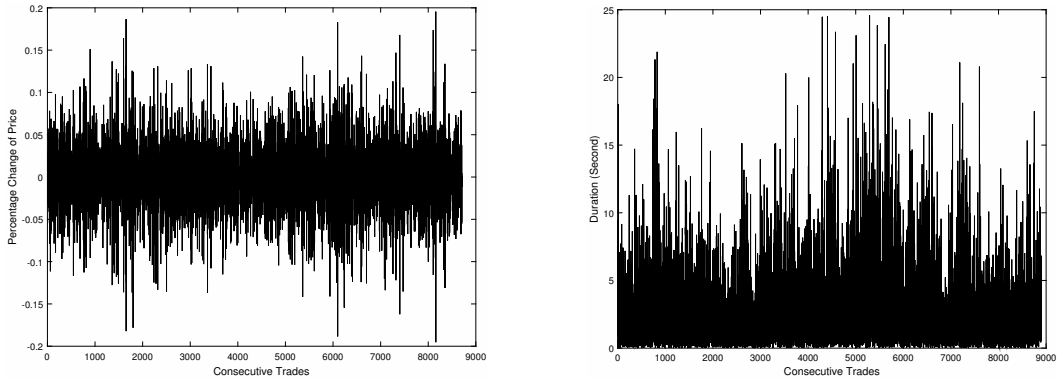
In this section, we illustrate the time endogeneity effect through simulations. The parameter values in Table 1.4 are used to carry out one particular simulation experiment. In each simulation

case, we independently generate 5,000 imaginary trading days. On each day, we simulated transactions for a total duration of 23,400 seconds. The U.S. stock market opens at 9:30 am and closes at 4 pm: a total of 6.5 hours, or 23,400 seconds. Our simulation fully mimics the real stock market operation. Notice that, in this particular case, we introduce time endogeneity by setting $\rho_4 = 0.2$.

Table 1.4: Simulation Parameter Values

Parameters	Values	Parameters	Values	Parameters	Values
ρ_1	0.8	a_{11}	0.0034	μ_1	-1
ρ_2	0	a_{12}	-0.0008	μ_2	-8.8
ρ_3	0	a_{21}	-0.0063	σ_1	0.02
ρ_4	0.2	a_{22}	0.0037	σ_2	0.02

Figure 1.4, panel (a) shows the price return dynamics in one specific trading day from our simulation case. Each vertical line indicates the change of the logarithmic price of one individual transaction, varying from -0.05% to 0.25%. Figure 1.4, panel (b) displays the trade durations from the same trading day as panel (a). One can observe that the durations between trades generated from our model are not constant, varying from 1 second to 25 seconds. This fits the real trading data for a stock with average liquidity.



(a) Simulated Return Plot

(b) Simulated Duration Plot

Figure 1.4: Model Simulation Plots

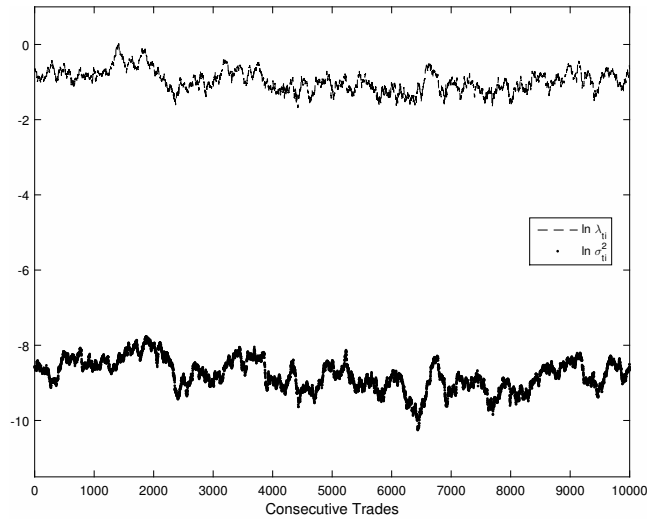


Figure 1.5: Latent Variables Evolution in the OU Process

Figure 1.5 depicts the trajectory of two latent variables: $\ln \lambda_{t_i}$ and $\ln \sigma^2_{t_i}$ during the same trading day. One can see that the OU process keeps these two variates fluctuating around an imaginary straight path, demonstrating its mean-reverting feature. Additionally, one can observe

that the two individual paths have similar patterns. Peaks and dips tend to occur at the same place. In fact, the persistence of these two latent variables is driven by the coefficient matrix A in equation (1.4). If one changes the values of A, the OU process could generate other types of behaviors, *e.g.*, the latent dynamics could evolve in opposite directions.

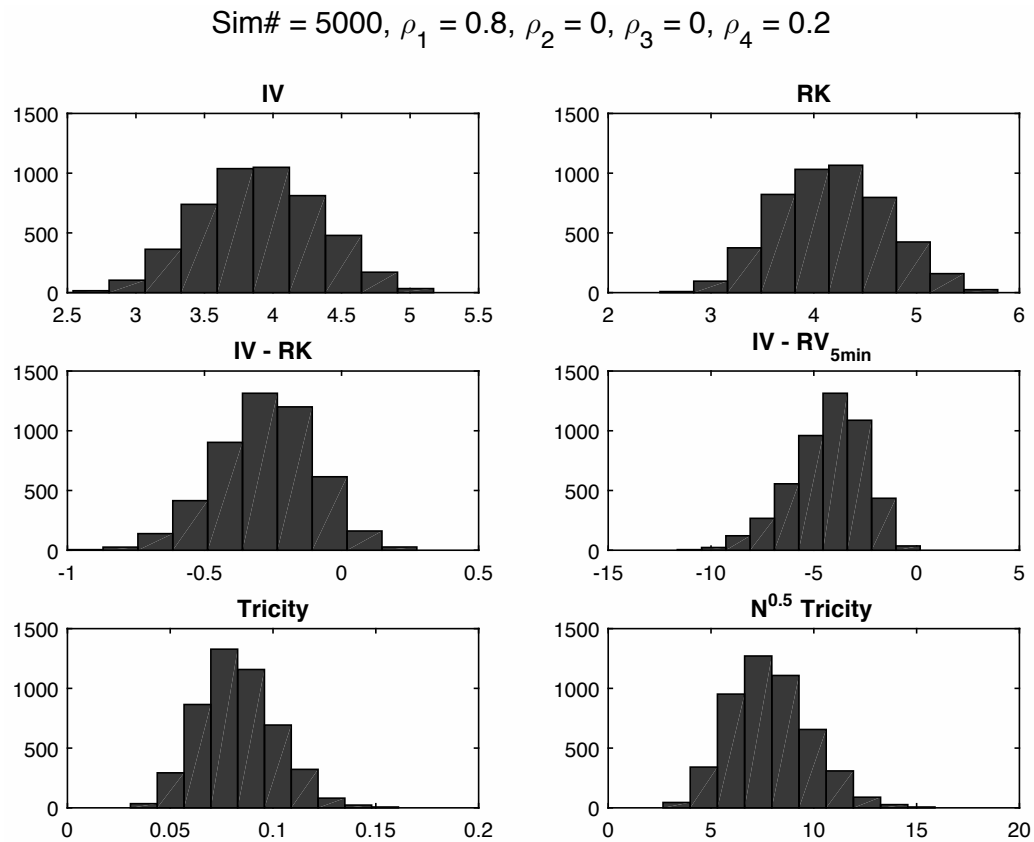


Figure 1.6: Volatility Proxy Performance Plots

Figure 1.6 displays the volatility analysis for all trading days from the simulation case. The IV histogram summarizes the spreads of all integrated volatility from 5000 iterations, centering around 4. The RK histogram plot (panel (1,2)) is beside IV's. As mentioned in the introduction section, RK is the best estimator of IV under the assumption that duration is exogenous. We can see that although the differences between IV and RK do not center around zero, the magnitude

of deviation is substantially smaller than the one created by $\mathrm{RV}_{5\mathrm{min}}$. This is direct evidence that RK is a better estimator than the commonly used $\mathrm{RV}_{5\mathrm{min}}$. Moreover, the simulation evidence proves that even the best estimator becomes a biased estimator in the case of time endogeneity. The tricity plot is interesting as well. It supports the theoretical finding of Li et al. (2014) that when time endogeneity effects exist, the source of estimation bias is from non-zero tricity. The last plot is the re-scaled tricity. In fact, Li et al. (2014) suggests that $\sqrt{N} \cdot$ Tricity contributes to the bias in the RV.

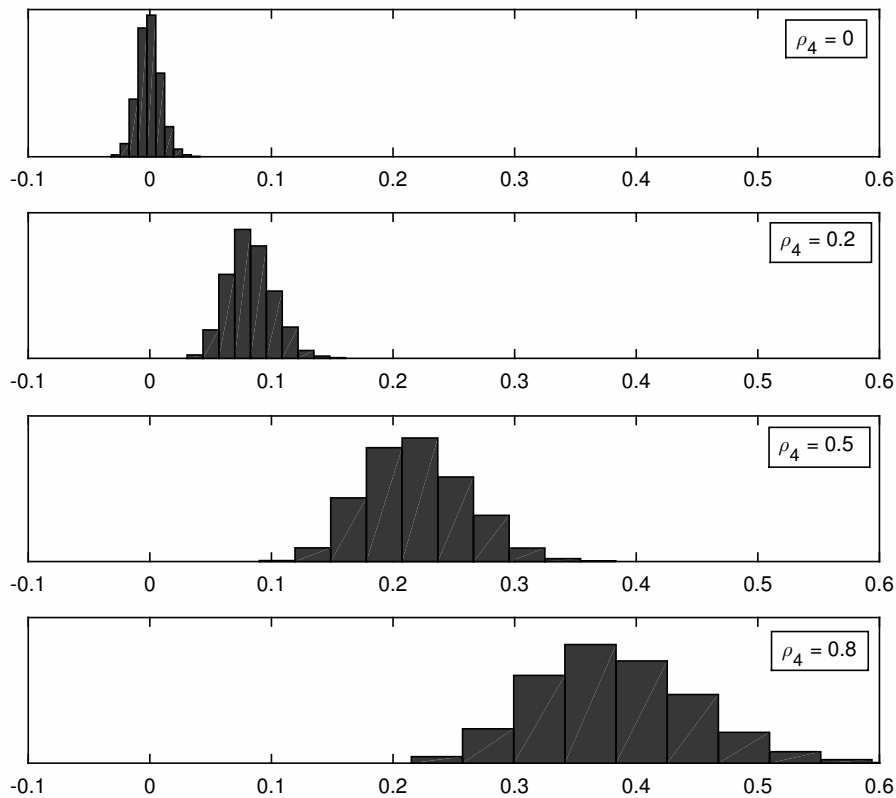


Figure 1.7: Tricity Deviation from Zero and Dispersion as Time Endogeneity Effect Grows

Figure 1.7 demonstrates the shape variations of Tricity spreads under the different values of ρ_4 when holding all other parameters fixed. Noticeably, the larger ρ_4, the further tricity deviates

away from zero and the greater variance it has. Hence, Figure 1.6 and Figure 1.7 directly support the point that our parametric model has the power of capturing endogenous time effects.

1.6 Conclusion

Li et al. (2014) define time endogeneity as a situation in which randomness in observation time matters because it implies a nonzero limit for tricity. Under the assumption of time being endogenous, conventional volatility proxies, *e.g.*, Realized Kernel and Realized Volatility, become asymptotically biased estimators. This paper introduces a model for irregularly spaced returns when durations are endogenous. The model contains a bivariate Ornstein-Uhlenbeck (OU) process that jointly models equity latent volatility and trading intensity. Together with two other processes of trading prices and durations, the time endogeneity is captured by our model. The mode has a linear state space representation. Hence, we implement QMLE along with the sign conditions. We filter the volatility with the Kalman filter to achieve an asymptotically unbiased volatility estimator when observation times are endogenous. Finally, our model demonstrates impressive prediction power through forecast accuracy tests.

CHAPTER 2

An Empirical Analysis of Option Implied Volatility and Stock Endogenous Time Volatility

2.1 Introduction

The Black-Scholes option pricing model provides elegant closed-from solutions. However, oftentimes it introduces significant option pricing errors in practice caused by restrictive assumptions of its own. The Black-Scholes model is built upon assumptions including constant volatility over option lifetime across different maturities, geometric Brownian motion framework, constant risk free rate, and non-arbitrage market setting. Consequently, if the Black-Scholes model were correct, then the volatility deduced from it should be the same across various maturities and strike prices, *i.e.* a flat implied volatility surface. Nevertheless, empirical features of the volatility surface, which is often referred as volatility smile, reveals that for a given maturity, options that are deep in the money or out of the money have higher implied volatilities than at-the-money options. One can regard implied volatility as the volatility reflecting to a realized option price, which can only be used in the Black-Scholes model. Mathematically, implied volatility from an

option is the inverse function of a Black-Scholes given its realized option price.

On the other hand, implied volatility plays an important role in both finance theory and financial market. It is considered to be closely related to stock market volatility as an efficient tool evaluating market's opinion towards a particular stock. Additionally, options are often traded on volatility with the implied volatility as a part of the effective price. Implied volatility also has important implications for risk management. For example, as volatility increases, the value at risk (VaR) goes up as well. Investors may want to adjust the risk exposure of their portfolio, whose volatility is predicted to increase.

Due to the importance of implied volatility as well as its weakness, a large body of literature has been developed around this topic. Heston (1993) extended the Black-Scholes model with a stochastic volatility setting, which is correlated with returns, to circumvent the issue arising from the original constant volatility assumption. The main contribution from his paper is that he used characteristic functions to derive a closed-form solution for the price of a European call option on an asset with stochastic volatility. A second path to address the same issue is from Derman and Kani (1994); Dumas et al. (1998); Rubinstein (1994). They approached the problem with a much lighter but an ad hoc tool: an OLS regression. The model is referred to as the deterministic volatility function (DVF) or the so-called Practitioner's Black-Scholes (PBS) by Christoffersen and Jacobs (2004). They found that option prices from a simple OLS regression of implied volatility on a polynomial in moneyness and maturity outperformed models whose diffusion term is a deterministic function of the strike price and maturity. It is worth noting that although DVF is a parsimonious model, it achieves impressive fit of actual option prices.

In this paper, we are trying to extend their DVF framework with volatility of the underlying equity to model implied volatility. The equity volatility estimator will be constructed from microstructure noise contaminated high-frequency trading data. This idea is motivated by Bergman et al. (1996), who established the theory that option prices should be correlated with the underlying asset volatility. Recent stimulating development of high-frequency equity volatility estimator provides financial econometrians profound insight into new ways of modeling implied

volatility. Before presenting our idea, we first introduce some basics of high-frequency equity volatility estimators.

