Intraday End of Day Volume Prediction

Alessio Sancetta*

May 22, 2019

Abstract

A model that predicts the total end of day volume and updates the prediction as new intraday information is observed is proposed. This model is a time varying coefficients model which is estimated using the framework of functional data. Semiparametric constraints such as monotonicity of the time varying coefficients can be imposed in the estimation. Results that allow us to derive confidence bands under a variety of scenarios including functions that lie at the boundary of the constrained set are given. In the empirical application we consider the end of day volume prediction of major Forex futures traded on the Chicago mercantile Exchange and show the benefits of the proposed methodology.

Key Words: functional regression, hypothesis testing, parameter at the boundary, restricted model.

JEL Codes: C58, C14

*Acknowledgements: I am grateful to Sriram Pyngas, Yuri Taranenko and Roger Wilson at UBS for an inspiring discussion. I thank the editor Andrew Patton and the referees for comments that have led to substantial improvements in content and presentation. I also benefited from comments from the participants at the High Voltage Econometrics Workshop, Palermo, Oct. 2018. E-mail: <asancetta@gmail.com>, URL: <http://sites.google.com/site/wwwsancetta/>. Address for correspondence: Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK

1 Introduction

The prediction of the end of day traded volume is an important quantity in finance. In particular, this is used by execution algorithms in high frequency trading. The traded volume allows us to estimate the characteristic size of the asset that needs to be either bought or liquidated and the related market impact (inter alia, Almgren and Chriss, 2000, Gatheral, 2010). In this case, as we observe the traded volume over the day, the variable of interest is the residual volume to the end of day. More precisely, let $V_i(t)$ be the volume traded so far at time t on day i. Let $V_i(T)$ be the end of day volume ($t < T$). The residual volume to the end of day is $Y_i(t) := V_i(T) - V_i(t)$. It would be wasteful to employ a model that predicts $V_i(T)$ without using information available until time t on day i. Usual econometric techniques fix an horizon and then estimate a model, and make predictions for that fixed horizon. Here, the horizon is the end of day, but we allow the prediction to change as t approaches the end of day T, as new observations become available, most crucially $V_i(t)$ among possibly others. To do so we recast the problem within the framework of functional data.

Intraday prediction of volumes or related quantities such as volatility has been extensively studied in the literature. One models the dynamics of the object of interest conditional on previous intraday information. The natural approach is to adapt or extend the multiplicative component GARCH (Engle and Sokalska, 2012) as done in Brownlees et al. (2011) and references therein. Here, we are not interested in the volume to be traded in the next minute or so. We wish to predict the total traded volume at the end of the day and possibly change our prediction during the day as we approach the end of the day. Both the next period prediction and the present problem are of interest with complementary goals in mind.

The contribution of the paper is to address the problem of prediction at a fixed point in time as information accumulates and the forecasting horizon shrinks. Having suggested a solution for the estimation problem, we also derive confidence bands for the estimator. Furthermore, prediction problems tend to benefit from the use of constrained estimation. Constraints can come in the form of monotonicity and/or convexity. We define two constrained estimators and show how to derive confidence bands when the constraint is binding under the null (i.e. the true infinite dimensional parameter lies on the boundary of the constraint). We also propose a test to assess the validity of the constraint. We apply our methodology to the prediction of end of day volumes of major Forex (Fx) futures traded on the Chicago Mercantile Exchange (CME). Our empirical

study shows that the residual volume to end of day depends on intraday variables with coefficients that are highly non-linear functions of the forecast horizon. We postulate a model that also depends on information from the traded volumes on the e-mini S&P500 futures and show that this variable improves the prediction out of sample. Finally, we compare the out of sample prediction to an autoregressive process (AR) that is based on daily volumes and estimated on the whole sample. This allows us to derive a lower bound on the relative value of using intraday information when predicting the end of day volume. As expected the improvement is huge. To assess the validity of the asymptotic inference carried out in the empirical study, we devise a simulation where the data generating process mimics the characteristics of the empirical data. We find that asymptotic inference can be reliable in relatively small samples.

The derivation of the asymptotic distribution of the unconstrained estimator is based on a standard martingale central limit theorem for martingale differences with values in a Hilbert space. The conditions can be verified under the sole existence of a second moment. The asymptotic distribution of the constrained estimators combines various known arguments from the finite dimensional literature applied to functional data, and it is again established under the sole condition of existence of second moments. These results on constrained estimation for functional data are new and have the practical goal of improving estimation and prediction.

Estimation of functional data is a mature subject, especially in the densely observed case (Wang et al., 2016, for a concise review and references). The literature on functional data regression tends to focus on the univariate case. However applications to multivariate financial problems have been considered by Kokoszka et al. (2015). Here, we address the multivariate case and show how to solve the problem using systems of equations. This is well suited for econometric analysis. We keep the estimation method purposely simple in order to make use of linear regression techniques in the implementation. As previously mentioned, the constrained estimation problem, when the true function is at the boundary of the constraint has not been addressed in the functional data case; results are well known for finite dimensional statistical problems (Geyer, 1994).

The comparison of constrained and unconstrained models is a well studied problem in econometrics (inter alia, Fan and Li, 1996, Zheng, 1996, Yatchew and Härdle, 2006). In the functional data framework with more than one functional regressor, many of these tests may struggle because of the use of a kernel smoother in high dimensions, which is sensitive to the choice of bandwidth. Hence, as in Yatchew (1992) we use sam-

ple splitting. Sample splitting has a long tradition in statistics and is not necessarily as inefficient as one might think (Cox, 1975, for an early reference). However, to make the problem applicable to general time series, we combine sample splitting with the predictive sequential (prequential) approach of Dawid (Dawid, 1997, Seillier-Moiseiwitsch and Dawid, 1993, and references therein). Hence, we construct a test statistic which is a martingale under the null hypothesis. The methodology is related to the usual Diebold-Mariano test (Diebold and Mariano, 1995), but is simpler to implement due to the martingale framework.

This paper also comes with a companion code that can be found in the GitHub repository URL:https://github.com/asancetta/FDRegression.

The plan for the paper is as follows. Section 2 introduces the methodology and gives the necessary details for estimation and inference. Section 3 applies the methodology to the prediction of end of day volumes of major Fx futures traded on the CME. A simulation study is presented in Section 4. It highlights the finite sample properties of the estimators presented here. Concluding remarks can be found in Section 5. The proofs of all the results can be found in the Appendix in Section A.1.

2 The Methodology

Let $\left\{\left(V_i\left(t\right)\right)_{t\in[0,T]} : i = 1, 2, ..., n\right\}$ be a sequence of cumulated traded volumes for a specific instrument over n days. In particular, $V_i\left(t\right)$ is the cumulated traded volume for day i, until time t. The end of day volume is $V_i\left(T\right)$ and $[0, T]$ is the trading interval within a day. This is the total daily volume traded by the end of day i. Define the residual volume to the end of day to be $Y_i\left(t\right) = V_i\left(T\right) - V_i\left(t\right)$. We are interested in $\mathbb{E}\left[Y_i\left(t\right)|X_i\left(t\right)\right]$, where $X_i\left(t\right)$ is some explanatory variable taking values in $\mathcal{X} \subseteq \mathbb{R}^K$. Our goal is to estimate $\mathbb{E}\left[Y_i\left(t\right)|X_i\left(t\right)\right]$ as a function of $X_i\left(t\right)$. The covariate on day i is $X_i = \left(X_i\left(t\right)\right)_{t\in[0,T]}$. We suppose that the variables $\left(Y_i\left(t\right)\right)_{t\in[0,T]}$ and $\left(X_i\left(t\right)\right)_{t\in[0,T]}$ are continuous stochastic processes. For each day i, these are a curves that evolve in time, not necessarily with stationary increments: take for example $X_i\left(t\right) = V_i\left(t\right)$, which is the traded volume until time t. Hence, we make use of the statistical theory of functional data (Bosq, 2000, Horváth and Kokoszka, 2012). In particular, for this high frequency problem, we can suppose that the data are densely observed in $[0, T]$ and any discretization is due to computational constraints. Despite non-stationarity (within the day), we can simply regress $Y_i\left(t\right)$ on $X_i\left(t\right)$ (for every t) as if we were to deal with panel

data: the index i can be seen as the time series dimension, while the index t could be viewed as a cross-sectional dimension. We suppose the following functional specification

$$Y_i(t) = X_i(t)' b_0(t) + \varepsilon_i(t) = \sum_{k=1}^{K} X_i^{(k)}(t) b_0^{(k)}(t) + \varepsilon_i(t), \tag{1}$$

where $b_0(t)$ is an unknown K-dimensional column vector valued function and the prime symbol $'$ stands for transpose. In (1), $\varepsilon_i(t)$ is the additive noise with mean zero when conditioning on $X_i(t)$. Throughout, $X_i^{(k)}(t)$ is the k^{th} entry in the K-dimensional vector $X_i(t)$, and similarly for $b_0^{(k)}(t)$.

To avoid degeneracies, we shall restrict estimation of $b_0(t)$ for t in a compact subset $\mathcal{T}$ inside $[0,T]$, for example, one minute after trading is initiated and one minute before trading ends on each day. In computer memory, functional data are usually stored as vectors. We consider N equispaced times $t_j < t_{j+1}$ such that $t_1 > 0$ and $t_N < T$, and denote this set of times by $\mathcal{T}_N := \{t_1, t_2, ..., t_N\}$. The details concerning actual implementation will be given in Section 2.3.2, and Section 3 will show the concrete empirical application.

To see the advantage in setting up the problem within a functional data framework, consider the following. Recall that $V_i = (V_i(t))_{t\in\mathcal{T}}$ is the cumulated volume trajectory on day i. We can approximate $(V_i(t))_{t\in\mathcal{T}}$ at N discrete points and consider a sample of N dimensional vectors $\left\{[V_i(t_1), V_i(t_2), ..., V_i(t_N)]' \in \mathbb{R}^N : i = 1, 2, ..., n\right\}$. It is well known that for N dimensional vector valued random variables, the central limit theorem and the law of large numbers would fail if N is not much smaller than n. This is because there are too many elements to control in the vector. On the other hand, the law of large numbers can hold for our vector even if n is much smaller than N as long as $n \to \infty$. For example, if V_i had smooth continuous trajectories, as we increased N, $V_i(t_{j-1})$ and $V_i(t_j)$ would converge to each other and the increase in N would not imply more elements to control. More generally, if the vector is obtained from data in a Hilbert space, then we can still approximate the whole trajectory arbitrarily well using a finite number of random variables (Bosq, 2000, for a formal treatment).

2.1 Notation and Background Information

The data considered are column vector valued functional data. Then, X_i takes values in a Hilbert space $\mathcal{H}^K$ with norm $|\cdot|_{\mathcal{H}^K}$ induced by the inner product $\langle X_1, X_2\rangle_{\mathcal{H}^K} = \sum_{k=1}^{K} \int_{\mathcal{T}} X_1^{(k)}(t) X_2^{(k)}(t)\, dt$ for any two $X_1, X_2 \in \mathcal{H}^K$. Hence, $\mathcal{H}^K$ is the space of vector

valued square integrable functions, though in due course we shall add continuity as well. For fixed $t \in \mathcal{T}$, we view $X_1(t)$ and $X_2(t)$ as elements in a K-dimensional Euclidean space with norm $|\cdot|_2$ induced by its canonical inner product $\langle X_1(t), X_2(t)\rangle = \sum_{k=1}^{K} X_1^{(k)}(t) X_2^{(k)}(t)$. The error term ε takes values in the Hilbert space $\mathcal{H}$ and the inner product is $\langle \varepsilon_1, \varepsilon_2\rangle_{\mathcal{H}} = \int_{\mathcal{T}} \varepsilon_1(t)\,\varepsilon_2(t)\,dt$; note that for fixed t, $\varepsilon_i(t)$ is a scalar. Except for the fact that the data are vector valued functions, the set up is the same of multivariate regression.

Let $C_{XX}(s,t) = \mathbb{E}X_i(s)\,X_i(t)'$, assuming stationarity with respect to the days $i = 1, 2, ..., n$. Even if $\mathbb{E}X_i(t)$ $t \in \mathcal{T}$ is not zero, for convenience we shall still refer to C_{XX} as the covariance function. With abuse of notation, we shall use the same symbol when referring to an operator and its kernel. The Hilbert-Schmidt norm $|\cdot|_{\mathcal{S}}$ of C_{XX} is defined by

$$|C_{XX}|_{\mathcal{S}}^2 = \int_{\mathcal{T}}\int_{\mathcal{T}} \sum_{k,l=1}^{K} \left|C_{XX}^{(k,l)}(s,t)\right|^2 dsdt,$$

where $C_{XX}^{(k,l)}(s,t)$ is the (k,l) entry in $C_{XX}(s,t)$.

2.2 Link to a Market Microstructures Intensity Model

The model in (1) can be derived from of a model of trading volume arrivals. To keep the discussion simple, suppose that trade size is constant, and with no loss of generality equal to one unit every time there is a trade. In this case, the cumulative volume $(V(t))_{t\in[0,T]}$ is a counting process taking values in the non-negative integers. For ease of notation, we drop the subscript i throughout this section, as the focus is on a fixed but arbitrary day i. Suppose that the counting process has compensator $\Lambda((0,t]) = \int_0^t \lambda(s)\,ds$, where $\lambda(t)$ is a positive predictable process bounded away from zero and infinity. Refer to λ as the intensity density. Intuitively, the higher is $\lambda(t)$ the higher is the probability of a trade during an infinitesimal interval $(t, t+dt]$ (Brémaud, 1988, for precise definitions). Given that $Y(t) = V(T) - V(t)$, we can write $Y(t) = \Lambda((t,T]) + M(t)$ where $M(t) = Y(t) - \Lambda((t,T])$. The process M is mean zero by construction when conditioning on the history up until time t.

If λ is a deterministic process, say a constant $\bar{\lambda}$ for simplicity of exposition, then, $Y(t) = \bar{\lambda} \times (T-t) + M(t)$ and this corresponds to the model (1) with $X(t) = 1$, $b_0(t) = \bar{\lambda} \times (T-t)$ and $\varepsilon(t) = M(t)$. It is easy to see how to extend this to the case of deterministic λ which is a function of time only. A deterministic intensity is equivalent

to say that the expected variation in volume during the day (often called volume profile by practitioners) is just the result of a deterministic intraday seasonal component. However, trade arrival appears to be a function of various variables that evolve over time. The seminal paper of Engle and Russell (1998) considers past durations, but also order book updates are relevant (e.g. Sancetta, 2018, and references therein). In this case, λ is a predictable stochastic process so that at time t, the conditional expectation of $\Lambda((t,T])$ is $\mathbb{E}_t\Lambda((t,T]) = X(t)' b_0(t)$, where $\mathbb{E}_t$ is the expectation conditioning on a history until time t and w.r.t. which the counting process and X are measurable at time t. Then, $\varepsilon(t) = M(t) + (1-\mathbb{E}_t)\Lambda((t,T])$. We have decomposed the error term $\varepsilon(t)$ in (1) into the process $M(t)$, plus an adjustment factor which is zero if the volume arrival intensity is deterministic. The adjustment captures the extent to which the future trajectory of the intensity is not deterministic. By definition $X(t)$ is uncorrelated with $\varepsilon(t)$. In summary, in (1) we are making the modelling assumption that the expectation of $\Lambda((t,T])$, conditional at time t, is the same as its expectation $\Lambda((t,T])$ conditioning on $X(t)$, and that this expectation is linear in the conditioning variables. This does imply that $X(t)$ and $\varepsilon(t)$ are uncorrelated.

Example Suppose that $\lambda(t) = g(Z(t))$ for some measurable function g and a stochastic process $\{Z(t) : t \in [0,T]\}$. The stochastic process does not need to be observable. However, we suppose that $\mathbb{E}_t g(Z(t+s)) = \mathbb{E}[g(Z(t+s))|X(t)] = X(t)' \dot{c}(t+s)$ where $\dot{c}(t) = dc(t)/dt$, for some differentiable vector valued map $t \mapsto c(t)$. Then, $\mathbb{E}_t \int_t^T \lambda(s)\, ds = \int_0^{T-t} X(t)' \dot{c}(t+s)\, ds$. By change of variables, the integral is $X(t)'(c(T) - c(t))$. Setting $b(t) = (c(T) - c(t))$ we recover the model in (1).

In the empirical section, we estimate an augmented version of the following model

$$Y_i(t) = b^{(1)}(t) + V_i(t)\, b^{(2)}(t) + V_{i-1}(t)\, b^{(3)}(t) + \varepsilon_i(t).$$

If the intensity measure $\Lambda((t,T])$ is deterministic, we have that $b^{(2)}(t) = b^{(3)}(t) = 0$ and $b^{(1)}(t) = \Lambda((t,T])$. If $b^{(3)}(t) = 0$ $t \in [0,1]$, the intensity is not deterministic, but satisfies

$$\mathbb{E}_t\Lambda((t,T]) = b^{(1)}(t)\left(1 + b^{(2)}(t)\right) + \left(V_i(t) - b^{(1)}(t)\right) b^{(2)}(t).$$

Supposing $b^{(2)}(t) > 0$ for $t \in (0,T)$, the conditional expectation is higher if the cumulated volume is higher than the deterministic coefficient value $b^{(1)}(t)$. If none of the

coefficients are zero, then,

$$\mathbb{E}_t\Lambda\left((t,T]\right) = b^{(1)}(t) + \left(V_i(t) - V_{i-1}(t)\right) b^{(2)}(t) + V_{i-1}(t)\left(b^{(2)}(t) + b^{(3)}(t)\right)$$

However, if $b^{(2)} = -b^{(3)}$, the above simplifies to

$$\mathbb{E}_t\Lambda\left((t,T]\right) = b^{(1)}(t) + \left(V_i(t) - V_{i-1}(t)\right) b^{(2)}(t). \tag{2}$$

In this case, the expected trajectory of the intensity measure depends on how the cumulated volume differs from the previous day cumulated volume. Under the assumption that $b^{(2)}$ is a positive function, we would expect higher volume if today's volume picks up relatively to yesterday's volume at the same time. This can be a reasonable modelling assumption.