A standard approach is to model the asset price in continuous time with an Itô process $d \ln S_t = \mu_t dt + \sigma_t dW_t$, where $\ln S_t$ is the logarithmic price, μ_t is the drift term, σ_t is the instantaneous volatility, and dW_t is the stochastic Wiener term, whose increments over a period of time Δt follow a normal distribution, $N(0, \Delta t)$. The process is observed at times t_i, $i = 0, 1, \cdots, N$ in the time interval $[0, 1]$. Thus the increment of $\ln S_t$ at the observation time t_i is defined as $\Delta X_i \equiv \ln S_{t_i} - \ln S_{t_{i-1}}$. By design, we see that the difference of log-prices, ΔX_i, is also the continuously compounded rate of return of the i-th trade. Within the same framework, additional terms (or processes) can be added to the Itô process to encompass unique features of the high-frequency data, including price jumps and microstructure noise. To evaluate the volatility of a financial asset, a natural way of defining the cumulative variance of asset prices over a period of time is simply to integrate the instantaneous volatility over the time interval. It is called integrated volatility (IV) in the literature, defined as: IV $\equiv \int_0^1 \sigma_t^2 dt$. Unfortunately, the instantaneous volatility is a latent variable. Searching for better volatility proxies has been an ongoing challenge in the financial econometrics research. Jacod and Protter (1998) laid the statistical theory foundation in their probability theory by proving that for an Itô process, Realized Variance (RV), defined as $[X, X]_t = \sum_{t_i \leq t} (\Delta X_i)^2$, asymptotically in-fill converges (sample size goes to infinite while observation time between two consecutive sample goes to zero) to the IV of the process. Barndorff-Nielsen and Shephard (2001, 2002) introduced this theory to financial econometrics. Their work has mushroomed more than 400 nonparametic volatility estimators of IV (see Liu et al., 2015). Several of these estimators have proven superior to the rest from various aspects. 5-minute Realized Volatility ($\text{RV}_{5\text{min}}$) is widely used due to its easiness of implementation (see Andersen and Benzoni, 2009, for details). Computing $\text{RV}_{5\text{min}}$ requires aggregating high-frequency data into 5 minute intervals and then summing up the squared returns at each 5-minute time interval: $\text{RV}_{5\text{min}} \equiv \sum (\Delta X_i^{5\text{min}})^2$, where $\Delta X_i^{5\text{min}}$ is the log-return of the ith interval on a consecutive 5-minute time grid. Bipower Realized Volatility (BPV) (see

Barndorff-Nielsen and Shephard, 2004) and Truncated Realized Volatility (TRV) (see Mancini, 2009) are representative choices dealing with price jumps. Finally, Realized Kernel (RK) (see Barndorff-Nielsen et al., 2011) serves as a robust estimator to both jumps and microstructure noises with substantially improved precision.

One of many prominent and valuable advantages of high-frequency data compared to lower frequency data is the inclusion of irregularly spaced transaction time. Intuitively, informed market participants would have clustered trading behavior when news hitting the market and flat trading behavior when no news hit the market. Both clustered and flat trades break the demand-supply balance in the market and affect the volatility of the financial products (see Easley and O'Hara, 1992). Hence, by including transaction duration, the time span between two consecutive transactions, one can generate a better volatility estimator. Representative models coupling with temporal feature includes: the autoregressive conditional duration (ACD) proposed by Engle and Russell (1998), the Stochastic Conditional Duration (SCD) model put forth by Bauwens and Veredas (2004), Abbring (2012)'s mixed hitting-time model, and Pelletier and Zheng (2013) and Wei and Pelletier (2015) bivariate Ornstein-Uhlenbeck (OU) process model for duration and price return. Li et al. (2014) made an interesting discovery that under the assumption that there exists an instantaneous correlation between duration and price return, the realized volatility is an asymptotically biased estimator of the integrated volatility. They documented that such time endogeneity between the transaction durations and the price process exists in financial data. In light of the discovery by Li et al. (2014), Pelletier and Weng (2015) developed a joint model for irregularly spaced returns and stock volatility. In their model, durations are endogenous, meaning the time span between two consecutive transactions is correlated with the stock price process. The model contains a bivariate Ornstein-Uhlenbeck (OU) process that jointly models equity latent volatility and trading intensity. Together with two other processes modeling trading prices and durations, the time endogeneity is captured by the model. They concluded that when treating transaction times as endogenous, their version of volatility estimator, the so-called endogenous time volatility estimator, is superior to most other realized volatility estimators in

terms of asymptotic biasness and forecasting performance.

As the flourishing development of high-frequency equity volatility estimators, the literature exploring their utility in option pricing is also blooming (see Bandi et al., 2008; Christoffersen et al., 2014). Most recent research by Christoffersen et al. (2014) investigated whether these forecasting improvements translate into added economic value. In their paper, they incorporated realized volatility, $\mathrm{RV}_{5\mathrm{min}}$ mainly, with a new class of affine discrete-time option valuation models. They concluded that realized volatility reduces the pricing errors significantly across moneyness, maturity, and volatility levels. Bandi et al. (2008) compared the profits from option pricing and trading with different nonparametric volatility. Main results from their findings is that estimators composed from flat-top kernels generally outperform other choices in terms of option pricing accuracy. Motivated by these work, we will investigate the benefits of integrating the time endogenous equity volatility with the DVF model in this paper. We are curious to see whether a more accurate volatility estimator of the underlying asset will lead to a more precise option pricing model.

The paper proceeds as follows: In Section 2, we present the endogenous time volatility estimator first. Then we incorporate it with four DVF model candidates. In Section 3, we discuss the estimation of the model. Section 4 contains a summary of an empirical study on IBM 2008 option transaction data as well as its high-frequency stock data. Section 5 concludes.

2.2 Model Specification

In this section, we first briefly introduce the endogenous time volatility estimator by Pelletier and Weng (2015). Second, we outline the DVF framework by Dumas et al. (1998) and incorporate it with the time endogenous volatility.

2.2.1 Time Endogenous Volatility

Following Pelletier and Weng (2015), the endogenous time volatility model consists of four layers. Loosely speaking, the first layer is a logarithmic price process linking observable prices with latent

instantaneous volatilities. The second layer is a dynamic process for the durations. Durations are generated by an exponential distribution with the conditional mean equal to the inverse of the latent trade intensity. The third layer of our model is a bivariate Ornstein-Uhlenbeck (OU) process for two latent variables: the logarithmic trade intensity and the logarithmic instantaneous volatility. The last layer allows for time endogeneity via a Gaussian copula. The details of the model are presented below.

The first layer of the endogenous time volatility model is a Wiener process of logarithmic equity price. The dynamics of the logarithmic price S_t at time t is defined as:

$$d \ln S_t = \sigma_t dW_{0,t}, \tag{2.1}$$

where $W_{0,t}$ is a Weiner term. σ_t is a latent instantaneous volatility as its coefficient. We ignore the drift term in our model for simplicity.

The second layer is a Poisson process for durations. λ_{t_i} is the stochastic intensity characterizing a point process. Intuitively, λ_{t_i} can be interpreted as the instantaneous probability to have a i-th trade occur at time t. It is defined as:

$$\lambda_{t_i} = \lim_{\Delta t \to 0} \left(\frac{1}{\Delta t} Prob[N(t + \Delta t) - N(t) = 1 | F_{t-}] \right), \tag{2.2}$$

where $N(t) = \sum_{i \geq 1} 1(T_i \leq t)$ is a counting process, summing up the total numbers of events up to and including time t. F_{t-} is all information available up to time t.

Naturally, we define the time span between the $(i-1)$-th trade and the i-th trade as duration d_i:

$$d_i \equiv T_i - T_{i-1}, \tag{2.3}$$

where T_i is the instant of the i-th occurrence of an event, satisfying $0 < T_1 < T_2 < \cdots$.

We assume d_i follows a conditional exponential distribution with mean $\lambda_{t_{i-1}}^{-1}$:

$$f(d_i|\lambda_{t_{i-1}}) = \frac{1}{\lambda_{t_{i-1}}} \cdot \nu_i \qquad \text{with} \quad \nu_i \sim \text{Exp}(1). \tag{2.4}$$

We could also write it as $f(d_i|\lambda_{t_{i-1}}) \sim \text{Exp}(\lambda_{t_{i-1}})$, which fits in the definition of a Poisson process.

The third layer in our model is a bivariate Ornstein-Uhlenbeck (OU) process for two latent variables: log-intensity λ_{t_i} and log-variance $\sigma_{t_i}^2$, which is denoted as:

$$dX_t = A(\mu - X_t)dt + SdW_{-0,t} \tag{2.5}$$

where:

$$X_{t_i} = \begin{bmatrix} \ln \lambda_{t_i} \\ \ln \sigma_{t_i}^2 \end{bmatrix}, \quad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

$$S = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \quad W_{-0,t} = \begin{bmatrix} W_{1,t} \\ W_{2,t} \end{bmatrix}.$$

The last layer of our model introduces the endogenous time effect to our model and links aforementioned ingredients all together via a Gaussian Copulas. The idea of Copulas is that joint distribution function can be written as a unique Copulas function of marginal distributions of each random variables under the condition that all marginal distributions are continuous (Sklar, (1959)). We impose contemporaneous correlations, denoted as ρ, to our model. The specifications of ρ's in Γ are defined as follows: ρ_1 is the instantaneous correlation between Wiener terms in $\ln S_{t_i}$ and $\ln \lambda_{t_i}$ dynamics. It is defined as:

$$\rho_1 \equiv < W_{0,t}, W_{1,t} > . \tag{2.6}$$

ρ_2 is the parameter controlling the leverage effect. It is defined as the instantaneous correlation

between Wiener terms in $\ln S_{t_i}$ and $\ln \sigma^2_{t_i}$ dynamics, denoted as:

$$\rho_2 \equiv < W_{0,t}, W_{2,t} > . \tag{2.7}$$

ρ_3 is the instantaneous correlation between Wiener terms in $\ln \lambda_{t_i}$ and $\ln \sigma^2_{t_i}$ dynamics. It is defined as:

$$\rho_3 \equiv < W_{1,t}, W_{2,t} > . \tag{2.8}$$

ρ_4 brings the time endogeneity effect to the model. If it is set to be zero then the model is counted for a exogenous time case. It is denoted as:

$$\rho_4 \equiv \mathrm{Corr}(\Phi^{-1}(F_{\exp}(\nu_i)), \Delta W_{0,t_i}/\sqrt{d_i}). \tag{2.9}$$

The Gaussian copulas density function in our case is defined as:

$$c(x) = \frac{1}{|\Gamma|^{1/2}} \exp\left[-\frac{1}{2}u'(\Gamma^{-1} - I_4)u\right], \tag{2.10}$$

where:

$$x = \begin{bmatrix} \Delta W_{0,t_i}/\sqrt{d_i} \\ \Delta W_{1,t_i}/\sqrt{d_i} \\ \Delta W_{2,t_i}/\sqrt{d_i} \\ \nu_i \end{bmatrix}, \quad \Gamma = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_4 \\ \rho_1 & 1 & \rho_3 & 0 \\ \rho_2 & \rho_3 & 1 & 0 \\ \rho_4 & 0 & 0 & 1 \end{bmatrix}, \quad u = \begin{bmatrix} \Delta W_{0,t_i}/\sqrt{d_i} \\ \Delta W_{1,t_i}/\sqrt{d_i} \\ \Delta W_{2,t_i}/\sqrt{d_i} \\ \Phi^{-1}(F_{\exp}\nu_i) \end{bmatrix}.$$

The normal quantile function u_i is given by: $u_i = \Phi^{-1}(F_i(x_i))$, where $F_i(\cdot)$ is the corresponding cumulative density function (cdf) of x_i.

To estimate the continuous-time model system, we need to discretize and log-linearize it as described in Pelletier and Weng (2015). They showed that the model could be written into a state-space representation. They employed the Kalman filter via QMLE to estimate the model, a brief outline of this procedure is showed in the next section. As a byproduct of the Kalman filter, we can recover latent variable $\ln \sigma^2_{t_i}$ through the Kalman smoother. Hence, the expected

intraday endogenous time volatility estimator can be obtained by:

$$\mathbf{E}(\sigma_{t_i}^2|I_T) = \exp(\mathbf{E}(\ln \sigma_{t_i}^2|I_T) + \mathrm{Var}(\ln \sigma_{t_i}^2|I_T)/2). \tag{2.11}$$

Therefore, the intra-day endogenous time volatility estimator is specified as:

$$\sigma_{\mathrm{en}}^2 = \sum_{i=1}^{N} \mathbf{E}(\sigma_{t_i}^2|I_T) * d_i. \tag{2.12}$$

where N is the total number of stock transactions in a particular trading day. We choose to integrate the endogenous time volatility obtained on the option execution day to the DVF framework showed below.

2.2.2 Deterministic Volatility Function

The second part of our model is to integrate σ_{en}^2 to Deterministic Volatility Function. The general procedure for Deterministic Volatility Function involves: (a) Given realized option prices, recover the Black-Scholes implied volatilities by inverting the Black-Scholes option pricing solution, (b) estimate coefficients of selected deterministic volatility functions, (in our case, DVF will include endogenous time volatility), (c) predict volatility based on the estimators obtained in step (b), (d) price options via Black-Scholes formula with volatility produced in step (c). Details are presented below.

The Black-Scholes European call option solution is:

$$C = e^{-q(T-t)} S_t \Phi(d_1) - e^{-r(T-t)} K \Phi(d_2), \tag{2.13}$$

where C stands for the price for a call option, S_t is spot price of underlying asset price, K is strke price, r is risk free rate, q is dividend yield rate, σ is volatility, and $T - t$ is option maturity.

Moreover Φ is the CDF of the standard normal distribution. d_1 and d_2 are equal to:

$$\begin{aligned} d_1 &= \frac{\ln(S_t/K) + (r - q + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \\ d_2 &= d_1 - \sigma\sqrt{T-t}. \end{aligned}$$

European option's put-call parity states that:

$$e^{-rT}K + C = e^{-qT}S_t + P \tag{2.14}$$

where P is the put option price with the same strike K and maturity T as the call option.

Given realized call (or put) option prices, C_{MKT}, these formulas therefore allow us to back out the value of σ, which is the implied volatility by the market price of that option,

$$\sigma_{\text{IV}} \equiv BS^{-1}(C_{\text{MKT}}, S_t, T, r, K, q). \tag{2.15}$$

Once we obtained Black-Scholes implied volatility from eq (2.15), the next step is to run a simple OLS regression of σ_{IV} on a polynomial in moneyness $M_t = S_t/K$, maturity T, and endogenous time volatility σ_{en}. Four selected regression candidates are given as follows:

$$\text{DVF 1} : \sigma_{\text{IV}} = \beta_0 + \beta_1 T + \beta_2 M_t, \tag{2.16}$$

$$\text{DVF 2} : \sigma_{\text{IV}} = \beta_0 + \beta_1 T + \beta_2 M_t + \beta_3 M_t^2 + \beta_4 M_t T, \tag{2.17}$$

$$\text{DVF 3} : \sigma_{\text{IV}} = \beta_0 + \beta_1 T + \beta_2 M_t + \beta_5 \sigma_{\text{en}}, \tag{2.18}$$

$$\text{DVF 4} : \sigma_{\text{IV}} = \beta_0 + \beta_1 T + \beta_2 M_t + \beta_3 M_t^2 + \beta_4 M_t T + \beta_5 \sigma_{\text{en}}. \tag{2.19}$$

Model 1 and model 2 attempt to capture variation in volatility attributed to both asset price and time. Model 3 and model 4 capture additional variation from treating the underlying asset trading processes are endogenous. We believe that σ_{en} will provide extra accuracy of fit compared to conventional DVF models. According to Dumas et al. (1998), adding quadratic form and cross

product terms of maturity and strike price will further improve the fitness of DVF. β_0 is the intercept term, β_1 and β_2 are the linear function of T and M_t respectively, while β_3 controls the curvature of the parabola shape of moneyness. Heuristically, one would expect a negative β_2 and a positive β_3. Li et al. (2014) found empirical evidence from high-frequency equity trading data that transaction times are endogenous, where only endogenous time volatility is the only consistent and asymptotically unbiased. Furthermore, the literature also suggests that underlying asset volatility has strong correlation to its Black-Scholes implied volatility. Hence, Incorporating σ_{en} will allow DVF to capture the variation from underlying asset volatility.