2.3 Functional Data Least Square and Quadratic Programming Estimation

For each $t \in \mathcal{T}$ we minimize

$$\sum_{i=1}^{n}\left(Y_i(t) - X_i(t)' b(t)\right)^2$$

with respect to (w.r.t.) $b(t) \in \mathbb{R}^K$. For each fixed t, $b(t)$ is a vector. The least square estimator is

$$\hat{b}(t) = \left[\hat{C}_{XX}(t,t)\right]^{-1} \hat{C}_{XY}(t,t),\ t \in \mathcal{T} \tag{3}$$

where $\hat{C}_{XY}(t,t) = \sum_{i=1}^{n} X(t) Y(t)/n$ and similarly for $\hat{C}_{XX}(t,t)$. We need to assume conditions such that $\hat{C}_{XY}(t,t)$ is invertible in probability for the above to be meaningful. Let $\mathcal{F}_i$ be the sigma algebra generated by $\{(X_{i-s}, \varepsilon_{i-s}) : s \geq 0\}$. The following condition will be used throughout the paper.

Condition 1 *The following hold:*

1. *The true model is (1), where the true coefficient b_0 is an element in $\mathcal{H}^K$;*
2. *The sequence $(X_i)_{i\in\mathbb{Z}}$ is stationary and ergodic, takes values in $\mathcal{H}^K$, satisfies $\max_{k\leq K}\left|X_i^{(k)}(s) - X_i^{(k)}(t)\right| \leq \kappa|s-t|^{\alpha}$ for some constant κ and $\alpha > 0$;*

3. $C_{XX}(s,t) := \mathbb{E}X_i(s) X_i(t)'$ $s,t \in \mathcal{T}$ *is well defined, and is such that* $C_{XX}(t,t)$ *has minimum eigenvalue* $\lambda_{\min}(t) \geq \underline{\lambda} > 0$ *for any* $t \in \mathcal{T}$*;*

4. *The sequence* $(\varepsilon_i)_{i\in\mathbb{Z}}$ *is stationary and ergodic martingale difference sequence that takes values in* $\mathcal{H}$ *and such that* $\mathbb{E}|\varepsilon_i|_{\mathcal{H}}^2 < \infty$*;*

5. $\mathbb{E}[\varepsilon_i(t) | X_i(t)$ *and* $\mathcal{F}_{i-1}] = 0$ *for every* $i \in \mathbb{Z}$*;* $\max_{k\leq K} \int_{\mathcal{T}} \mathbb{E}\left|\varepsilon_i(t) X_i^{(k)}(t)\right|^2 dt < \infty$*;* $C_\sigma(s,t) = \mathbb{E}\varepsilon_i(s)\varepsilon_i(t) X_i(s) X_i(t)'$ *is continuous* $s,t \in \mathcal{T}$*.*

Condition 1 provides a reasonable balance between simplicity and a realistic setup for the current problem. The moment conditions are essentially minimal. Different conditions can be used at the cost of additional notation. Point 2 requires the regressors to be continuous, which is not the case if we use volumes as regressors. However, given that we sample at equally spaced times and not continuously, we can pretend that $V_i(t)$ is a continuous version of the cumulative volume when used as a regressor. The implication of Point 3 in Condition 1 is most easily seen when considering the case $K = 1$. Then, it means that $\inf_{t\in\mathcal{T}} \mathbb{E}X_i^2(t) > 0$. Following the stylized model in Section 2.2, ε_i can depend on X_i, however, the expectation of $\varepsilon_i(t)$ is equal to zero when conditioning on $X_i(t)$ and the past.

The OLS estimator satisfies the following central limit theorem.

Theorem 1 *Let* $\hat{b}$ *be as in (3). Under Condition 1,*

$$\sqrt{n}\left(\hat{b} - b_0\right) \to G_X$$

weakly in $\mathcal{H}^K$*, where* $G_X = (G_X(t))_{t\in\mathcal{T}}$ *is a mean zero Gaussian process with a.s. continuous sample paths in* $\mathbb{R}^K$ *and with matrix covariance function*

$$[C_{XX}(s,s)]^{-1} C_\sigma(s,t) [C_{XX}(t,t)]^{-1},\ s,t \in \mathcal{T} \tag{4}$$

where $C_\sigma(s,t) := \mathbb{E}\varepsilon_i(s)\varepsilon_i(t) X_i(s) X_i(t)'$*.*

Note that Point 5 in Condition 1 does not imply that $C_\sigma(s,t) = \mathbb{E}\varepsilon_i(s)\varepsilon_i(t)\mathbb{E}X_i(s) X_i(t)$. In order to construct confidence bands, we need an estimator of (4). Define the residuals

$$\hat{\varepsilon}_i(t) = Y_i(t) - X(t)'\hat{b}(t) \tag{5}$$

and the estimator $\hat{C}_\sigma(s,t) = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i(s)\hat{\varepsilon}_i(t) X_i(t) X_i(s)$ for $C_\sigma(s,t)$.

Lemma 1 *Under Condition 1, if also* $\mathbb{E}\left(|\varepsilon_i|_{\mathcal{H}}^4 + |X_i|_{\mathcal{H}^K}^4\right) < \infty$, $\hat{C}_\sigma(s,t) \to C_\sigma(s,t)$ *and* $\hat{C}_{XX}(s,t) \to C_{XX}(s,t)$ *in probability under the Hilbert-Schmidt norm* $|\cdot|_{\mathcal{S}}$.

2.3.1 Constrained Estimation

Given the nature of the problem, it is natural to impose constraints on the estimated functional coefficient $\hat{b}$. In practice the estimation suffers from noise in a finite sample. We may have prior believes regarding the shape of some of the entries in b_0. This constraints reduce noise and can improve prediction. For example, $b_0^{(1)}(t)$ is expected to be monotonically decreasing as the residual volume to end of day is monotonically decreasing. The reader can refer to the empirical study (Section 3) where plots of the estimated functional coefficients are reported.

Motivated by these remarks, we can impose restrictions on the estimator. Such restrictions can be in the form of shape constraints and/or smoothness constraints. Consider the problem

$$\inf_{b \in \mathcal{R}} \left\{ -\int_{\mathcal{T}} b(t)' \hat{b}(t)\, dt + \frac{1}{2} \int_{\mathcal{T}} b(t)' b(t)\, dt \right\}. \tag{6}$$

Here $\mathcal{R}$ is a closed convex subset of $\mathcal{H}^K$. For example, in Section 3.2 we consider estimation under monotonicity constraints. The set up introduced here allows us to address such constrained estimation. In practice, $\hat{b}$ is estimated at a finite number of points $t \in \mathcal{T}_N$. Hence, the above is a quadratic programming problem. (Computational details will be discussed in Section 2.3.2.) We cannot directly use Theorem 1, to find the asymptotic distribution of the estimator unless we suppose that $b_0 \in \text{int}(\mathcal{R})$, where $\text{int}(\mathcal{R})$ is the interior of $\mathcal{R}$. For example, if a function is monotonically decreasing, the set $\mathcal{R}$ would include all functions that are non-increasing. The interior of $\mathcal{R}$ comprises strictly decreasing functions. However, $\mathcal{R}$ also contains functions that are constant over $\mathcal{T}$ or some of its subsets. In this case, b_0 would lie on the boundary of $\mathcal{R}$ and the usual central limit theorem does not hold (Geyer, 1994). Hence, under the sole condition that $b_0 \in \mathcal{R}$, we need to introduce additional concepts used for constrained estimation. Let $\mathcal{C}_0$ be the tangent cone of $\mathcal{R}$ at b_0. The tangent cone is defined as the set of all $\delta = \lim_{n\to\infty} \epsilon_n^{-1}(\beta_n - b_0)$ for a sequence $(\beta_n) \in \mathcal{R}$ such that $\beta_n \to b_0$ (in $|\cdot|_{\mathcal{H}^K}$ norm) and a real positive sequence $\epsilon_n \to 0$. In compact notation $\mathcal{C}_0 = \limsup_{\epsilon\to 0} \epsilon^{-1}(\mathcal{R} - b_0)$. Essentially we set $\delta = \epsilon_n^{-1}(b - b_0)$ for all $b \in \mathcal{R}$. If $b_0 \in \text{int}(\mathcal{R})$ then $\mathcal{C}_0 = \mathcal{H}$ because we are free to choose b in a small enough ball centered at b_0. Hence, we can just use

OLS and Theorem 1 applies. The role of ϵ_n^{-1} is to blow up the ball centered at b_0 so that it becomes the whole of $\mathcal{H}$. This is indeed what happens when $\epsilon_n = n^{-1/2}$ in $\delta = \epsilon_n^{-1}(b - b_0)$. Asymptotically, this means that constrained and unconstrained estimation have the same distribution. If b_0 is at the boundary of $\mathcal{R}$, then we can only choose b in a direction that points inside $\mathcal{R}$ (feasible direction) when we carry out the estimation. In this case the asymptotic distribution is different from the standard Gaussian limit because certain (random) directions of $\sqrt{n}(b - b_0)$ are not allowed by the constraint; the term random is used because b should minimize a random function. We have the following.

Theorem 2 *Let $\mathcal{R}$ be a closed convex subset of $\mathcal{H}^K$. Suppose that $b_0 \in \mathcal{R}$ and let $\tilde{b}$ be the solution of (6). Under Condition 1,*

$$\sqrt{n}\left(\tilde{b} - b_0\right) \to \arg\inf_{\delta \in \mathcal{C}_0}\left\{-\int_{\mathcal{T}} \delta(t)' G_X(t) + \frac{1}{2}\int_{\mathcal{T}} \delta(t)' \delta(t)\, dt\right\}$$

where G_X is the Gaussian process defined in Theorem 1.

Theorem 2 can be used to derive confidence bands for the constrained estimator using simulations. Using continuity of the Gaussian process, we simulate paths at a finite number of points $\mathcal{T}_N$ in $\mathcal{T}$. This means simulating vectors. The details for the vector implementation are given in Section 2.3.2. For expository reason, it is best to consider the continuous time setting in this section. We simulate from G_X, then we (approximately) solve the constrained problem

$$\inf_{\delta \in \mathcal{C}_0}\left\{-\int_{\mathcal{T}} \delta(t)' G_X(t) + \frac{1}{2}\int_{\mathcal{T}} \delta(t)' \delta(t)\, dt\right\}.$$

If b_0 is not fully specified under the null we have to use a suitable sample estimator in the construction of $\mathcal{C}_0$. The solution of the above problem is $\tilde{\delta}$. We repeat many times, as many as the number of simulated paths of G_X, and derive confidence bands from the empirical quantiles of $\tilde{\delta}$ which we have approximated at $t \in \mathcal{T}_N$ only.

We also consider the alternative constrained minimization problem

$$\inf_{b \in \mathcal{R}}\left\{-\int_{\mathcal{T}} b(t)' \hat{C}_{XY}(t,t)\, dt + \frac{1}{2}\int_{\mathcal{T}} b(t)' \hat{C}_{XX}(t,t)\, b(t)\, dt\right\}. \tag{7}$$

Once we obtain the estimator $\hat{b}$ it is more practical to solve (6) rather than (7). Nevertheless, (7) provides a single framework for constrained and unconstrained estimation

and usually works better in practice.

Theorem 3 *Suppose that $b_0 \in \mathcal{R}$ and let $\tilde{b}$ be the solution of (7). Under Condition 1,*

$$\sqrt{n}\left(\tilde{b} - b_0\right) \to \arg\inf_{\delta \in \mathcal{C}_0} \left\{ -\int_{\mathcal{T}} \delta(t)' G(t)\, dt + \frac{1}{2}\int_{\mathcal{T}} \delta(t)' C_{XX}(t,t)\, \delta(t)\, dt \right\}$$

in distribution, where $(G(t))_{t\in\mathcal{T}}$ is a mean zero vector valued Gaussian process with matrix covariance function $\mathbb{E}G(s)\,G(t)' = C_\sigma(s,t)$ $s,t \in \mathcal{T}$.

In the problems we consider, the constraint imposes linear restrictions such as monotonicity and convexity. It is instructive to show that the tangent cone is easy to derive in this cases. For simplicity, let $K = 1$ and suppose that $b \in \mathcal{R} \subset \mathcal{H}^K$ if and only if $R(b) \leq r$ for some bounded linear functional $R : \mathcal{H} \to \mathbb{R}$ and some constant r. Then, by linearity,

$$r \geq R(b) = R(b_0) + R(b - b_0).$$

If b_0 is at the boundary of $\mathcal{R}$, the constraint is binding and $R(b_0) = r$, implying $R(b - b_0) \leq 0$. Hence, when the constraint is linear and the true b_0 is on the boundary, $\delta \in \mathcal{C}_0$ if and only if $R(\delta) \leq 0$. We may impose a constraint at each value $t \in \mathcal{T}$ of $b(t)$, or some subset of $\mathcal{T}$. To avoid additional notation, it is easier to discuss this in the practical implementation of Section 2.3.2, where we discretize the functions and store their values into vectors as in usual econometric analysis.

2.3.2 System of Linear Equations and Practical Implementation

In what follows we rely on the machinery of seemingly unrelated regression (SUR) and apply it to the context of vector valued functional data and constrained estimation using quadratic programming. The only difference is that in SUR the number of equations is smaller than the number of observations. When dealing with functional data, the reverse is usually true.

In reverse order from Section 2.3, we shall discuss the implementation starting from the quadratic programming problem (7) as this requires more general notation which is best to introduce from the start. We then consider the simpler problem (6). The small sample performance of the latter can be poor. However, it is simpler than (7) to implement in large samples.

Recall that the estimator is computed on a fixed time grid $\mathcal{T}_N := \{t_1, t_2, ..., t_N\}$. In order to write the quadratic programming problem, let $\boldsymbol{Y}_v$ be the $n \times 1$ column vector with i^{th} entry $Y_i(t_v)$. Let $\boldsymbol{Y}$ be the $nN \times 1$ vector obtained from stacking the vectors $\boldsymbol{Y}_v$ one below the other: $\boldsymbol{Y} = [\boldsymbol{Y}'_1, \boldsymbol{Y}'_2, ..., \boldsymbol{Y}'_N]'$. Similarly define $\boldsymbol{X}_v$ to be the $n \times K$ matrix with i^{th} row $X_i(t_v)'$. Then, define

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ 0 & \boldsymbol{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{X}_N \end{bmatrix}.$$

This is the standard notational set up for a system of regression equations. We consider each intraday time t_v to represent a separate equation. We solve the following quadratic programming problem

$$\min_{\boldsymbol{b} \in \mathbb{R}^{KN}} \left\{ -\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y} + \frac{1}{2}\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} \right\} \tag{8}$$

such that $\boldsymbol{R}\boldsymbol{b} \leq \boldsymbol{r}$ where $\boldsymbol{R}$ is a $q \times NK$ matrix, $\boldsymbol{r}$ is a $q \times 1$ vector and the inequality is meant elementwise. This is an elementary problem. For example, if we want to impose a monotone decreasing constraint on $b^{(1)}(t)$ and we have $K = 2$ variables, the first row of $\boldsymbol{R}$ is $(-1, 0, 1, 0, 0, ..., 0)$. The i^{th} row of $\boldsymbol{R}$ is just the first row of $\boldsymbol{R}$ with entries shifted $2(i-1)$ places to the right, e.g. for $i = 2$, $(0, 0, -1, 0, 1, 0, 0, ..., 0)$. In this case we have $q = N - 1$ restrictions. Let $\boldsymbol{P}_k$ be the $N \times NK$ matrix that picks up the coefficients corresponding to the the k^{th} variable. For example $\boldsymbol{P}_k\tilde{\boldsymbol{b}}$ has entries equal to $\left(\tilde{b}^{(k)}(t)\right)_{t \in \mathcal{T}_N}$ where the tilde is used if the estimator is a constrained one. Hence, our constrained estimator of $b_0(t_v)' = \left(b_0^{(1)}(t_v), b_0^{(2)}(t_v), ..., b_0^{(K)}(t_v)\right)$ is just the v^{th} row of $\left(\boldsymbol{P}_1\tilde{\boldsymbol{b}}, \boldsymbol{P}_2\tilde{\boldsymbol{b}}, ..., \boldsymbol{P}_K\tilde{\boldsymbol{b}}\right)$. We denote by $\hat{\boldsymbol{b}}$ the estimator from (8) under no constraint.

If $\boldsymbol{b}_0$ is the true parameter (using vector notation) and is such that $\boldsymbol{R}\boldsymbol{b}_0 < \boldsymbol{r}$ (elementwise), then the tangent cone is the whole of $\mathbb{R}^{KN}$. On the other hand if $\boldsymbol{R}_{\mathcal{I}}\boldsymbol{b}_0 = \boldsymbol{r}_{\mathcal{I}}$ where $\mathcal{I} \subseteq \{1, 2, ..., q\}$, the constraint is binding for the indexes in $\mathcal{I}$. Here, $\boldsymbol{R}_{\mathcal{I}}$ denotes the subset of the rows in $\boldsymbol{R}$ with row index in $\mathcal{I}$. Similarly we define $\boldsymbol{r}_{\mathcal{I}}$. In this case, the tangent cone of $\left\{\boldsymbol{b} \in \mathbb{R}^{NK} : \boldsymbol{R}\boldsymbol{b} \leq \boldsymbol{r}\right\}$ at $\boldsymbol{b}_0$ is the set of $\boldsymbol{\delta} \in \mathbb{R}^{NK}$ such that $\boldsymbol{R}_{\mathcal{I}}\boldsymbol{\delta} \leq \mathbf{0}_{|\mathcal{I}|}$ where $\mathbf{0}_{|\mathcal{I}|}$ is the column vector of zeros with dimension given by the dimension of $\mathcal{I}$.