The last piece of the model is to replace the parameter σ in the Black-Scholes equation with the new σ_{DVF} to obtain the option prices, C_{DVF}, given different maturities and strike prices. To do so, after estimating DVF, we need to use equation (2.20) to generate a set of backward looking σ_{DVF}. A minimum value of the local volatility rate is imposed to prevent negative values.

$$\sigma_{\text{DVF}} = \max(0.001, \text{DVF}(\hat{\beta}, T, K, \sigma_{\text{en}})), \tag{2.20}$$

$$C_{\text{DVF}} = e^{-q(T-t)} S_t \Phi(d_1) - e^{-r(T-t)} K \Phi(d_2), \tag{2.21}$$

where

$$\begin{aligned} d_1 &= \frac{\ln(S_t/K) + (r - q + \sigma_{\text{DVF}}^2/2)(T-t)}{\sigma_{\text{DVF}}\sqrt{T-t}}, \\ d_2 &= d_1 - \sigma_{\text{DVF}}\sqrt{T-t}. \end{aligned}$$

To evaluate the performance of DVF candidates, we calculate the value of mean-squared dollar ($MSE) objective function (see Christoffersen and Jacobs, 2004) and relative prediction accuracy (RPA) function. DVF with the smallest $MSE and RPA indicates better forecasting

performance. These two loss functions are specified as:

$$\$\text{MSE} = \frac{1}{NT}\sum_{i}^{N}\sum_{t}^{T}(C_{\text{MKT},i,t} - C_{\text{DVF},i,t})^2, \tag{2.22}$$

$$\text{RPA} = \frac{1}{NT}\sum_{i}^{N}\sum_{t}^{T}\frac{|C_{\text{MKT},i,t} - C_{\text{DVF},i,t}|}{C_{\text{MKT},i,t}}, \tag{2.23}$$

where $C_{\text{MKT},i,t}$ denotes the market price of an individual call option i at time t. $C_{\text{DVF},i,t}$ follows the same notation.

2.3 Estimation

In this section, we present main estimation results from Pelletier and Weng (2015) endogenous time volatility model. First part is the state space representation of their continuous-time model after discretization and linearization. Then they employed the Kalman filter to do Quasi-Maximum Likelihood Estimation (QMLE) using similar techniques to the ones in Harvey et al. (1994); Pelletier and Zheng (2013). To recover the endogenous time volatility estimator we will use the Kalman smoother algorithm from De Jong and Shephard (1995).

2.3.1 Endogenous Time Volatility Model Estimation

After discretization and linearization conditioning on the signs of returns, the model has a state space representation as follows [1]:

Observation Equation:

$$Y_{t_i} = \begin{bmatrix} \mu^{\psi}_{*,t_i} \\ -1.2704 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} X_{t_{i-1}} + \begin{bmatrix} \psi_{*,t_i} \\ \varepsilon_{*,t_i} \end{bmatrix}, \tag{2.24}$$

[1] The subscript $*$ indicates the use of the sign.

where

$$Y_{t_i} = \begin{bmatrix} \ln(d_i) \\ \ln(y_{t_i}^2) - \ln(d_i) \end{bmatrix}, \quad X_{t_i} = \begin{bmatrix} \ln(\lambda_{t_i}) \\ \ln(\sigma_{t_i}^2) \end{bmatrix}.$$

State Equation:

$$X_{t_i} = [I_2 - \text{expm}(-Ad_i)]\mu + \mu_{*,t_i}^{\Omega} + \text{expm}(-Ad_i)X_{t_{i-1}} + \Omega_{*,t_i}, \tag{2.25}$$

where

$$\mu_{*,t_i}^{\Omega} = 0.7979 \cdot \Sigma_{t_i}^{\zeta,\Omega}(d_i)^{-0.5} \cdot \text{sign}(y_i) \quad \text{with} \quad \text{sign}(y_i) = \begin{cases} 1 & \text{if } y_i \text{ is positive} \\ -1 & \text{if } y_i \text{ is negative} \end{cases}.$$

The variance-covariance matrix of the innovation terms conditional on the signs of the returns is:

$$\text{Var}\begin{bmatrix} \psi_{*,t_i} \\ \varepsilon_{*,t_i} \\ \Omega_{*,t_i} \end{bmatrix} = \begin{bmatrix} \sigma_{\psi,*,t_i}^2 & \Sigma_{*,t_i}^{\psi,\varepsilon} & \mathbf{0}_{1\times 2} \\ \Sigma_{*,t_i}^{\psi,\varepsilon} & \sigma_{\varepsilon,*}^2 & \Sigma_{*,t_i}^{\varepsilon,\Omega\prime} \\ \mathbf{0}_{2\times 1} & \Sigma_{*,t_i}^{\varepsilon,\Omega} & \Sigma_{*,t_i}^{\Omega} \end{bmatrix}_{4\times 4},$$

where

$$\begin{aligned} \Sigma_{*,t_i}^{\varepsilon,\Omega} &= 1.1061 \cdot \Sigma_{t_i}^{\zeta,\Omega}(d_i)^{-0.5} \cdot \text{sign}(y_i), \\ \Sigma_{*,t_i}^{\Omega} &= \Sigma_{t_i}^{\Omega} - \mu_{*,t_i}^{\Omega}\mu_{*,t_i}^{\Omega}{}'. \end{aligned}$$

Then we employ the Kalman filter (see De Jong, 1991) to estimate the model then to compute endogenous time volatility estimators via the Kalman Smoother.

2.3.2 Kalman Filter

We follow the Kalman filter documented in De Jong (1991) and De Jong and Shephard (1995). This version of the Kalman filter allows the innovation terms from the state equation and the observation equation to be correlated. The state space representation is given by the following system of equations:

Observation Equation: $$y_t = X_t\beta + Z_t\alpha_t + G_t u_t;$$

State Equation: $$\alpha_{t+1} = W_t\beta + T_t\alpha_t + H_t u_t.$$

where u_t are independent $N(0, \sigma^2 I)$ variables. β, Z_t, G_t, T_t, and H_t are parameters of interest. To estimate the state space model, they implements the following Kalman filter to record e_t, D_t and K_t:

Innovation: $$e_t = y_t - X_t\beta - Z_t a_t$$

Innovation Covariance: $$D_t = Z_t P_t Z_t' + G_t G_t'$$

Kalman Gain: $$K_t = (T_t P_t Z_t' + H_t G_t') D_t^{-1}$$

$$a_{t+1} = W_t\beta + T_t a_t + K_t e_t$$

$$P_{t+1} = T_t P_t L_t' + H_t J_t'$$

$$L_t = T_t - K_t Z_t$$

$$J_t = H_t - K_t G_t$$

where $a_1 = W_0\beta$ and $P_1 = H_0 H_0$. As the by-product of Kalman Filter, the log-likelihood function based on $Y = \{y_1, y_2, \cdots, y_t\}$ is:

$$\ln L(Y) = -\frac{1}{2}\{\sum_{t=1}^{T} \ln|D_t| + \sum_{t=1}^{T} e_t' D_t^{-1} e_t\}.$$

Thus, the corresponding mappings to De Jong's Kalman filter algorithm are:

$$X_t\beta = \begin{bmatrix} \mu^{\psi}_{*,t_i} \\ -1.2704 \end{bmatrix}, \quad Z_t = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$W_t\beta = [I_2 - \text{expm}(-Ad_i)]\mu + \mu^{\Omega}_{*,t_i}, \quad T_t = \text{expm}(-Ad_i),$$

$$G_tG_t' = \begin{bmatrix} \sigma^2_{\psi,*,t_i} & \Sigma^{\psi,\varepsilon}_{*,t_i} \\ \\ \Sigma^{\psi,\varepsilon}_{*,t_i} & \sigma^2_{\varepsilon} \end{bmatrix}_{2\times 2}, \quad G_tH_t' = \begin{bmatrix} \mathbf{0}_{1\times 2} \\ \\ \Sigma^{\varepsilon,\Omega'}_{*,t_i} \end{bmatrix}_{2\times 2}, \quad H_tH_t' = \Sigma^{\Omega}_{*,t_i}.$$

2.3.3 Kalman Smoother

To get smoothed $\sigma^2_{\text{en},t}$, we store e_t, D_t, and K_t after implementing the Kalman filter displayed above. The Kalman smoother algorithm is as follows: first we set $r_T = 0$ and $U_T = 0$. Then we run for $t = T, T-1, \cdots, 1$,

$$\begin{aligned}
C_t &= H_t(I - G_t'D_t^{-1}G_t - J_t'U_tJ_t)H_t', \\
\varepsilon_t &\sim N(0, \sigma^2 C_t), \\
V_t &= H_t(G_t'D_t^{-1}Z_t + J_t'U_tL_t), \\
r_{t-1} &= Z_t'D_t^{-1}e_t + L_t'r_t - V_t'C_t^{-1}\varepsilon_t, \\
U_{t-1} &= Z_t'D_t^{-1}Z_t + L_t'U_tL_t + V_t'C_t^{-1}V_t, \\
\eta_t &= H_t(G_t'D_t^{-1}e_t + J_t'r_t) + \varepsilon_t \sim p(H_tu_t|\cdot).
\end{aligned}$$

In our case, $\ln\sigma_t^2$ is presented by α_t. Hence, $\tilde{\alpha_{t+1}} = W_t\beta + T_t\alpha_t + \eta_t$. $\tilde{\alpha_t}$ stands for the smoothed version of latent variable α_t. Furthermore, $\text{Var}(\ln\sigma^2_{t_i}|I_T)$ in equation (2.11) corresponds to C_t. The final step is to input the series of $\tilde{\alpha}_t$ to equation (2.11).

2.4 Empirical Results

2.4.1 Summary Statistics

We choose IBM (ticker: IBM) stock for our analysis. The sample period is from Jan. 03, 2008 to Dec. 31, 2008. To compute daily endogenous time volatilities over the period, we follow the data cleaning procedure described in Barndorff-Nielsen et al. (2009) to process the original data. We first remove all trading records outside the 9:30 am - 4 pm EST window when the exchange is open. Second, we delete entries with corrected trades, whose correction indicator is different from zero, as well as those with an abnormal sale condition (where COND has a letter code other than 'E' and 'F'). Third, if multiple transactions have the same time stamp, we only keep one trade record with the volume-weighted price.

Other than the standard data cleaning procedure, market microstructure noise is a native feature of the high-frequency data. Examples of microstructure noise include bid-ask spread, discreteness of price changes, data recording error, and rounding effects, *etc.* Hansen and Lunde (2006) point out that auto-correlated market microstructure noise contaminates observed returns, which biases the realized variance at ultra-high frequencies. A common treatment to mitigate microstructure noise in the realized volatility literature is to employ calendar-time (every n minutes) sampling scheme under the assumption of observation times being exogenous. It is widely believed that the lower frequency of sampling schemes, the smaller magnitude of the noise, the lower noise-to-signal ratio. However, a obvious drawback of the calendar-time sampling scheme is the complete loss of temporal information in the raw data. For the purpose of preserving temporal information and simultaneously diminishing the market microsctructure noise, we employ the tick-time (every m trades) sampling scheme. Similar to the calendar-time sampling scheme, it combines returns and durations of all trades contained in each m-trade interval. Fukasawa (2010) proves that under the assumption of no rounding error existing and observation times being exogenous, the tick-time sampling scheme provides an appropriate RV estimator asymptotically converging to IV on price grids, including both transaction price grids and quote

price grids.

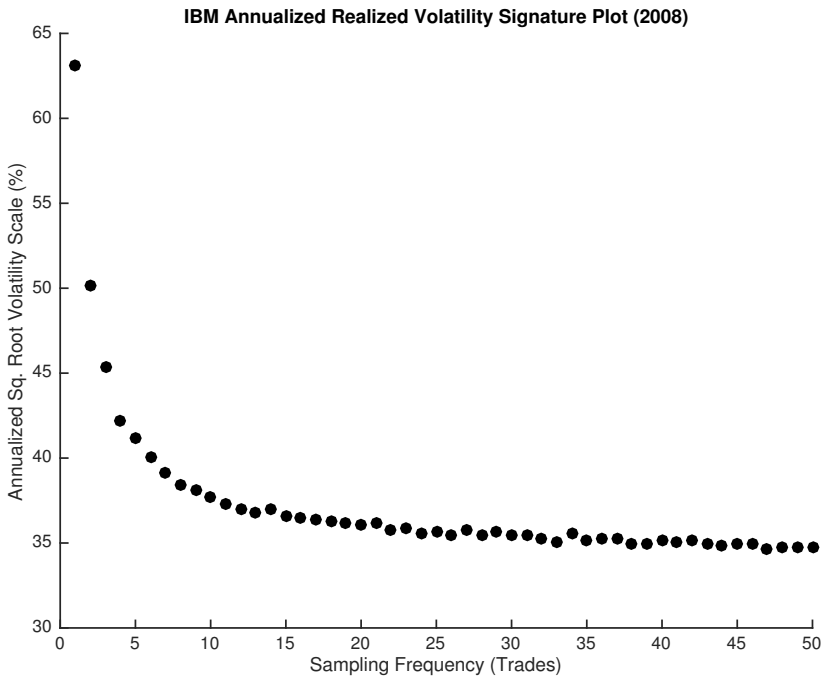


Figure 2.1: Annualized IBM 2008 Realized Volatility Signature Plot

Figure 2.1 depicts the volatility signature for different tick-time sampling frequencies. The y-axis indicates the annualized square root of volatility. The x-axis refers to the sampling frequency by m-trade. One can observe that volatility starts to stabilize when sampling frequency is above the 15-trade mark, indicating that the microstructure noise is sufficiently diminished by aggregating individual returns at a frequency of 15 trades or higher. We choose the 'every 50 trades' tick-time sampling scheme for our empirical study.

Table 2.1 includes the descriptive statistics for the returns and durations over the sample period. The number in the column name indicates the sampling frequency used for the analysis. For example, '50-Trade Return' represents the return series where an observation consists of the returns over 50 consecutive trades. We observe positive skewness and large excess kurtosis

Table 2.1: IBM 2008 Every-50-Transaction Summary Statistics

	50-Trade Return	50-Trade Duration
Obs. #	155,238	155,238
Mean	0.000016	37.865041
Std.	0.10565	26.81197
Min	-1.85162	1.71000
Max	2.30749	484.81300
Median	0.00000	31.14100
Skewness	0.09005	1.67156
Kurtosis	13.670	7.952
ACF(1)	-0.01552	0.66412
ACF(2)	-0.00909	0.60722
ACF(3)	-0.00794	0.58433

from the single-trade data. The 50-trade sampling scheme mitigates the issue. Both the return distribution and duration distribution have heavier tails than a normal distribution. It is worth noting that autocorrelation for the return series becomes negligible after order one in the single-trade frequency data set. However, for the 50-trade frequency data set, the return series appears uncorrelated. The duration series, on the other hand, shows increasing autocorrelation after being aggregated.