The matrix covariance $C_{XX}(s,t)$ is estimated by $\hat{\boldsymbol{C}}$, which is a matrix with $(k + (v-1)K, l + (u-1)K)$ entry equal to $\hat{C}_{XX}^{(k,l)}(t_v, t_u)$, which is the (k,l) entry in $\boldsymbol{X}'_v\boldsymbol{X}_u/n$. The covariance $C_\sigma(s,t)$ is estimated by $\hat{\mathbf{C}}_\sigma$, which is a matrix with

$(k+(v-1)K, l+(u-1)K)$ entry equal to $\hat{C}_\sigma^{(k,l)}(t_v,t_u) = n^{-1}\sum_{i=1}^n \tilde{\varepsilon}_i(t_v)\tilde{\varepsilon}_i(t_u)X_i^{(k)}(t_v)X_i^{(l)}(t_u)'$; here $\left\{(\tilde{\varepsilon}_i(t))_{t\in\mathcal{T}_N} : i=1,2,...,n\right\}$ are the residuals from the constrained estimation. With this notation we compute

$$\hat{\boldsymbol{C}}_b := \text{blockdiag}\left(\hat{\boldsymbol{C}}\right)^{-1}\hat{\boldsymbol{C}}_\sigma\text{blockdiag}\left(\hat{\boldsymbol{C}}\right)^{-1}$$

which is an estimator of (4). Here, $\text{blockdiag}\left(\hat{\boldsymbol{C}}\right)$ is the block diagonal matrix with block diagonal elements equal to the ones of $\hat{\boldsymbol{C}}$ along $(k+(v-1)K, k+(v-1)K)$ for $k=1,2,...,K$, $v=1,2,...,N$ and zero otherwise. In particular, the $(k+(v-1)K, l+(u-1)K)$ entry in $\hat{\boldsymbol{C}}_b$ is the (k,l) entry in (4) evaluated at (t_v,t_u). The variance of each function is $\text{diag}\left(\text{blockdiag}\left(\hat{\boldsymbol{C}}\right)^{-1}\hat{\boldsymbol{C}}_b\text{blockdiag}\left(\hat{\boldsymbol{C}}\right)^{-1}\right)$ where $\text{diag}(\cdot)$ stands for the diagonal matrix of its argument. Note that neither $\hat{\mathbf{C}}_\sigma$ or $\hat{\boldsymbol{C}}$ are invertible when the set of points in $\mathcal{T}_N$ is very dense (i.e. $N\to\infty$). Hence, it is not possible to construct a more efficient estimator using generalized least square. In order to construct confidence bands using Theorem 2, we need to simulate a zero mean Gaussian random vector with covariance matrix $\hat{\boldsymbol{C}}_\sigma$.

Large Datasets. In (6), the constrained estimator is directly derived from the unconstrained one. Using a system of equations, this is the solution to

$$\min_{\boldsymbol{b}\in\mathbb{R}^{KN}}\left\{-\boldsymbol{b}'\hat{\boldsymbol{b}}+\frac{1}{2}\boldsymbol{b}'\boldsymbol{b}\right\} \tag{9}$$

such that $\boldsymbol{R}\boldsymbol{b}\leq\boldsymbol{r}$. Recall the $\hat{\boldsymbol{b}}$ is the estimator from (8) under no constraint.

For large datasets, estimation based on (8) might pose some challenges because of the large memory needed to construct the matrix $\boldsymbol{X}'\boldsymbol{X}$. Estimation based on $\hat{\boldsymbol{b}}$ and then (9) does not pose such a problem. The OLS estimator $\hat{\boldsymbol{b}}$ can be computed from (3) for each $t\in\mathcal{T}_N$. This requires N OLS estimations. Hence, $\boldsymbol{P}_v\hat{\boldsymbol{b}} = (\boldsymbol{X}_v'\boldsymbol{X}_v)^{-1}\boldsymbol{X}_v'\boldsymbol{Y}_v$ where the matrix $\boldsymbol{P}_v$ was implicitly defined above. In this case it might be more intuitive to redefine $\hat{\boldsymbol{b}}$ as $\left[\left(\boldsymbol{P}_1\hat{\boldsymbol{b}}\right)',\left(\boldsymbol{P}_2\hat{\boldsymbol{b}}\right)',...,\left(\boldsymbol{P}_N\hat{\boldsymbol{b}}\right)'\right]'$ and then solve (9) which is independent of the sample size. Intuition together with data analysis carried out by the author suggest that when $\hat{\boldsymbol{b}}$ is very close to the true coefficient (e.g. little noise), and the constraint is true, then (9) provides good results, which are comparable to (8). However, when the noise level is high and the constraint might not be true (or only approximately true), (9) can give rather poor results relatively to (8). Whenever

possible, it is safer to use (8) unless $\hat{\boldsymbol{b}}$ itself appears to be already rather smooth and satisfying the constraint. This is relatively easy to establish by visual inspection of $\hat{\boldsymbol{b}}$.

It is also clear that the extent to which problem (9) leads to suboptimal results is related to how close $C_{XX}(t,t)$ is to being diagonal for every t.

2.4 Testing Restrictions

Confidence bands do not allow us to formally test for the validity of restrictions unless we completely specify a possible value for b_0. As mentioned in the Introduction, test statistics for testing functional forms have been proposed in the case of parametric versus non-parametric alternatives in the case of real valued data. Many of these methods use some kernel smoothing in order to approximate a conditional expectation. When the number of covariates K is greater than one, the procedure may require a rather careful choice of smoothing parameter. Moreover, in the functional data context we have the additional computational challenge of dealing with the index parameter $t \in \mathcal{T}$. In these circumstances, sample splitting can be an option that is simple to implement. The estimators are computed from an estimation sample $\{(Y_i, X_i) : (-n+1) \leq i \leq 0\}$ of size n, say $\hat{b}_{est}$ and $\tilde{b}_{est}$, where the subscript stresses this fact. The relative performance is assessed on a subsequent testing sample $\{(Y_i, X_i) : 1 \leq i \leq m\}$ so that our full sample has size $n+m$. Define

$$D_i := \int_{\mathcal{T}} \left(Y_i(t) - X_i(t)' \tilde{b}_{est}(t)\right)^2 dt - \int_{\mathcal{T}} \left(Y_i(t) - X_i(t)' \hat{b}_{est}(t)\right)^2 dt$$

and the test statistic

$$S_m := S_m\left(\tilde{b}_{est}, \hat{b}_{est}\right) := \frac{\sum_{i=1}^m D_i}{\sqrt{\sum_{i=1}^m D_i^2}}. \tag{10}$$

Define the event $B := \left\{\left|\tilde{b}_{est} - \hat{b}_{est}\right|_{\mathcal{H}^K} \in [\epsilon, \epsilon^{-1}]\right\}$ for some arbitrary but fixed $\epsilon > 0$. Note that $(D_i)_{i \geq 0}$ can be degenerate under the null if $\left|\tilde{b}_{est} - \hat{b}_{est}\right|_{\mathcal{H}^K} \to 0$. This is why we introduce the event B. If $\left|\tilde{b}_{est} - \hat{b}_{est}\right|_{\mathcal{H}^K} \to 0$, the two estimators will perform similarly out of sample. The null hypothesis is

$$H_0 : \lim_m \sum_{i=1}^m \frac{\mathbb{E}_{i-1}[D_i|B]}{\sqrt{m}} \leq 0,$$

in probability. The null is expressed in terms of conditional expectations allowing for some finite sample approximation error. The null essentially says that asymptotically the constrained estimator does not lead to higher forecasts errors. The alternative is

$$H_1 : \lim_m \sum_{i=1}^{m} \frac{\mathbb{E}_{i-1}\left[D_i|B\right]}{\sqrt{m}} = \infty$$

in probability, to ensure that the power of the test goes to one. The critical values are obtained from an application of the following martingale central limit theorem.