The IBM option trading data has the same sampling period as its equity from Jan. 03, 2008 to Dec. 31, 2008. Each transaction record contains the following information: the underlying asset ticker, the underlying asset spot prices, the exchange of the quote, option type (call or put), option expatriation date, the strike price of the option, the option bid-ask prices, the option trading volume, and open interests. All options are either European call or put options, which distribute almost symmetrically in the sample set. Table 2.2 shows that the original data set contains 46,660 daily observations in 2008. Their moneyness ranges from 0.38 to 4.32 with their maturities varying from 1 day to 593 days. To clean up data, all observations are subject to following conditions: First, options not satisfying lower boundary condition, $C \geq S_t - Ke^{T-t}$, are discarded. Second, options having moneyness above 2 are dropped, since options with moneyness outside of 2 are generally not reflected as "true" option value. Third, options with maturity less than 3 days are filtered since those options are highly sensitive to price-volatility bias. Finally,

Table 2.2: Original IBM 2008 Traded Options Summary Statistics

	All Options			Call Options			Put Options		
				Moneyness					
Range	0-0.8	0.8-1.2	>1.2	0-0.8	0.8-1.2	>1.2	0-0.8	0.8-1.2	>1.2
Obs.	8448	21860	14441	5254	11053	6945	3194	10807	7496
Mean	0.69	0.98	1.52	0.68	0.98	1.52	0.72	0.98	1.52
Std.	0.08	0.11	0.33	0.09	0.11	0.33	0.06	0.11	0.33
Min	0.38	0.80	1.20	0.38	0.80	1.20	0.44	0.80	1.20
Max	0.80	1.20	4.32	0.80	1.20	4.32	0.80	1.20	4.32
Median	0.72	0.97	1.45	0.70	0.96	1.45	0.73	0.97	1.45
Skewness	-0.94	0.24	3.47	-0.70	0.26	3.37	-1.10	0.22	3.55
Kurtosis	3.26	1.91	22.66	2.62	1.92	21.75	0.43	1.91	23.45
				Maturity (days)					
Range	< 40	40-70	>70	< 40	40-70	>70	< 40	40-70	>70
Obs.	15,016	4,578	26,589	7,771	2,352	13,887	7,245	2,226	12,702
Mean	20.26	53.74	231.18	20.36	53.73	230.78	20.14	53.76	231.61
Std.	11.05	8.78	147.59	10.97	8.82	147.17	11.14	8.74	148.04
Min	1	41	71	1	41	71	1	41	71
Max	39	69	593	39	69	593	39	69	503
Median	20	53	161	21	53	161	20	54	161
Skewness	-0.02	0.11	0.89	-0.04	0.13	0.89	-0.01	0.10	0.89
Kurtosis	1.83	1.73	2.49	1.85	1.72	2.51	1.82	1.74	2.46

options with implied volatility less than 1% or greater than 100% are deleted.

For implied volatility computation, we use annualized 3-month treasure bill rate as the constant risk-free rate base upon equation (2.15). The present value of actual dividends paid before maturity is subtracted from the stock value. Hence, the continuous compound daily yield rate, q, is derived from the underlying stock historical dividend data. Finally, we take the average price of bid-ask price as the option trading price in equation (2.15). For DVF regressions, we normalized all covariates before estimation. [2]

Consequently, Table 2.3 summarizes the statistics of the screened IBM option transactions based on three different categories: option types, moneyness, and maturities. The remaining data amounts to 43,969 observations, 94.2% of original 46,660 observations. We can see that

[2]For every set of covariate, we subtract its mean, then divided by its standard deviation.

Table 2.3: Filtered IBM 2008 Options Summary Statistics

	All Opiotns			Call Options			Put Options		
				T<20					
Moneyness	<0.94	0.94-1.06	$\geq$ 1.06	<0.94	0.94-1.06	$\geq$ 1.06	<0.94	0.94-1.06	$\geq$ 1.06
Obs.	2084	1015	2338	1294	508	992	790	507	1346
Avg. Price	8.40	3.36	9.81	0.12	3.15	22.78	21.97	3.58	0.25
Avg. Im. Vol.	0.49	0.33	0.60	0.48	0.33	0.60	0.50	0.33	0.59
				$20 \leq$ T $<$ 40					
Moneyness	<0.94	0.94-1.06	$\geq$ 1.06	<0.94	0.94-1.06	$\geq$ 1.06	<0.94	0.94-1.06	$\geq$ 1.06
Obs.	2959	4185	3529	1727	2340	1681	1232	1845	1848
Avg. Price	9.95	8.50	12.92	0.40	1.55	26.41	23.34	17.31	0.66
Avg. Im. Vol.	0.38	0.37	0.51	0.38	0.37	0.54	0.38	0.36	0.49
				T$\geq$40					
Moneyness	<0.94	0.94-1.06	$\geq$ 1.06	<0.94	0.94-1.06	$\geq$ 1.06	<0.94	0.94-1.06	$\geq$ 1.06
Obs.	12630	4454	13734	6987	2230	6836	5643	2224	6898
Avg. Price	15.61	11.01	17.67	3.20	11.11	32.34	30.97	10.91	3.14
Avg. Im. Vol.	0.31	0.31	0.41	0.32	0.32	0.43	0.29	0.30	0.39

option prices increase as moneyness and maturities go up for call options and opposite for put options. The most liquid options traded in the market are options with long term maturities (T$\geq$40). Finally, as expected, a volatility smile can be clearly observed across the moneyness and maturity categories.

2.4.2 Analysis of the Results

This subsection focuses on the detailed proceeding we deploy to identify whether endogenous time equity volatility as an additional source can improve the pricing error of IBM options.

First, to obtain daily endogenous time volatility estimator, which has been included in DVF 3 and DVF 4 models, we first estimate all parameters in equation (2.1) to equation (2.10) using IBM high-frequency equity transaction data, whose sampling period ranges from Nov. 17, 2008 to Dec. 31, 2008. We combine every 50 transactions together as one data point to mitigate microstructure noise. The reason why we do not choose to include all data from 2008 is that the endogenous time volatility model is too computationally intensive. Second, after obtaining those parameter values, we generate daily IBM stock endogenous time volatility for each business day in 2008 via equation (2.12). The third step is to estimate all four DVF models on a one-month rolling

window basis of IBM option data and IBM daily endogenous time volatility. More specifically, the one-month rolling window mechanism works as follow: we use one month length data to estimate DVF models and use it to predict the following day option prices. The sampling length is always one month long. However, the starting date and ending date of the sampling period is rolling. Each DVF regression has been estimated only once based on corresponding observations in that period. Once obtained estimated parameters of four different DVF models, we feed the following day option data into the estimated DVF models to generate predicted implied volatility, σ_{IV}. The final step is replacing σ with predicted σ_{IV} in call or put Black-Scholes to get predicted option prices, $C_{\mathrm{DVF},i,t}$. \$MSE and RPA, which are defined in equation (2.22) and equation (2.23), are reported in Table 2.7.

Table 2.4, Table 2.5, and Table 2.6 display the parameter estimates of DVF regression analysis based on all IBM options, only call options data, and only put options data respectively. The sample period is from Nov. 17, 2008 to Dec. 15, 2008 across all regressions involved. These tables contain estimates and standard errors from four DVF candidates as well as model R-square, Root Mean Squared Error (RMSE), and log likelihood.

Table 2.4 displays the regression estimation results of four DVF model candidates on all IBM options, irrespective of their types. We find that all estimates are statistically significant across four DVF models. At the same time, their signs are consistent. Longer maturity tends to lower the implied volatility, which meets the expectation from the volatility smile plot. Moneyness, on the other hand, contributes positively to the implied volatility. Endogenous time volatility estimator as a new covariate serves a less prominent but still significant role to the implied volatility than moneyness and maturity. The pairwise comparison results of DVF 1, DVF 3 and DVF 2, DVF 4 show that although the magnitude of its estimates are smaller than the other covariates, adding endogenous time volatility estimator to DVF model will improve model fit and prediction power.

Table 2.4: DVF Estimation Results for All IBM Options Traded from 2008-11-17 to 2008-12-15

	DVF 1		DVF 2		DVF 3		DVF 4	
Parameters	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
Intercept	0.68518	0.01224	0.63875	0.023903	0.69408	0.01278	0.64796	0.02420
Maturity	-0.34619	0.01150	-0.24732	0.02913	-0.34514	0.01150	-0.24642	0.02912
Moneyness	0.20313	0.00842	0.22117	0.02642	0.20296	0.00842	0.2206	0.02641
Moneyness^2			0.01922	0.01236			0.01933	0.01235
Endogenous Volatility					-0.16503	0.06845	-0.16389	0.06836
Moneyness × Maturity			-0.08784	0.02455			-0.08774	0.02453
Observations	4,458		4,458		4,458		4,458	
R-Square	0.200		0.203		0.201		0.204	
RMSE	0.374		0.374		0.374		0.374	
Log Likelihood	-1,946		-1,938		-1,942		-1,935	

Table 2.5: DVF Estimation Results for IBM Call Options Traded from 2008-11-17 to 2008-12-15

	DVF 1		DVF 2		DVF 3		DVF 4	
Parameters	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
Intercept	0.65956	0.01472	0.61472	0.02791	0.67000	0.01542	0.62552	0.02832
Maturity	-0.33298	0.01464	-0.21738	0.03532	-0.33198	0.01463	-0.21718	0.03529
Moneyness	0.24627	0.01289	0.25973	0.03896	0.24606	0.01288	0.25901	0.03893
Moneyness^2			0.027856	0.01570			0.027806	0.01569
Endogenous Volatility					-0.19169	0.08517	-0.18786	0.08499
Moneyness × Maturity			-0.10954	0.03118			-0.10884	0.03116
Observations	2,576		2,576		2,576		2,576	
R-Square	0.213		0.217		0.215		0.219	
RMSE	0.354		0.353		0.353		0.353	
Log Likelihood	-975.5		-968.7		-972.9		-966.3	

Table 2.5 shows the regression results based on IBM call options only. On the other hand, Table 2.6 is generated based on IBM put options only. Their estimation results are generally

similar to Table 2.4. These results repeatedly manifest that the endogenous time volatility factor adds small contribution to better DVF model fit. It is worth noting that endogenous time volatility factor is statistically insignificant if we only regress on put option data over this particular period.

Table 2.6: DVF Estimation Results for IBM Put Options Traded from 2008-11-17 to 2008-12-15

	DVF 1		DVF 2		DVF 3		DVF 4	
Parameters	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.	Estimates	S.E.
Intercept	0.57070	0.01778	0.50600	0.03778	0.57605	0.01846	0.68662	0.04589
Maturity	-0.36281	0.01833	-0.29311	0.05025	-0.36182	0.01835	-0.29148	0.05027
Moneyness	0.19322	0.01643	0.21363	0.05595	0.19306	0.01643	0.21306	0.05595
Moneyness^2			0.010882	0.02039			0.011153	0.02039
Endogenous Volatility					-0.12665	0.11246	-0.12828	0.11245
Moneyness × Maturity			-0.05758	0.04017			-0.058133	0.04016
Observations	1,882		1,882		1,882		1,882	
R-Square	0.190		0.191		0.190		0.192	
RMSE	0.4		0.4		0.4		0.4	
Log Likelihood	-946.8		-945.6		-946.2		-944.9	

Table 2.7: Economic Value of DVF Models (on 200-Day Average)

	DVF 1	DVF 2	DVF 3	DVF 4
$MSE	1.6684	1.2741	1.7774	1.3464
RPA	78.15%	62.43%	79.33%	63.52%

The economic value of our improved DVF model is reflected by $MSE and RPA (see Table 2.7). $MSE and RPA are calculated independently on each day to measure back testing performance. This table illustrates the average $MSE and RPA values over 200 trading days from Feb. 4, 2008

to Nov. 14 2008. By comparing predicted option prices with their market price, we find that DVF with endogenous time volatility factor helps a little regarding in-sample fit. However, the original DVF model still beats our proposed model in terms of out-of-sample forecast performance.

2.5 Conclusion

This paper explores an application of the so-called endogenous time volatility estimator developed by Pelletier and Weng (2015). In their paper, they fixed the bias issue of conventional equity volatility estimators, *e.g.* realized volatility estimator, realized kernel volatility estimator, caused by the endogeneity in transaction times. We look into the utility of incorporating it with the Deterministic Volatility Function (DVF) framework by Dumas, Fleming, and Whaley (1998). Their approach is designed to circumvent the issue that implied volatility derived from Black-Scholes model often performs poorly in practice due to the assumption of underlying asset volatility being constant over the option lifetime. We add the endogenous time volatility of the underlying asset as additional covariate to the DVF model. In the paper, we briefly specify the continuous-time model of the endogenous time volatility estimator. We employ the Kalman smoother algorithm to obtain the estimator. Then we integrate the endogenous time volatility estimator into four DVF pairwise comparison models. For empirical analysis, we apply IBM option and high-frequency equity data in 2008. We find that DVF with endogenous time volatility factor helps a little regarding in-sample fit. However, the original DVF model still beats our proposed model in terms of out-of-sample forecast performance.

CHAPTER 3

New Technology-Industry Concordances Using Linked Micro-Level Data on Patents and Firms

3.1 Introduction

Patent data provides unique aspects for studying the economics of innovation, for example by allowing the study of public goods aspects and dynamics of knowledge spillovers or by measuring the ex-post effects of industrial policies or price changes on induced innovation (see Calel and Dechezleprêtre, 2014; Dechezleprêtre and Glachant, 2013; Jaffe et al., 2000; Johnstone et al., 2010). Most innovation policy research using patents and economic data either uses macro-level aggregate data on patent counts by country-year (see Johnstone et al., 2010) or firm-level data (see Calel and Dechezleprêtre, 2014). However, there has also been a persistent need for better "meso-scale" data on innovation activity at the industrial level (see Lybbert and Zolas, 2014). Such meso-scale data is important for industrial innovation policy (see Lybbert and Zolas, 2014; Warwick and Nolan, 2014), because they permit analysis of innovation policy across different economic sectors without requiring detailed enterprise-level data (which do exist, but often with

limited coverage). Applications of technology-industry concordance in literature include the study of innovation in global value chains study (see Lybbert and Zolas, 2014) and analyses of geographic and sectoral clustering of innovation activity (see Johnson and Brown, 2004). This need has led to efforts to map the types of technologies embodied by patented inventions and the industries to which they are related. To meet this need, a number of technology-industry concordance systems have been developed to link technology classification systems for patents to industrial classification systems for enterprises.

When designing concordance systems, it is important to precisely define what is meant when we say patented inventions in a given technology class are *linked* or *used* in a given industry. Existing concordance systems have defined these linkages based on expert assessment and machine learning methods applied to textual descriptions of the technologies and industries. Expert assessment was the original means of generating these concordances - most notably in the Yale Technology Concordance (YTC), discussed below (see Kortum and Putnam, 1997) - and consists essentially of reading patent documents individually or descriptions of a whole class of technologies, then reading the descriptions of economic sectors in industrial classification systems, and subjectively judging which industries are most relevant for the patents in a given technology area. The primary utility of this approach is that it allows the tacit knowledge of experts to be used in defining these linkages. The downsides of this approach are the inherent subjectivity and labor intensiveness of this process. This in particular limits the ability to keep such concordances up-to-date, as new technology classes are introduced. For example, the YTC is based on a fixed sample of Canadian patents examined between 1978 and 1993. While newer concordances using this methodology have been developed (discussed below), both subjectivity and and labor intensiveness remain issues that may introduce bias and limit the saliency of such concordances in an environment with rapidly increased rates of patenting (see Kortum and Lerner, 1999; Li, 2012).

In response to the limitations of concordances based on expert assessment, text-mining methods have been proposed to automate the process of assigning patents - and thereby sets

of technologies - to different industries. Of particular note, Lybbert and Zolas propose an "algorithmic links with probabilities" (ALP) method, in which patent title and abstract text is mined for terms associated with economic sector descriptions in official industry classification systems. This text-mining approach provides computer-generated indicators linking individual patents to industries. Lybbert and Zolas then use these generated data to computer technology-level frequencies of patents belonging to different industries to impute Bayesian conditional probabilities of patent belonging to a given technology also belong to a given industry (and vice versa).

However, patents may be linked to different industries using information other than textual descriptions of technologies and industries. Of particular relevance are the applicants listed on patent documents and their associated industries. Patent applicant information has been digitized and readily available in statistical databases since the 1990s, but only recently has a concerted effort been mounted, by the USPTO, EPO, WIPO, NBER and the OECD among others, to categorize the type of applicant in global patent databases and - in the case of private enterprises - the economic sectors which they belong. This type of information provides an independent measure of the technology-industry relationship. Patent applicants are the parties that profit from the conferred intellectual property rights, and so, for example, knowing that the agricultural seed and biotech company Monsanto (ISIC and NACE code = 0130) has many patents associated with the IPC category for biotechnology suggests that some proportion of the patents in this IPC code correspond to the agricultural seed or biotechnology industries. This method of categorizing patented inventions into different industries does not rely on how the patent is described in the title or abstract (which applicants may intentionally obfuscate), but rather depends on whether a patent has some revealed economic value to a given industry based on an enterprise in that industry having found it worth it to expropriate the intellectual property.

In this paper we use global linked patent-enterprise data from the EPO and the OECD to study technology-industry concordances. We conduct two types of analysis. First, we investigate

whether and to what extent previous concordances, based on textual analysis, predict technology-industry linkages based on applicant-level data. Second, we construct two new concordances based on applicant-linked data: a Raw OECD Counts (ROC) concordance and an OECD Bayesian Averaged Concordance (OBAC), which uses Bayesian Model Averaging (BMA) to increase the predictive power of our concordance system. The motivation for using the BMA method in the OBAC weights is the recognition that multiple IPC codes jointly predict industry membership, but that including 643 regressors in a single regression presents an infeasible and unreliable estimation problem. BMA is a method designed for such situations with large numbers of predictor variables, selecting subsets of these variables in an array of regression models and then updating the forecasting weight placed on any given model using Bayes' rule. As we show in our results, this theoretically motivated (as opposed to ad-hoc) methodology provides more precision in probabilistic assignment of technology classes to firms (and vice versa).

The rest of this paper is organized as follows. In Section 2, we provide background on industry-technology concordances for patent data. In Section 3.3, we give a detailed discussion of the proposed algorithms based on linked patent-firm data. Section 3.4 contains a summary of the databases we use as well as the data cleaning procedure. Section 3.5 presents empirical findings. Section 3.6 concludes.

3.2 Background on Technology-Industry Classification Systems, Existing Concordances, and Their Uses

The dominant technology classification system for patents, and the one we use in this paper, is the International Patent Classification (IPC) system. In contrast, there are a myriad of systems for classifying enterprises into economic sectors. Based on the databases we employ, the Nomenclature of Economic Activities (NACE) and the International Standard Industrial Classification (ISIC) are the systems used in this analysis. In this section we briefly describe these systems, before discussing the different concordances - primarily, the ALP and YTC concordances - that have

been developed to match them.

3.2.1 IPC Technology Classification System and PATSTAT

The IPC system has been the dominant means of classifying patents according to the type of technology they embody. The system is designed for assessing the legal property rights claims of patent applications. Patent examiners use the IPC system to identify "prior art" in related technology areas when considering patent applications for new inventions, and what aspects of these inventions are novel. The IPC system was adopted by all major patent offices in the world in the 1970s as part of the Strasbourg Agreement, an international treaty administered by the World Intellectual Property Organization (WIPO). [1]

The IPC system classifies patents hierarchically, with eight sections at the top level of the hierarchy (*e.g.* Section C corresponds to "Chemistry; Metallurgy") and approximately 70,000 subdivisions at the lowest level. These subdivisions are 8 digit codes using letters and numbers (*e.g.* C12N 15/09 corresponds to "Mutation or genetic engineering using recombinant DNA technology"). Each patent document is classified by the examiner according to these subdivisions, and patents frequently have multiple IPC classifications. Following prior literature, we focus our analysis on a middle level of the IPC hierarchy: the 4-digit IPC classifications, of which there are approximately 650 (643 distinct codes in the patent database used in this paper, see below). The IPC system is continually revised, with new technology classifications added as needed and previous patents reclassified in the IPC system as new classifications are introduced. [2]

3.2.2 Industrial Classification Systems

Industrial classification systems are used for categorizing enterprises to specific industries. In contrast to the IPC system, industrial classification systems are designed to be used for statistical analysis, *e.g.* by government agencies and research organizations. The dominant

[1] Details of IPC system can be found on WIPO website: `http://www.wipo.int/classifications/ipc/en/`.

[2] The IPC system document: `http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf`.

industrial classification systems include the International Standard Industrial Classification (ISIC) administered by the United Nations, [3] the Nomenclature of Economic Activities (NACE) system administered by Eurostat, [4] and the Standard Industrial Classification (SIC) system and the North American Industrial Classification System (NAICS) used in North America. [5]

In this paper we focus first on the NACE (Revision 2) system, because this is the primary one used in the economic database analyzed here (see below). We also include the ISIC system (Revision 4), because the ISIC system is used in the main previous technology-industry concordance to which we compare our approach and because the ISIC system is more widely recognized outside of Europe. Both systems classify enterprises to industries according to primary (or textitcore), secondary and ancillary activities, with these activities determined by the economic value added to the enterprise from each activity and ranked via a hierarchical "top-down" method (see NACE document from Eurostat). A relatively straightforward concordance exists between NACE and ISIC, [6] permitting this comparison. The highest hierarchy of the ISIC system is coded in letters from "A" to "U". Lower levels of the hierarchy consist of 2-, 3-, and 4-digit codes. By design, NACE is a derivation of ISIC: Categories at all levels of NACE are defined either to be identical to, or to form subsets of, single ISIC categories. The first level and the second level of ISIC Rev. 4 (sections and divisions) are identical to sections and divisions of NACE Rev. 2. The third and fourth levels (groups and classes) of ISIC Rev. 4 are subdivided in NACE Rev. 2 according to European requirements. However, groups and classes of NACE Rev. 2 can always be aggregated into the groups and classes of ISIC Rev. 4. The aim of the further breakdowns in NACE Rev. 2, is to obtain a classification more suited to the structures of the European economies. Regarding the way agencies update ISIC and NACE tables, the correspondence table are available online.

[3] To explore ISIC system with complete notes see: `http://unstats.un.org/unsd/publication/seriesM/seriesm_4rev4e.pdf`.

[4] Detailed document about NACE system is located at `http://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF`.

[5] The U.S. Census website on SIC/NAICS: `https://www.census.gov/eos/www/naics/`.

[6] The official correlation document: `http://unstats.un.org/unsd/cr/registry/regso.asp?Ci=70`.

[7] [8] Firms in Orbis be classified across multiple NACE codes (at the 4-digit level). Additional concordances exist between ISIC and SIC or NAICS, [9] but for concision we do not delve into those concordances here.

3.2.3 Existing Technology-Industry Concordance Systems

The primary purpose of technology-industry concordance systems is to permit inference about how patenting activity within one IPC technology category relates to different industries. Such concordances can loosely be thought of as answering the following question: If the only information available for a given patent was that it belonged to IPC technology code t, does it belong (or how likely is it to belong) to industry i? Answering this question enables the analyst to use the aforementioned meso-scale data such as IPC-level patent counts (easily accessible in online databases, such OECD.Stat) in economic analysis by allocating patent counts to different industries, without having to go through the significantly more time-consuming process of analyzing individual patents.

Existing technology concordances focus on matching technologies to industries using textual analysis of patent documents and industry descriptions. Originally, this matching was manually done by patent examiners or other experts. This is the case for the majority concordances: the YTC (Kortum and Putnam, 1997), the "DG Concordance" (Schmoch et al., 2003), the OECD concordance.[10] The key difference between these approaches are first that the YTC, in particular, uses the Canadian Standard Industrial Classification (cSIC) to develop matches with the IPC system, whereas the other concordances use the more widely applicable ISIC system. The other key difference is that the YTC uses probabilistic matching between technology and industry, whereas the other approaches attempt to generate deterministic one-to-one correspondences

[7] The correspondence between ISIC Rev.4 and ISIC Rev.3.1: `http://unstats.un.org/unsd/cr/registry/regso.asp?Ci=61`.

[8] The correspondence between NACE Rev. 2 and NACE Rev. 1.1: `http://ec.europa.eu/eurostat/documents/1965800/1978760/CORRESPONDENCETABLENACEREV.2NACE-REV.1.1.pdf/df9cd8a8-0b4a-4197-bad7-727a0b9fd59b`.

[9] Other industry system concordances are introduced on the U.S. Census website: `https://www.census.gov/eos/www/naics/concordances/concordances.html`.

[10] `http://www.wipo.int/edocs/mdocs/classifications/en/ipc_ce_32/ipc_ce_32_10-annex1.doc`.

based on industry and technology descriptions. As mentioned above, the use of expert assessment in all of these approaches means that these concordances are inherently subjective but also that they are likely to include tacit knowledge that may be omitted from automated approaches. See Lybbert and Zolas (2014) for a more detailed descriptions of these approaches.

The only automated, computer-based approach that has heretofore been developed is the ALP concordance (Lybbert and Zolas, 2014). This approach first applies text-mining techniques to analyze patent abstracts in relation to industry descriptions in ISIC (and, potentially, other industrial classification systems). Using these methods, they generate patent-specific industry classifiers. From these individual classifiers, Lybbert and Zolas then generate concordance weights at the 3- and 4-digit IPC levels by computing raw, "specificity," and "hybrid" weights based on conditional and unconditional patents within technology and/or industry classes. The raw weights are simply conditional frequencies, *e.g.* the percentage of biochemistry (IPC4 code = C12N) patents that are linked to a given industry, such as "research and development on biotechnology" (ISIC=7210). In analyzing the raw weights generated from this procedure, Lybbert and Zolas note that some technology classes relate to a more diverse set of industries, whereas other technologies are more specific to a small subset of industries. To account for this specificity, they modify the raw weights using Bayes rule in combination with a uniform prior distribution of the likelihood that a given industry associates with all IPC codes. This approach generates posterior weights which downweight IPC codes with many patents relative to IPC codes with fewer patents. Finally, the hybrid weights are an ad-hoc blending of the raw and specificity weights. Lybbert and Zolas finally impose ad hoc cutoff of 2% , below which the weights are set to zero and normalized. This is to reduce very frequency, obviously false positives (Type I errors).

One can download all existing ALP concordances tables from the WIPO website. [11] The specific ALP concordance table file we use for comparing our results is "isic_rev4_4_to_-ipc4.txt". The ALP weights exhibit an (intentional) pattern in which only a few technology classifications have positive weights for a given industry classification. For example, under ISIC

[11] http://www.wipo.int/econ stat/en/economics/publications.html.

code: 111 ("Growing of cereals (except rice), leguminous crops and oil seeds"), ALP weights are positive for only 24 out of the 643 4-digit IPC codes. We show results comparing our weights (described below) to ALP weights in section 3.5.1.

3.3 Proposed New Concordance Algorithms

The two new concordances we propose in this paper introduce two innovations, one related to the underlying data and the other to the statistical methodology. First, we generate new industry classifiers of individual patents based on applicants' core industries, instead of using text-mining approaches. Second, we introduce BMA as a statistical forecasting tool for isolating the relevant subsets of IPC codes for a given industry, this providing a more systematic approach of dealing with the Type I error problem described by Lybbert and Zolas. Being based on OECD data (described below) matching patent applicants to entities, we title these new weights a *Raw OECD Concordance* (ROC), which calculates simple conditional frequenceies like the raw ALP weights but using the new applicant-based classifier, and the *OECD Bayesian Averaged Concordance* (OBAC), which uses the new classifiers and BMA tools. In this section, we describe the formulas and methods used in these weights, before describing the underlying data used in the next section.

The ROC weights employ the same formula as the raw ALP weights (just using different industry classifiers), computing the ratio of all patented inventions [12] in a given IPC (4-digit) code *and* in a specific industry divided by the total number patent families in that IPC code. Thus, the ROC likelihood that a patent with IPC code t belongs to industry i is:

$$\omega_{i|t}^{\text{ROC}} \equiv \frac{n_{i,t}}{n_t} \tag{3.1}$$

where $n_{i,t}$ is the total number of patent families with a particular technology code t and linked to the industry i. And n_t is the total number of patent families classified as technology t.

[12] We define inventions based on patent family with all families share a particular IPC (4-digit) code.

The OBAC weights are calculated using BMA. This forecasting method incorporates model uncertainty and usually improves predictive power of regression-based approaches substantially as compared to any single regression model (see Hoeting et al., 1999). Instead of choosing one dominant regression model, the BMA algorithm exhausts all possible combinations of the regressors, in our application consisting of subsets of the 643 IPC4 codes. The basic idea of BMA algorithm is as follows: First, all possible regression models are assigned a prior probability of being the true model. Second, Bayesian computational methods are used to iteratively sample different models and compute approximate posterior model probabilities and means for coefficients of all covariates. Based on several different information criterion, the algorithm generates weights of each model (in our case, combinations of IPC codes) to construct the posterior mean. The basic mathematical object at the foundation of BMA methods is the predictive probability density $f(y|D)$ for outcome y (in our case, industry classifiers) conditional on data D (in our case IPC codes):

$$f(y|D) = \sum_{k=1}^{K} f_k(y|D)\text{Prob}(M_k|D),$$

where $f_k(y|D)$ is the k-th predictive density. This is an average of the posterior distribution under each of the models considered, weighted by their posterior model probability, $\text{Prob}(M_k|D)$. $M_1, \cdots, M_K$ are the models considered. The posterior probability for model M_k is given by:

$$\begin{aligned}
\text{Prob}(M_k|D) &= \frac{\text{Prob}(D|M_k)\text{Prob}(M_k)}{\sum_{l=1}^{K}\text{Prob}(D|M_l)\text{Prob}(M_l)}, \\
\text{Prob}(D|M_k) &= \int \text{Prob}(D|\theta_k, M_k)\pi(\theta_k|M_k)d\theta_k,
\end{aligned}$$

where θ_k are the model k's parameters, and $\pi(\theta_k|M_k)$ is the prior for θ_k in model k. It is worth noting that BMA algorithm will still be computationally heavy when the number of covariates is large, and the standard approach is the therefore to set a maximum for the number of covariates that can be included in any single model (see Hoeting et al., 1999, for details).