Theorem 4 *Suppose Condition 1,* $\sup_{i\geq 1} \mathbb{E}\left[|X_i|^{4\eta}_{\mathcal{H}^K} + |X_i\varepsilon_i|^{2\eta}_{\mathcal{H}^K}|B\right] < \infty$ *for some* $\eta > 1$*, and*

$$\frac{1}{m}\sum_{i=1}^{m} D_i^2 \to c \in (0,\infty) \tag{11}$$

in probability on the event B*. If* $\sum_{i=1}^{m} \frac{\mathbb{E}_{i-1}[D_i|B]}{\sqrt{m}} = o_p(1)$*, then, on the event* B*,* S_m *converges in distribution to a standard normal random variable.*

3 Predicting the End of Day Volume of CME Fx Futures

We consider the prediction of the end of day volume for the front month of major Fx futures contract traded on the CME. This is an important problem. Spot Fx is traded over the counter and no volumes on the primary electronic communication networks are reported. Only the last traded price is reported, on sliced time stamps, i.e. there are no streamed trade prices. Hence, CME traded volumes can be used an indication of volumes. The most liquid major Fx products on the CME are the futures on EURUSD, JPYUSD, USDCHF, GBPUSD, AUDUSD, NZDUSD, USDCAD. The respective CME tickers are 6E, 6J, 6S, 6B, 6A, 6N, 6C. To possibly improve our predictions, we also use the volumes from the S&P500 e-mini contract (ticker ES) as explanatory variable.

3.1 Data Description

The sample is for the period 01/Apr/2013-30/Sep/2013. For this period we have 127 days of available data. The data comprises all the messages sent by the Chicago CME. From these we compute the cumulative trading volumes during each day at a one

Table 1: Summary Statistics. One-minute volumes have been computed for every day and appended (vectorized) for each day. The sample size of the vectorized data is 68040. Summary statistics have then been computed: mean, standard deviation (std), skewness (skew), kurtosis (kurt), and the sample autocorrelations at lags 5 and 100 (acf(5), acf(100)).

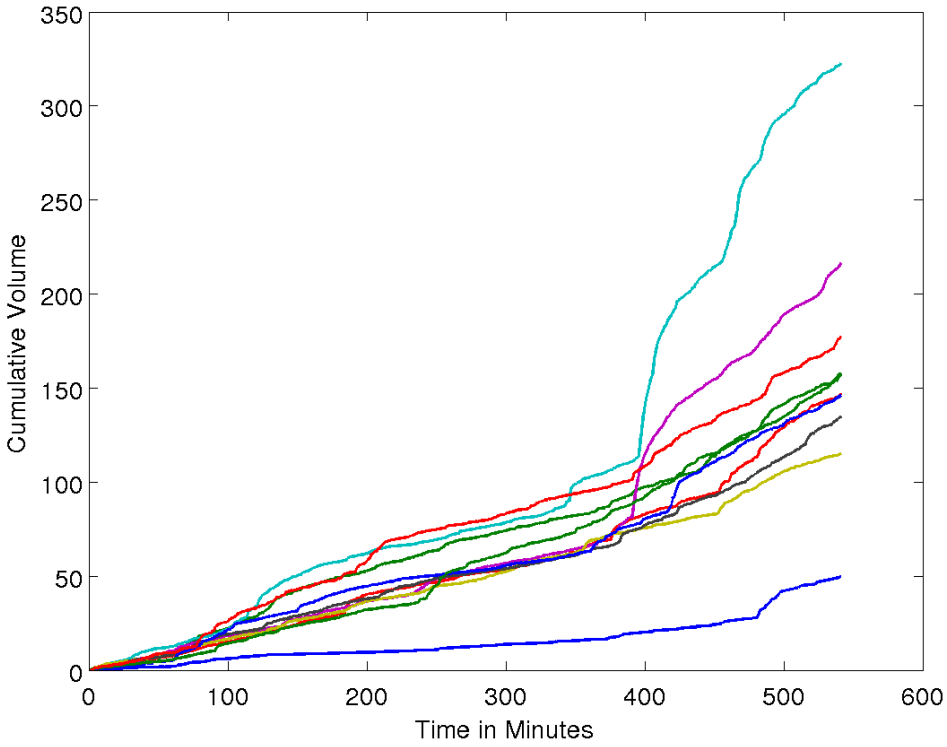
	mean	std	skew	kurt	acf(5)	acf(100)
6E	0.27	0.41	7.79	126.55	0.37	0.05
6J	0.16	0.26	9.00	180.65	0.38	0.09
6S	0.04	0.07	10.35	265.75	0.30	0.06
6B	0.13	0.21	12.18	329.21	0.26	0.02
6A	0.11	0.15	5.89	75.27	0.34	0.06
6N	0.02	0.03	6.21	71.85	0.19	0.02
6C	0.07	0.11	11.13	291.72	0.29	0.06
ES	1.29	2.59	4.18	29.94	0.65	-0.05

minute frequency. We focus on the hours of the day when there is most activity in London. Hence, when we mention start and end of day, we mean 7:00am to 4:00pm London time. The latter corresponds to the London spot Fx fixing. A model that uses information outside these hours would be more complex. Trading on the CME during different hours of the day tends to be characterized by players with specific regional characteristics. We leave such extension to future research.

Table 1 reports basic summary statistics for volumes at one minute frequency. Recall that our interest is nevertheless in the end of day volume, which is the total sum of one minute volumes. The summary statistics show that the volumes cannot be well approximated by a compound Poisson process. This could be due to unaccounted intraday seasonal patters that induce strong sample autocorrelation. Hence, we also computed the seasonal component of one-minute volumes and divided the original one-minute volumes by it. The resulting acf did not change the overall picture of persistent volumes even at lags as high as 100.

Figure 1 plots the cumulative volumes for 6E, i.e. $\left(V_i\left(t\right)\right)_{t\in\mathcal{T}_N}$. From the picture it becomes obvious why the process cannot be modelled by a compound Poisson process. Hence, it is natural to consider each day as one observation, and recast the problem within a functional data framework.

Figure 1: Cumulative Volume 6E. The cumulative volume for 6E is plotted for a small sample of days.



3.2 Model Description and Hypotheses

The model is

$$\begin{aligned} Y_i(t) &= b^{(1)}(t) + V_i(t)\, b^{(2)}(t) + V_{i-1}(t)\, b^{(3)}(t) + \\ &\quad + V_{i-1}(T)\, b^{(4)}(t) + V_i^{ES}(t)\, b^{(5)}(t) + \varepsilon_i(t). \end{aligned} \tag{12}$$

The variable $V_i(t)$ is the volume traded until time t on day i for a given futures contract. The variable $V_i^{ES}(t)$ denotes the cumulative volume up to time t of ES. Recall that $Y_i(t) = V_i(T) - V_i(t)$. This model requires the estimation of 5 functions. Volumes are expressed in thousands in this empirical section. The set $\mathcal{T}$ is mapped into $[0,1]$ by linear transformation with end points representing 7:01am and 3:59pm, respectively. We consider one minute sampling so that $\mathcal{T}_N$ is a set with 539 elements. Estimation is carried out by OLS in the case of unconstrained estimation and using (8) for the constrained estimation. We also computed the constrained estimator from (9), but the results were poor as expected for such sample size.

We consider constraints implied by the following hypotheses.

Hypothesis 1. $b^{(5)}(t) = 0$, i.e. the volume of ES does not enter the equation.

Hypothesis 2. Hypothesis 1 holds and $b^{(1)}(t)$ is monotonically decreasing. This is reasonable, as the residual volume to the end of day is monotonically decreasing by construction.

Hypothesis 3. Hypotheses 1 and 2 hold. Moreover, $b^{(1)}(t)$ is also convex for $t \leq 1/2$ and concave for $t > 1/2$. This corresponds to the case where relatively more volume is transacted at the beginning of trade in London and when we get closer to the 4:00pm London spot Fx fixing. Also, $b^{(2)}(t)$, $b^{(3)}(t)$ are convex monotonically decreasing functions, while $b^{(4)}(t)$ is decreasing. The monotone decreasing constraints on $b^{(2)}(t)$, $b^{(3)}(t)$ mean that the volumes executed so far will progressively have lower importance for the prediction of remaining volumes even when we consider previous day volumes. However, the decrease is at a decreasing rate. The decreasing constraint on $b^{(4)}(t)$ should follow for the reasons just discussed.

Hypothesis 4. Hypotheses 1 and 2 hold, and $b^{(2)}(t)$, $b^{(3)}(t)$ and $b^{(4)}(t)$ are constant. In this case, the only time varying coefficient is $b^{(1)}$.

Hypothesis 5. Same as Hypothesis 3, but with $b^{(5)}$ constrained to be monotonically decreasing rather than zero.

3.2.1 Model Evaluation and Results

We split the sample in estimation and test sample. We estimate (12) on the estimation sample under the various hypotheses including a fully unconstrained one. We then compare the estimators on the test sample. In particular we use the first 70 trading days for estimation and the remaining 57 days for testing. For each model we construct the out of sample prediction using the $b(t)$ estimated on the estimation sample. We have an unconstrained model and 5 hypotheses.