We implement BMA using a logit regression model for the predicted probability that patent

j belongs to industry i, conditional on patent j's IPC codes:

$$\text{Prob}(y_j^i = 1) = \frac{1}{1 + \exp(-\sum_t \beta_{i,j}^{(M_k)} d_j^t)}, \tag{3.2}$$

with:

$$y_j^i = \begin{cases} 1 & \text{if patent } j \text{ is in industry } i \\ 0 & \text{otherwise} \end{cases},$$

$$d_j^t = \begin{cases} 1 & \text{if patent } j \text{ has a IPC code } t \\ 0 & \text{otherwise} \end{cases}.$$

and where (M_k) denotes the k-th logit model candidates used to predict the probability of a patent coming from industry i given its technology class t. Each logit model candidate has different combination of covariates with associated coefficient vector $\beta_{i,j}^{(M_k)}$. The posterior probability p_{M_k} is the normalized Bayesian weights assigned to each regression model. We restrict and normalize these weights to be positive for only the best 5 models, based on their BIC.

Once we obtained the posterior model weights $p_{(M_k)}$ and predictive coefficients $\beta_{i,j}^{(M_k)}$ from BMA, we calculate the OBAC weights as the weighted mean predicted probabilities of belonging to industry i conditional on belonging to technology t:

$$\omega_{i|t}^{\text{OBAC}} \equiv \sum_{k=1}^{K} \frac{p_{(M_k)}}{n_t} \sum_{j \in \Omega_t} F\left(\beta_0^{(M_k)} + \beta_t^{(M_k)} + \sum_{s \neq t} \beta_s^{(M_k)} d_j^s\right), \tag{3.3}$$

where K is the number of models with positive posterior weights in the BMA, $F(\cdot)$ is the logit function defined in eq.(3.2) and Ω_t is the set of all patent families belong to a particular technology t (which has size n_t).

3.4 Data Preparation

We first describe the databases used in this analysis, before describing some additional data preparation procedures that were necessary to generate the concordances described above.

3.4.1 Data Sources

Our patent data come from the the European Patent Office's (EPO's) Worldwide Patent Statistical Database (PATSTAT). This database contains bibliographic information for approximately 70 million patent applications from around 90 countries and more than 100 patent offices going back at least to 1990 (and often much longer). The PATSTAT version we use in this paper was issued in April, 2014. In contrast to online web-search tools for finding individual patents, PATSTAT is intended for large-scale statistical analysis. The entire database consists of over 20 tables, including patent titles and abstracts, filing and granting dates, citation information, inventor and applicant information, technology classifications (IPC codes), as well as patent family identifiers (since multiple applications are usually related to a single invention, e.g. with applications being denied and then resubmitted or applications being submitted in multiple offices for the same invention).

In this paper we use the technology classifications, applicant information and family identifiers from PATSTAT. To group patents into families, we use identifiers from the EPO's master documentation database, DOCDB. All of our analysis is done at the family-level, meaning for example that each observation j in equation 3.3 is a patent family (not a single application). The technology classification tables are straightforward, with all IPC codes associated with any applications in a given family being extracted and linked with that family. Patent applicant data in PATSTAT correspond to the original applicants for a given application. Given our objectives, we use all applicants on any application in a given family to tie that family to industries using firm-level data described below. Original applicant identifiers and names in PATSTAT require a significant amount of cleaning, with slight variations in names for the same applicant often leading to a different applicant identifier. To deal with this issue, the OECD has created a

separate Harmonized Applicant Name (HAN) table, which we use here (see Thoma et al., 2010, for details). We use OECD HAN database, Feb. 2015 version, which provides a grouping of patent applicants names resulting from a cleaning and matching of names. Names of applicants were originally extracted from EPO's PATSTAT, October 2011, and the results were propagated to the subsequent editions of PATSTAT (2012, 2013 and 2014) to enhance the coverage.

To link patent applicants to industries, we use databases from the OECD's Microdatalab (Squicciarini et al., 2013). This resource links PATSTAT to a limited version of the ORBIS commercial firm-level database from Bureau Van Dijk (BvDEP). Ninety-nine percent of firms it covers are private companies. The ORBIS database is not an exhaustive database of all companies around the world, although the aim of BvDEP is to increase its coverage in all countries (see Ribeiro et al., 2010). We use the OECD version of the Orbis database in this paper available for use by researchers via a cooperative data-sharing agreement between BvDEP and OECD between 2008 and 2011. The OECD versions of the database differ in many respects to the raw data, and in this paper we employ only the industry classification information from the database. For more information on the structure and format of this database (see Ragoussis and Gonnard, 2012). For our purposes the key information in OECD's Microdatalab is a table linking HAN identifiers to firm names and identifiers in Orbis, which contains industry classifiers for each firm (using the NACE Rev. 2 system).

To link two databases, we perform all data queries through MySQL and SAS. Throughout the data-cleaning step, two major restrictions are imposed. First, we exclude all patents granted to universities from the PATSTAT database. Second, all patent families being granted after 2008 as well as incomplete records have been discarded in the data cleaning step. The reason why we choose 2008 as the cutoff year is because OECD Orbis database we used is the 2008 version.

3.4.2 Additional Data Processing

After matching the patent and firm databases, we obtain a raw sample of around 3.9 million observations (patent families). This sample contains 643 unique IPC4 codes and 171 unique

NACE codes. To estimate the OBAC concordance for the full dataset requires that we estimate equation 3.2 for each of these 171 industries, including all of the 643 IPC4 codes as potential regressors in each of these regressions. This equates to $171 \cdot 2^{643} \approx 6.2 \cdot 10^{195}$ possible regression models. Even using BMA to discriminate among potential subsets of regressors, this universe of possible models is too large to exhaist in estimation, requiring some additional data trimming. We first restrict the sample to those observations having at least one industry code and one technology class, which reduces the sample to around 2.3 million observations. We then discard those IPC4 codes with a frequency in the sample of less than 1% and those NACE codes with a frequency of less than 0.5%. This leaves our final OBAC concordance having 37 IPC4 codes 57 NACE codes. [13] The table of these retained IPC4 and NACE codes can be found in our website as supplementary materials.

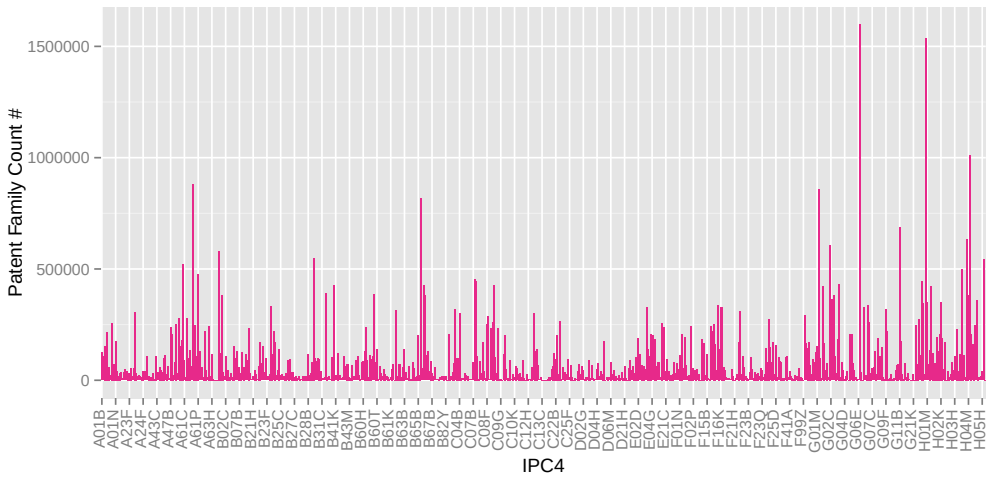
3.4.3 Sample Validity

Although Orbis has broad global coverage of companies, the patents linked to it represent 8% of the patent families in PATSTAT. To assess potential sample selection issues with our proposed concordances, we provide compare the subsample used in our analysis with the full PATSTAT database.[14] Figure 3.1 illustrates that the IPC4 technolohgical and geographic distributions in the raw PATSTAT population are very similar to the final IPC4 distribution in the matched sample. There appears to be no important IPC4s missing from the matched sample. In Figure 3.2 we also compare the top 15 most-requent IPC4 codes in the raw PATSTAT with that in the matched sample in Figure 3.2, panels (a) and (b). Among the 10 largest IPC4 codes, the two plots are essentially the same. Furthermore, regarding the geographic representation (by country) of the applicants, the map based on the raw PATSTAT is similar to the scale-adjusted map on the matched sample as displayed in Figure 3.1 panels (b) and (d). Fourthly, the matched data has very good global coverage, as seen in Figure 1, panel (d). Firm-level data from North

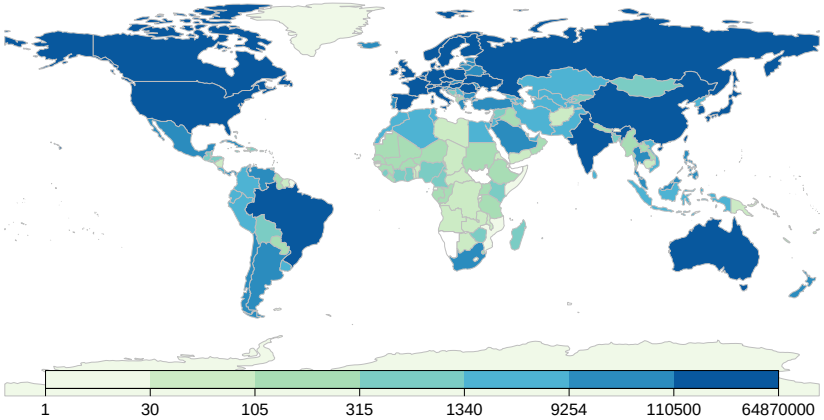
[13] In the future, we will seek to expand the number of included IPC codes, by reducing the cutoff 0.5% or less, which dramatically increases computational time.

[14] See Squicciarini et al. (2013) for a more exhaustive discussion of sample representativeness using the Microdatalab matched subsamples.

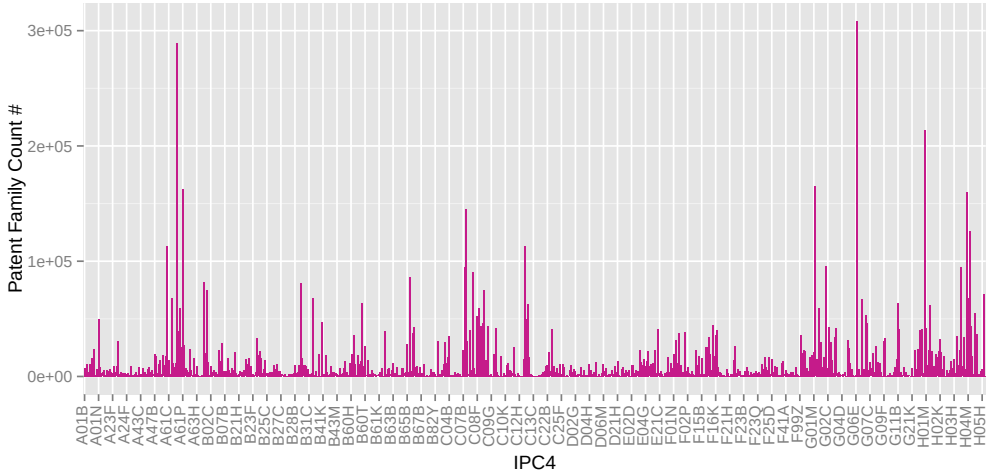
America, Europe and East Asia has better coverage than that in India and Brazil. Finally, from Figure 3.2 panel (c), we see that "Manufacture of pharmaceutical preparations" industry appears to associated with the greatest frequency in patenting (which is not surprising), followed by "Manufacture of instruments and appliances for measuring, testing and navigation" and "Manufacture of electronic components" respectively. Overall, we conclude that this provides evidence that the matched data is sufficiently representative of the raw PATSTAT database, at least for the purpose at hand.