Figure 2 shows the functions for 6E using the model from Hypothesis 1 (which is unconstrained but omits the ES variable). The figures also include the model from Hypothesis 3, and it plots confidence bands around the model from Hypothesis 3, using the results from Theorem 1. The estimates for Hypothesis 3 are much smoother. However, we also note that imposing restrictions might not always lead to improvements. The constraint on $b^{(3)}$ implied a monotonically decreasing impact of $V_{i-1}(t)$. The plot of $b^{(3)}$ shows that the opposite could be true for 6E and this comment applies to all other contracts except for 6N. In particular, to some degree we can see that the trajectory of $b^{(3)}$ resembles the one of $-b^{(2)}$. In this case, if we actually had $b^{(2)} = -b^{(3)}$, then we would recover the model in (2). This would imply that a higher cumulative volume today, relative to yesterday, should lead to higher residual volume to end of day. Finally, the constraint on $b^{(4)}$ might hold at the boundary of the parameter and be constant. This comment also applies to all other contracts.

For illustration purposes, the plot for $b^{(4)}$ also includes confidence bands when we suppose that the true function is constant. In this case, the function lies at the boundary of the space of monotonically decreasing functions. Hence, the confidence bands are derived using Theorem 3 as opposed to Theorem 1. This latter bands are centered around the average of the estimated parameter $\tilde{b}^{(4)}$, i.e. $b_0^{(4)} = \int_{\mathcal{T}} \tilde{b}^{(4)}(t)\,dt$, from Hypothesis 3. These confidence bands are the dash-dot lines that tend to be slightly narrower than the usual confidence bands and they are not necessarily symmetric around the true value. It is not unreasonable to think that the function could be constant.

We compare the constrained estimators to the unconstrained. To this end, we

Figure 2: Estimated Functions. The estimated functions for $b^{(k)}$, $k = 1, 2, 3, 4$ from the model in (12) are plotted in the case of the estimators from Hypothesis 1 and 3. The estimators from Hypothesis 3 are smoother but close to the ones from Hypothesis 1. Dashed lines represent the 95% confidence bands round the estimator from Hypothesis 3, computed using Theorem 1. For $b^{(4)}$ for the purpose of illustration, confidence bands under the assumption that the true $b_0^{(4)}$ is constant (i.e. on the boundary of the constraint) are also plotted as dash-dotted lines. These bands are computed using Theorem 3 and plotted around the average of the estimator from Hypothesis 3.

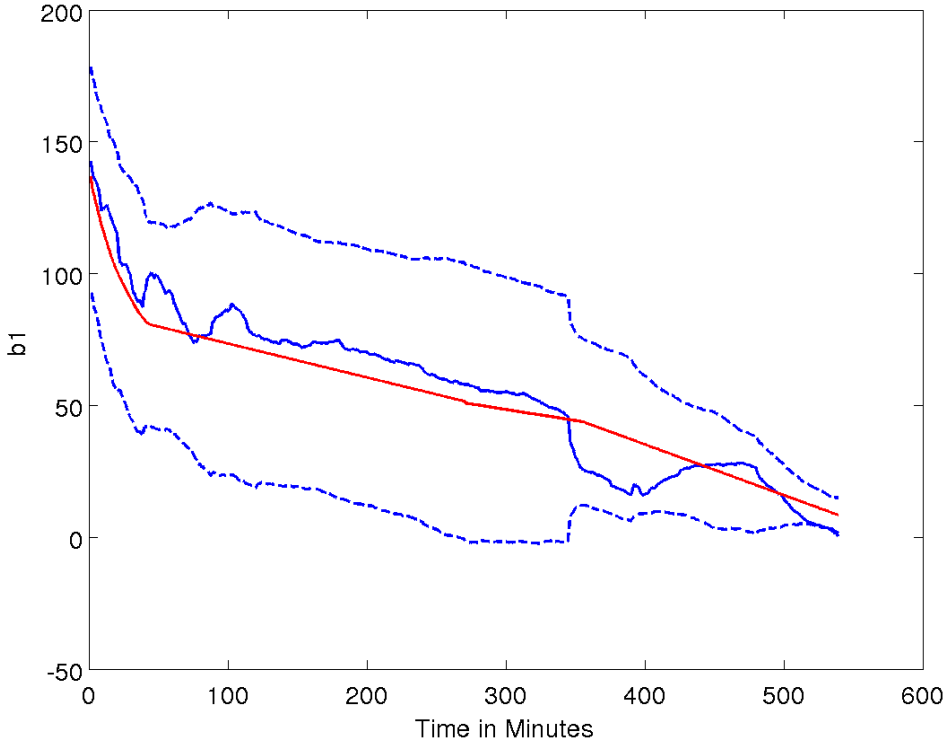
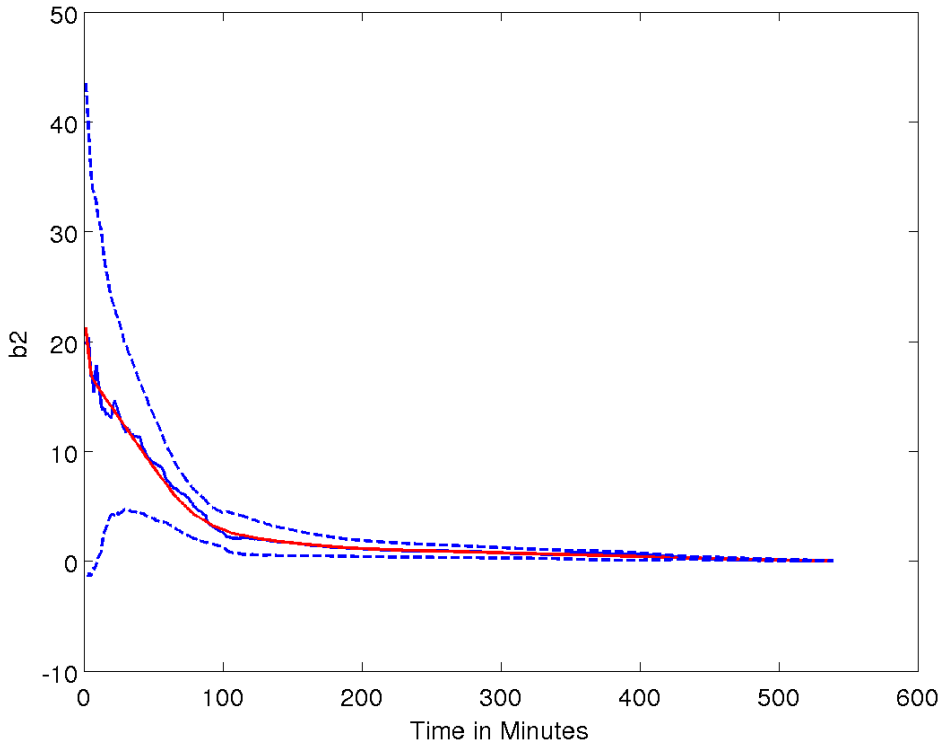