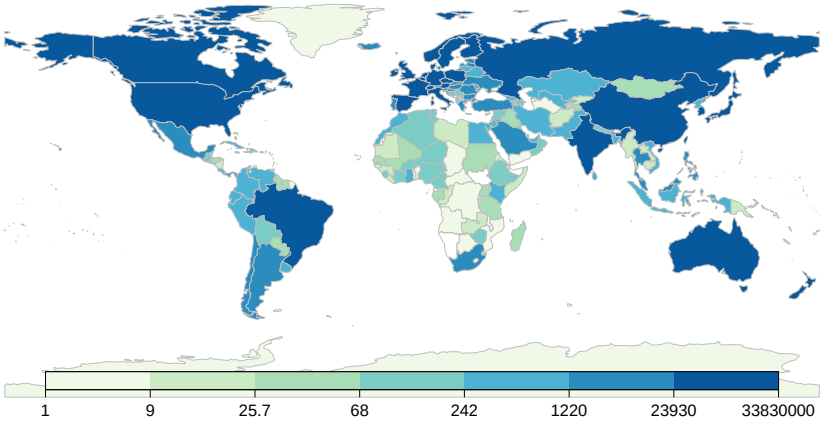
(a) PATSTAT Patents IPC4 Distribution



(b) PATSTAT Patents Geographical Distribution

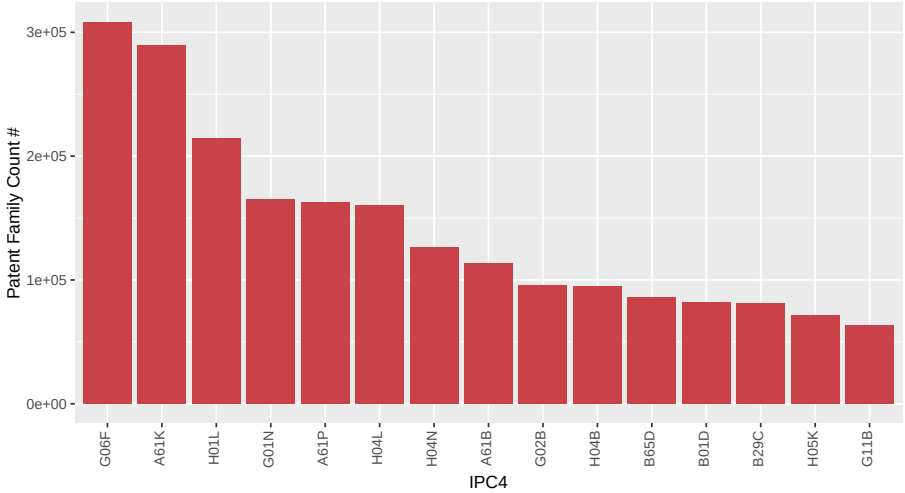


(c) Matched Patents IPC4 Distribution

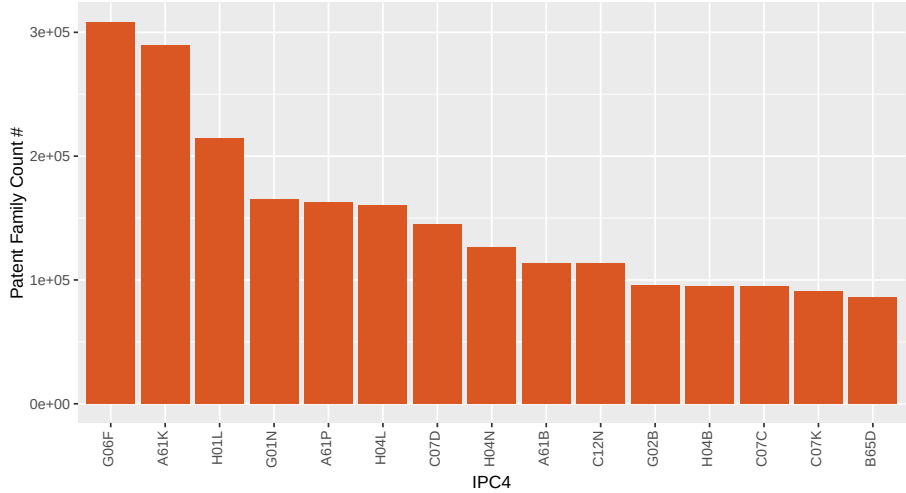


(d) Matched Patents Geographical Distribution

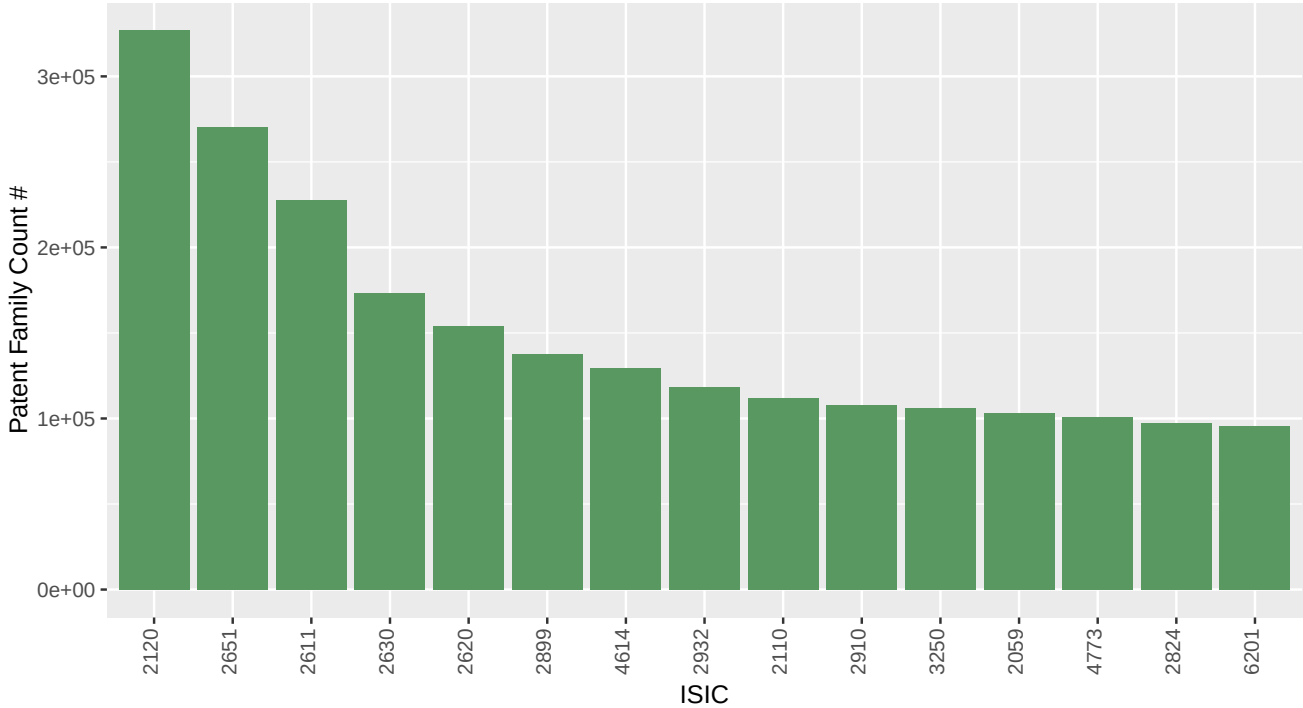
Figure 3.1: PATSTAT Vs. Matched: IPC4 Distribution and Patents Geographical Distribution Patterns



(a) TOP 15 IPC4 Distribution from All PATSTAT Patents



(b) TOP 15 IPC4 Distribution from Matched Patents



(c) TOP 15 NACE Distribution from Matched Patents

Figure 3.2: TOP 15 IPC4 Areas and NACE Fields Plots

3.5 Results and Sample Output

We present the results of our analysis in two stages, first examining how well the existing ALP concordance predicts industry membership at the patent-level, based on our independent ASM data. We then present sample output for the two new concordances proposed.

3.5.1 Does the Prior ALP Concordance Predict Applicant Sector Matches?

We examine the predictive power of the ALP concordance by estimating a series of binary probability regression models, in which the patent-specific ALP weights are calculated based on the patent's listed IPC codes. These tailored ALP weights are then used as explanatory variables in the regression models. To assess the performance of the ALP concordance in predicting industry-technology links using applicant sector matching, we examine the sign and statistical significance of the estimated regression coefficients associated with these weights across the industries for which we have sufficient data. The general statistical approach is to estimate the following regression model:

$$\text{Prob}(y_j^i = 1) = \Phi(\beta_0^i + \beta_{\text{ALP}}^i \cdot \tilde{\omega}_j^i), \tag{3.4}$$

where

$$y_j^t = \begin{cases} 1 & \text{patent } j \text{ is in industry } i \\ 0 & \text{otherwise} \end{cases}.$$

$\Phi(\cdot)$ is the binary probability link function. For robustness we estimate logit, probit and linear probability model specifications for $\Phi(\cdot)$. The β's are regression coefficients, and $\tilde{\omega}_j^i$ is the tailored ALP weight associating individual patent j with industry i. The latter is calculated as the sum of all patent j's ALP weights in a given industry i, divided by the total ALP weights also condition

on industry i:

$$\tilde{\omega}_j^i = \frac{\sum_t d_j^t \omega_t^i}{\sum_t \omega_t^i}, \qquad \text{with} \qquad d_j^t = \begin{cases} 1 & \text{if patent } j \text{ has a IPC code } t \\ 0 & \text{otherwise} \end{cases}.$$

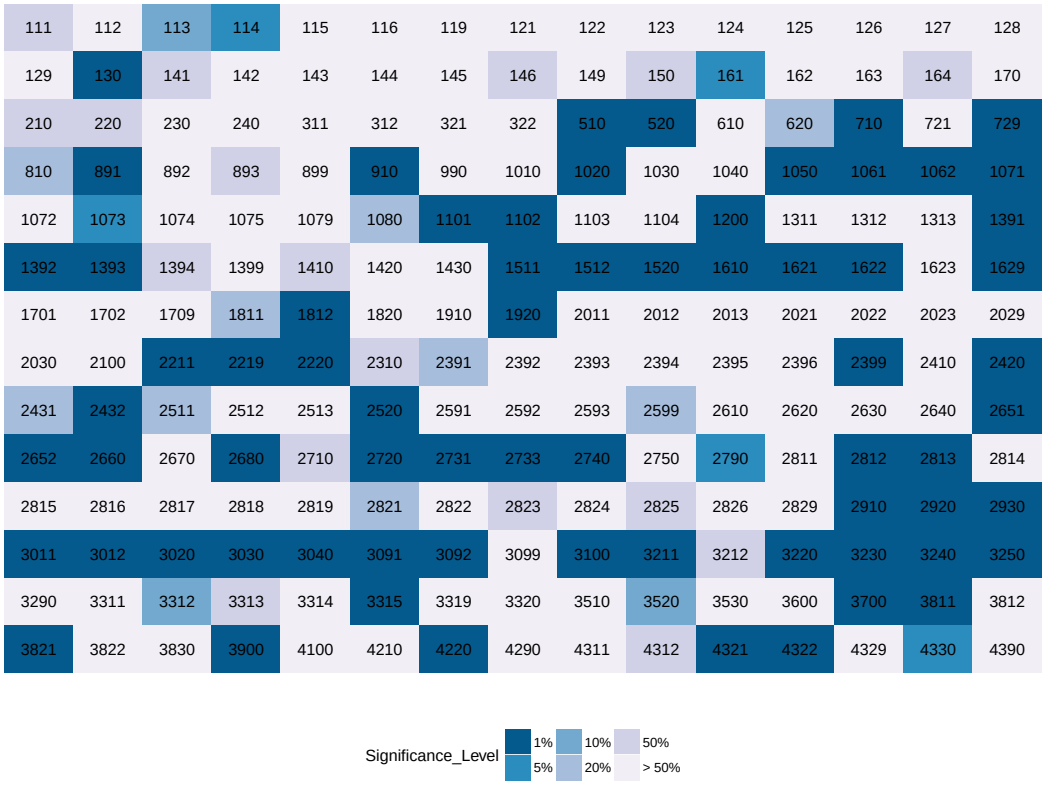


Figure 3.3: ALP Concordance Significance Test Results

Figure 3.3 shows the test results regarding ALP ISIC-IPC4 weights for the logit specification (results from the probit and linear probability specifications are very similar). ALP weights were only obtainable for ISIC4 codes up to "4390", which belongs to the last subcategory of the "F" section (referring to the construction industry). This means that our test captures roughly

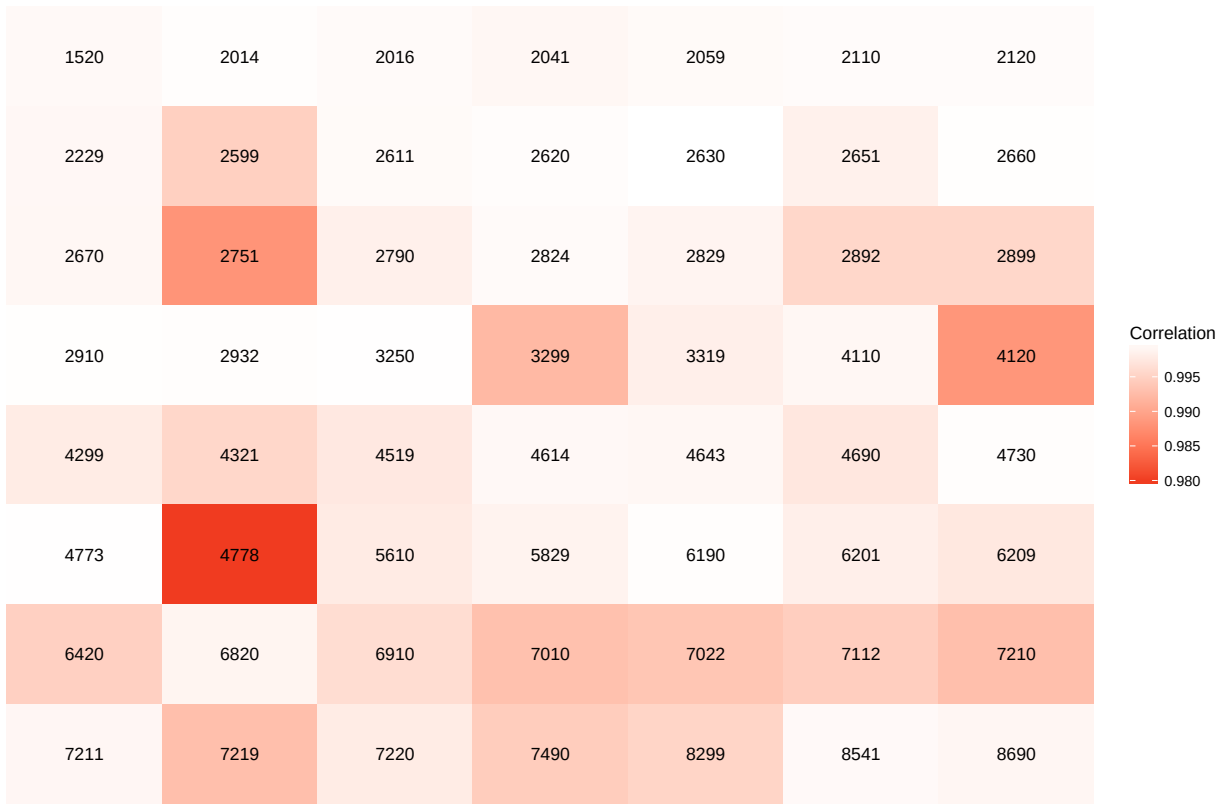
50% of the 419 ISIC4 codes. The darker a cell color is, the more statistically significance the corresponding ALP weight. The legend at the bottom indicates the significance levels revealed by the tests. We find only 73 out of 210 ALP ISIC-IPC4 weights, roughly 35%, achieve the 5% significance level. Additional details about the test results can be found from the appendix table.

3.5.2 Example Output and Patterns in Proposed Concordances

Table 3.1: Top Five Industry Areas For Example IPC Code (C12N: "Mutation or Genetic Engineering")

NACE	Label	ROC	OBAC
2120	Manufacture of pharmaceutical preparations	0.173	0.236
8541	Post-secondary non-tertiary education	0.118	0.160
7211	Research and experimental development on biotechnology	0.098	0.136
7220	Research and experimental development on social sciences and humanities	0.048	0.065
7490	Other professional, scientific and technical activities n.e.c.	0.037	0.051

To demonstrate the output from our proposed concordances, we begin with an example technology: genetic engineering. Table 3.1 lists the top 5 industries probabilistically linked to patent via its applicants, conditional on the patent having a IPC4 code corresponding to "micro-organisms; propagating, preserving or maintaining micro-organisms; mutation or genetic engineering; culture media." The the OBAC weights predict that such a patent has a 24% chance of being granted to a pharmaceutical company, a 16% chance to a university or a 13.5% chance to biotechnology research institution and so on. Figure 3.4 illustrates that within 56 different industry areas, ROC and OBC are extremely highly correlated irrespective of the algorithm difference. Even the least correlation indicated by dark red is 0.98 based on the firm level data. Light color shows that majority of ROC and OBC's correlation are greater than 0.995.



Note: the number in each cell is a NACE code

Figure 3.4: Correlation Between ROC and OBC

Another suggestive pattern observed in table 3.1 is that the OBAC weights (for the top 5 industries) are consistently larger than ROC weights. This pattern suggests that the OBAC weights may concentrate more predictive probability mass among the most likely industries or technologies. To confirm this pattern, we compute a measure of concentration, the Herfindahl index, for the ROC and OBAC weights. To measure the concentration of technologies in a given industry, the Herfindahl index formula becomes:

$$H_i = \sum_t w_{i|t}^2. \tag{3.5}$$

where $w_{i|t}$ is the weight (predictive conditional probability) for industry i conditional on technology t. An equivalent formula can be calculated measuring the concentration of industries across technologies. We calculate these indices for both the ROC and OBAC weights and then plot them against one another in figures 3.5 and 3.6. In 3.5 each dot corresponds to a unique NACE code. The x-axis corresponds to the Herfindahl index for the OBAC-based Herfindahl index for that NACE code, whereas the y-axis corresponds to the ROC-based Herfindahl index. The solid line corresponds to the 45 degree line, revealing that the OBAC weights are consistently more concentrated than the ROC weights. The same pattern emerges in Figure 3.6, where each dot instead corresponds to a technology. This provides a strong motivation for using the OBAC weights (and machine learning classifications in general) over simply using the raw data. Supplementary material with full ROC and OBAC tables are in appendix section.

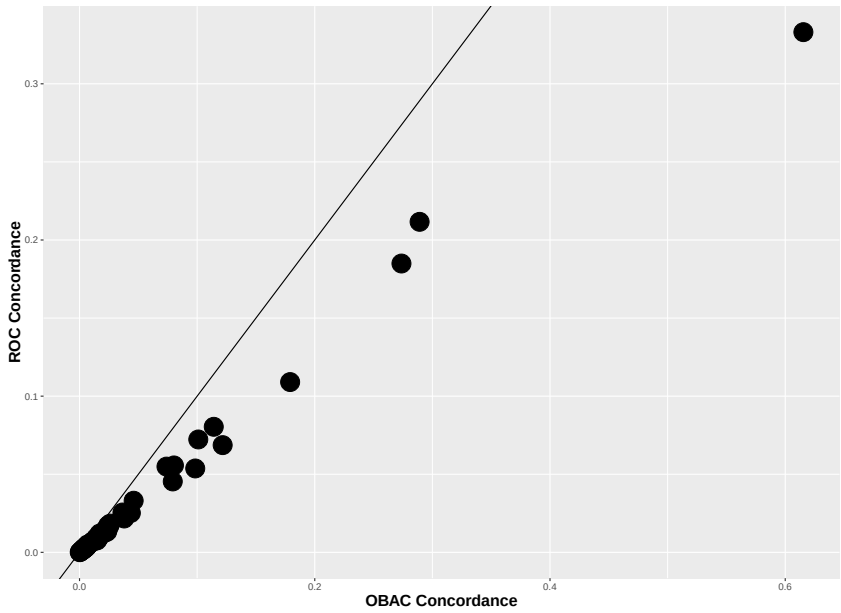


Figure 3.5: Herfindahl Index for Industry Sectors between OBAC and ROC

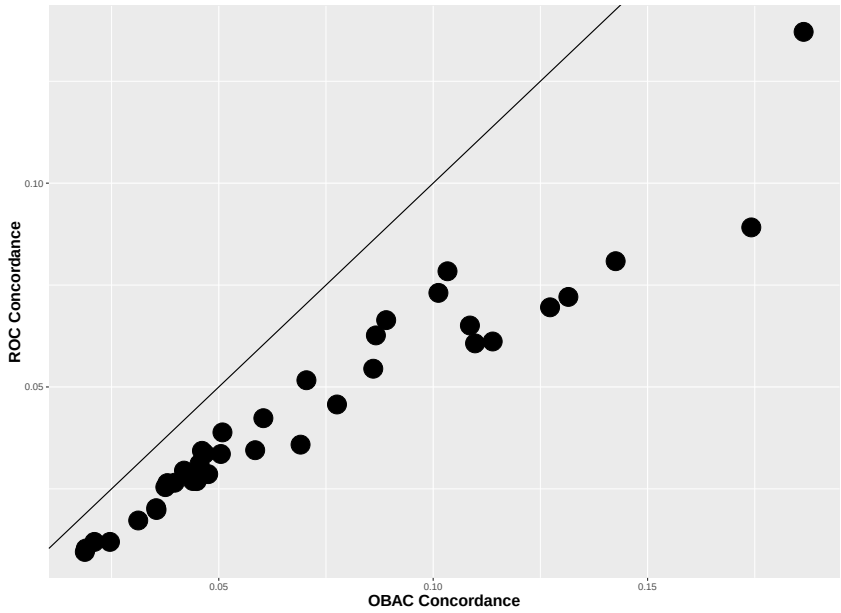


Figure 3.6: Herfindahl Index for Technology Classifications between OBAC and ROC

3.6 Discussion

There are two general results in this paper that have implications for future economic analysis using patent data. First, existing text-based concordances appear to frequently lack significant predictive power in identifying technology-industry links based on applicant sector matching (ASM). Second, machine-learning regression methods (here represented by BMA) appear to offer a more systematic way of probabilistically assigning technologies represented in patent data to different economic sectors.

The apparent orthogonality between textual analysis and applicant sector matching is a finding that deserves further investigation. While text-mining methods carry obvious qualifications (many discussed by Lybbert and Zolas), ASM has its own caveats. ASM should strictly be interpreted simply as the fact that an enterprise in a given industry filed for a patent with specific set of technology classifications. Such matches do not necessarily reveal an inherent value of the patent for that industry: a firm may have filed for a patent opportunistically, rather than because the

invention was integral to its core business. Our data do not show us whether this patent was commercialized or not, or whether the patent was sold to (and subsequently commercialized by) an enterprise in a different industry. Cases of the former instance may result in Type I errors that we simply cannot control for with our data. Cases of the latter instance would suggest Type II errors, i.e. failing to link patents to the industries into which they are sold. Concordances based on text-mining, being generated from entirely different data, are not subject to these issues. So it is likely that some links found by text-mining are still valid even when not confirmed by ASM (and vice versa). Both text-based and applicant sector matching incur measurement error. Given that the measurement error from each type of matching is likely orthogonal to the other, a logical direction for future research is to examine the combination of text-based and applicant sector matching.

The second contribution of this paper, the introduction of machine learning methods into technology-industry concordances, is one which may improve all algorithmic concordances, regardless of whether they are based on text-based or ASM. We find that these methods, at least in the case of BMA, tend to concentrate the predictive probability mass among the most likely industries or technologies, relative to the raw frequencies. This provides a policy motivation for adopting the OBAC over the ROC. Future work could explore whether this pattern persists when BMA (or another machine learning method) is applied to the text-based matches. Another possible use of machine learning in this context is to consider the intersection of multiple technologies as indicators of industry membership. Of course, 643 unique IPC4 codes implies over 200,000 possible two-way interaction terms that may included in any given regression - a number far too vast for coherent estimation. Yet machine learning methods are designed for isolating subsets of relevant regressors in such contexts.

REFERENCES

Abbring, J. H. (2012): "Mixed Hitting-Time Models," *Econometrica*, 80, 783–819.

Abramowitz, M. and I. A. Stegun (1965): *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.*

Aït-Sahalia, Y. and J. Yu (2009): "High Frequency Market Microstructure Noise Estimates and Liquidity Measures," *The Annals of Applied Statistics*, 3, 422–457.

Andersen, T. G. and L. Benzoni (2009): "Realized Volatility," in *Handbook of Financial Time Series*, ed. by T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen, Berlin, Heidelberg: Springer Berlin Heidelberg.

Andersen, T. G. and T. Bollerslev (1997): "Intraday Periodicity and Volatility Persistence in Financial Markets," *Journal of Empirical Finance*, 4, 115–158.

——— (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Bandi, F. M., J. R. Russell, and C. Yang (2008): "Realized volatility forecasting and option pricing," *Journal of Econometrics*, 147, 34–46.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2009): "Realized Kernels in Practice: Trades and Quotes," *Econometrics Journal*, 12, C1–C32.

——— (2011): "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of The Covariation of Equity Prices with Noise and Non-Synchronous Trading," *Journal of Econometrics*, 162, 149–169.

Barndorff-Nielsen, O. E. and N. Shephard (2001): "Non-Gaussian Ornstein-UhlenbeckâĂŘBased Models and Some of Their Uses in Financial Economics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*

——— (2002): "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 253–280.

——— (2004): "Power and Bipower Variation with Stochastic Volatility and Jumps," *Journal of Financial Economics*, 2, 1–37.

Bauwens, L. and D. Veredas (2004): "The Stochastic Conditional Duration Model: A Latent Variable Model for The Analysis of Financial Durations," *Journal of Econometrics*, 119, 381–412.

Bergman, Y. Z., B. D. Grundy, and Z. Wiener (1996): "General Properties of Option Prices," *The Journal of Finance*, LI, 1573–1610.

Bollerslev, T. (1986): "Generalized autoregressive conditional heteroskedasticity," .

——— (2001): "Financial Econometrics: Past Developments and Future Challenges," *Journal of Econometrics*, 100, 41–51.

Calel, R. and A. Dechezleprêtre (2014): "Environmental Policy and Directed Technological Change: Evidence from the European carbon market," *Review of Economics and Statistics*, 1–73.

Campbell, J. Y., A. W. Lo, and A. C. Mackinlay (1997): *The Econometrics of Financial Markets*, Princeton University Press.

Chen, F., F. X. Diebold, and F. Schorfheide (2013): "A Markov-switching Multifractal Inter-Trade Duration Model, with Application to US Equities," *Journal of Econometrics*, 177, 320–342.

Christoffersen, P. F., B. Feunou, K. Jacobs, and N. Meddahi (2014): "The Economic Value of Realized Volatility: Using High-Frequency Returns for Option Valuation," *Journal of Financial and Quantitative Analysis*, 49, 663ïijƊ697.

Christoffersen, P. F. and K. Jacobs (2004): "Which GARCH Model for Option Valuation?" *Management Science*, 50, 1204–1221.

Corsi, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7, 174–196.

De Jong, P. (1991): "The Diffuse Kalman Filter," *The Annals of Statistics*, 19, 1073–1083.

De Jong, P. and N. Shephard (1995): "The Simulation Smoother for Time Series Models," *Biometrika*, 82, 339–350.

Dechezleprêtre, A. and M. Glachant (2013): "Does Foreign Environmental Policy Influence Domestic Innovation? Evidence from the Wind Industry," *Environmental and Resource Economics*, 58, 391–413.

Derman, E. and I. Kani (1994): "The Volatility Smile and Its Implied Tree," *Goldman Sachs Quantitative Strategies Research.*

Diebold, F. X. and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Dumas, B., J. Fleming, and R. E. Whaley (1998): "Implied Volatility Functions : Empirical Tests," *The Journal of Finance*, LIII, 2059–2106.

Easley, D. and M. O'Hara (1992): "Time And the Process of Security Price Adjustment," *The Journal of Finance*, 47, 577–605.

Engle, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1007.

——— (2000): "The Econometrics of UltraâĂŘHighâĂŘFrequency Data," *Econometrica*, 68, 1–22.

Engle, R. F. and J. R. Russell (1998): "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162.

Fukasawa, M. (2010): "Realized Volatility with Stochastic Sampling," *Stochastic Processes and their Applications*, 120, 829–852.

Ghysels, E., C. Gouriéroux, and J. Jasiak (2004): "Stochastic Volatility Duration Models," *Journal of Econometrics*, 119, 413–433.

Hamilton, J. D. (1994): *Time Series Analysis*, Princeton University Press.

Hansen, P. R. and A. Lunde (2006): "Realized Variance and Market Microstructure Noise," *Journal of Business & Economic Statistics*, 24, 127–161.

Harvey, A. C., E. Ruiz, and N. Shephard (1994): "Multivariate Stochastic Variance Models," *The Review of Economic Studies*, 61, 247–264.

Harvey, A. C. and N. Shephard (1996): "Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns," *Journal of Business & Economic Statistics*, 14, 429–434.

Heston, S. L. (1993): "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options," *The Review of Financial Studies*, 6, 327–343.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. Volinsky (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417.

Hull, J. and A. White (1987): "The Pricing of Options on Assets with Stochastic Volatilities," *The Journal of Finance*, 42, 281–300.

Jacod, J. and P. E. Protter (1998): "Asymptotic Error Distributions for the Euler Method for Stochastic Differential Equations," *The Annals of Probability*, 26, 267–307.

——— (2011): *Discretization of Processes*, Springer.

Jaffe, A. B., M. Trajtenberg, and M. S. Fogarty (2000): "Knowledge Spillovers and Patent Citations : Evidence from a Survey of Inventors," *American Economic Review*, 90, 215–218.

Johnson, D. K. N. and A. Brown (2004): "How the West Has Won: Regional and Industrial Inversion in U. S. Patent Activity," *Economic Geography*, 80, 241–260.

Johnstone, N., I. Haščič, and D. Popp (2010): "Renewable Energy Policies and Technological Innovation: Evidence Based on Patent Counts," *Environmental and Resource Economics*, 45, 133–155.

Kiefer, N. (1988): "Economic Durationd Data and Hazard Functions," *Journal of Economic Literature*, XXVI, 646–679.

Kim, S., N. Shephard, and S. Chib (1998): "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *The Review of Economic Studies*, 65, 361–393.

Kortum, S. and J. Lerner (1999): "What is behind the recent surge in patenting?" *Research Policy*, 28, 1–22.

Kortum, S. and J. Putnam (1997): "Assigning Patents to Industries: Tests of the Yale Technology Concordance," *Economic Systems Research*, 9, 161–176.

Li, X. (2012): "Behind the recent surge of Chinese patenting: An institutional view," *Research Policy*, 41, 236–249.

Li, Y., P. A. Mykland, E. Renault, L. Zhang, and X. Zheng (2014): "Realized Volatility When Sampling Times Are Possibly Endogenous," *Econometric Theory*, 30, 580–605.

Liu, L. Y., A. J. Patton, and K. Sheppard (2015): "Does Anything Beat 5-minute RV? A Comparison of Realized Measures Across Multiple Asset Classes," .

Lybbert, T. J. and N. J. Zolas (2014): "Getting patents and economic data to speak to each other: An 'Algorithmic Links with Probabilities' approach for joint analyses of patenting and economic activity," *Research Policy*, 43, 530–542.

Mancini, C. (2009): "Non-Parametric Threshold Estimation for Models with Stochastic Diffusion Coefficient and Jumps," *Scandinavian Journal of Statistics*, 36, 270–296.

Patton, A. J. (2011): "Volatility Forecast Comparison Using Imperfect Volatility Proxies," *Journal of Econometrics*, 160, 246–256.

Patton, A. J. and K. Sheppard (2009): "Evaluating Volatility and Correlation Forecasts," in *Handbook of Financial Time Series*, 801–838.

Pelletier, D. and Q. Weng (2015): "Returns, Durations, and Time Endogeneity," .

Pelletier, D. and H. Zheng (2013): "Joint Modeling of High-Frequency Price and Duration Data," .

Ragoussis, A. and E. Gonnard (2012): "THE OECD-ORBIS DATABASE TREATMENT AND BENCHMARKING PROCEDURES," .

Renault, E., T. van der Heijden, and B. J. Werker (2013): "The Dynamic Mixed Hitting-Time Model for Multiple Transaction Prices and Times," *Journal of Econometrics*, 180, 233–250.

Renault, E. and B. J. Werker (2004): "Stochastic Volatility Models with Transaction Time Risk," *CentER Discussion Paper*.

Ribeiro, S. P., S. Menghinello, and K. D. Backer (2010): "The OECD ORBIS Database: Presponding to the Need for Firm-Level Micro-Data in the OECD," .

Rubinstein, M. (1994): "Implied Binomial Trees," *The Journal of Finance*, LXIX, 771–818.

Schmoch, U., F. Laville, P. Patel, and R. Frietsch (2003): "Linking Technological Areas to Industrial Sectors," *European Economic Review*, 47, 687–710.

Squicciarini, M., H. Dernis, and C. Criscuolo (2013): "Measuring Patent Quality: Indicators of Technological and Economic Value," .

Thoma, G., S. Torrisi, A. Gambardella, D. Guellec, B. H. Hall, and D. Harhoff (2010): "Harmonizing and combining large datasetsâĂŞan application to firm-level patent and accounting data," .

Warwick, K. and A. Nolan (2014): "Evaluation of Industrial Policy," *OECD Science, Technology and Industry Policy Papers*, No. 16.

Wei, W. and D. Pelletier (2015): "A Jump-Diffusion Model with Stochastic Volatility and Durations," 1–44.

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1084.

Žežula, I. (2009): "On Multivariate Gaussian Copulas," *Journal of Statistical Planning and Inference*, 139, 3942–3946.