Figure 2: Estimated Functions. Continued



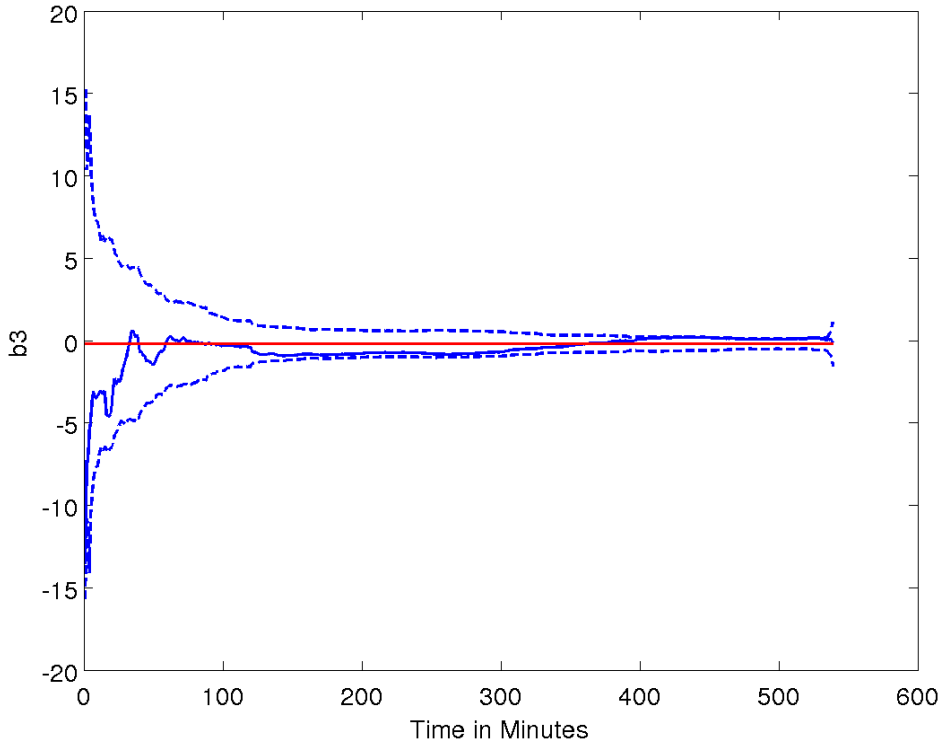
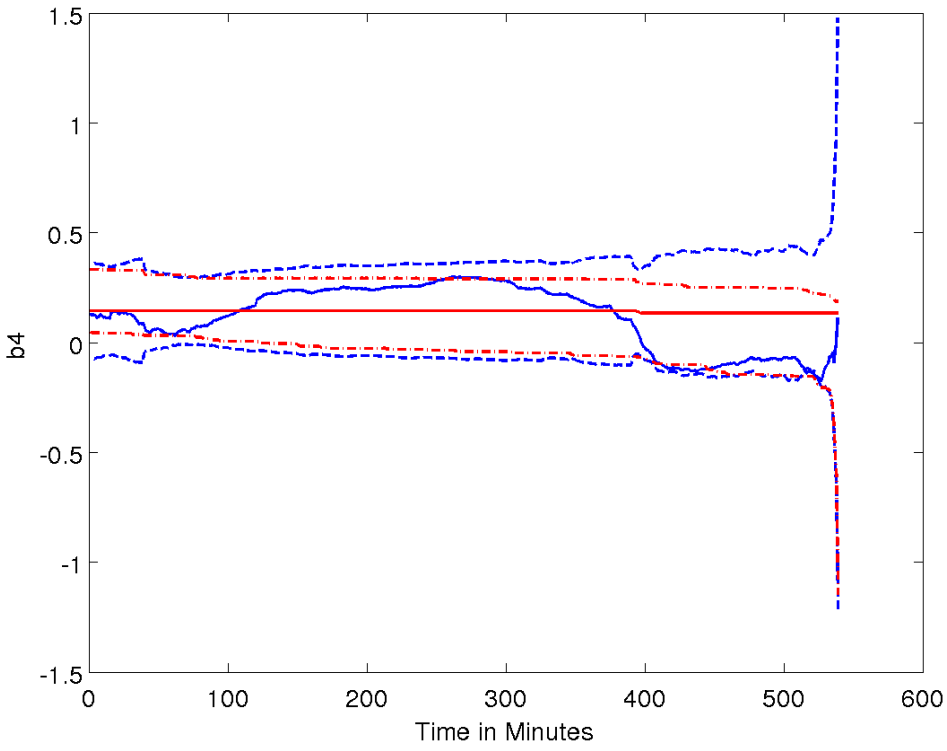


Figure 2: Estimated Functions. Continued



compute the mean square error $MSE(b) = \frac{1}{m}\sum_{i=1}^{m} SE_i(b)$, where

$$SE_i(b) = \frac{1}{N}\sum_{j=1}^{N}\left(Y_i(t_j) - X_i(t_j)' b(t_j)\right)^2$$

is the intraday square error. The performance of a constrained estimator $\tilde{b}$ relative to a benchmark parameter $\hat{b}$, is evaluated in terms of the percentage of relative improvement

$$PRI\left(\tilde{b}, \hat{b}\right) := 100\frac{MSE\left(\hat{b}\right) - MSE\left(\tilde{b}\right)}{MSE\left(\hat{b}\right)} \tag{13}$$

which is a number in $[0, 100]$, with a larger number signifying a larger improvement. Here, $\tilde{b}$ is one of the constrained models, while $\hat{b}$ is the unconstrained estimator. Similarly, we conduct the model performance evaluation test using Theorem 4 and the statistic $S_m = S_m\left(\tilde{b}, \hat{b}\right)$ where the arguments make explicit which models we compare. Note that Theorem 4 is valid even if we compare two different constrained models. Table 2 reports the results. There is evidence that the constrained models perform relatively well, but we have to reject the null that the coefficients, except for the intercept, are constant. Overall, H3 and H5 provide the best performance. The model H5 is as H3 but also includes the volume of ES as a regressor. To further test the restriction imposed by H3, we also report the results for H3 versus H5. In this case, looking at the PRI, in agreement with anecdotal evidence from spot Fx traders, we infer a benefit in using model H5 versus H3 for 6A and 6N. However, when we look at the test statistic S_m, we cannot reject the null that H3 performs as well as H5.

Finally, to assess the validity of the martingale assumption in Condition 1, we used the estimated b to construct residuals over the whole sample for each of the hypotheses. These residuals can then be used to conduct a Box-Pierce statistic for functional data (Horváth and Kokoszka, 2012, Ch.7, see also Sancetta, 2015, Th.2). Using five lags, nothing was found to be significant at the 5% level, except for three functional Box Pierce test statistics out of 42. There are 42 test statistics because for each of the 7 instruments, 6 Box-Pierce tests for functional data are carried out, one for each of the 5 hypotheses and one for the the unconstrained model.

Table 2: Hypotheses Comparison. The PRI in (13) and the test statistic in (10) are computed for Hypotheses 1-5, where H1 stands for Hypothesis 1 and so on. The improvement for PRI and the test statistic S_m are relative to the unconstrained estimator of the model (12) and also relative to H5 in the case of H3. Large positive values of PRI are an improvement in percentage points. Large positive values of S_m mean that the restriction imposed by the hypothesis is rejected. Conversely, a large negative value favours the hypothesis/constraint in terms of out of sample performance relatively to the unconstrained estimator or H5. The asymptotic distribution of S_m is standard normal under the null.

	H1	H2	H3	H4	H5	H3
	versus unconstrained model					versus H5
	PRI					
6E	0.22	0.24	5.14	-31.04	5.35	-0.22
6J	2.43	2.45	2.75	-44.97	2.67	0.08
6S	3.27	3.33	3.79	-14.44	3.68	0.11
6B	0.05	0.05	1.17	-13.13	1.18	-0.02
6A	-0.60	-0.60	0.89	-20.29	1.55	-0.67
6N	-0.03	0.01	0.21	-16.59	1.51	-1.32
6C	0.12	0.12	1.34	-59.32	1.08	0.27
	S_m					
6E	-0.55	-0.59	-2.67	2.81	-2.74	0.76
6J	-2.05	-2.06	-2.19	4.12	-2.24	-0.49
6S	-2.48	-2.52	-2.00	2.01	-1.95	-1.69
6B	-0.06	-0.06	-0.61	1.76	-0.62	1.00
6A	0.55	0.54	-0.83	2.95	-1.91	0.54
6N	0.06	-0.02	-0.38	1.84	-3.11	1.49
6C	-0.25	-0.26	-1.87	3.74	-1.85	-0.35

3.3 Comparison to Volume Prediction Using Daily Data

We compare predictions with the model that only predicts the end of day volumes

$$V_i(T) = c_0 + \sum_{l=1}^{p} c_l V_{i-l}(T) + \varepsilon_i(T) \tag{14}$$

where the number of lags is $p = 8$.

To give a positive bias in favour of the daily model, the coefficients c_l $(l = 0, 1, ..., 8)$ in (14) are estimated using the whole sample. Despite this, it is clear that (14) shall not outperform (12), even if the parameters of the latter are estimated on the estimation sample. The goal is to derive a reasonable lower bound on the value of using intraday information. To this end we compare (14) to the prediction from the model in Hypothesis 1, which is the unconstrained model except for not benefiting from the inclusion of volumes from ES. The end of day volume prediction made from (14) is $Pred_{i,daily} := \hat{c}_0 + \sum_{l=1}^{p} \hat{c}_l V_{i-l}(T)$, where the coefficients are replaced by the estimated ones denoted by a hat. As an illustrative example, considering only the last five days in the test sample, Figure 3 shows the plot of the actual volumes, the prediction $Pred_{i,daily}$ and the prediction that uses intraday information $Pred_{i,intraday}(t) = V_i(t) + X_i(t)' \hat{b}(t)$. From Hypothesis 1, recall that $X_i(t)' = [1, V_i(t), V_{i-1}(t), V_{i-1}(T)]'$, in this case.

The intraday square error when using a daily prediction is

$$SE_{i,daily} = (V_i(T) - Pred_{i,daily})^2.$$

Note that

$$\frac{1}{N}\sum_{j=1}^{N} (Y_i(t_j) - (Pred_{i,daily} - V_i(t_j)))^2 = (V_i(T) - Pred_{i,daily})^2. \tag{15}$$

We can use again the PRI computed on the test sample:

$$PRI := 100\frac{MSE_{daily} - MSE_{intra}}{MSE_{daily}} \tag{16}$$

where MSE_{daily} is computed from $SE_{i,daily}$ and MSE_{intra} is the MSE from the intraday model (12) under Hypothesis 1. From (15) we see that the daily prediction for $Y_i(t)$ is $(Pred_{i,daily} - V_i(t_j))$. This quantity can be negative, while $Y_i(t) > 0$. Hence, we also

Figure 3: Volume Predictions. Predictions based on the daily model (14) and the intraday model (12) under Hypothesis 1 are plotted together with the actual end of day volumes for the last five days of the sample.

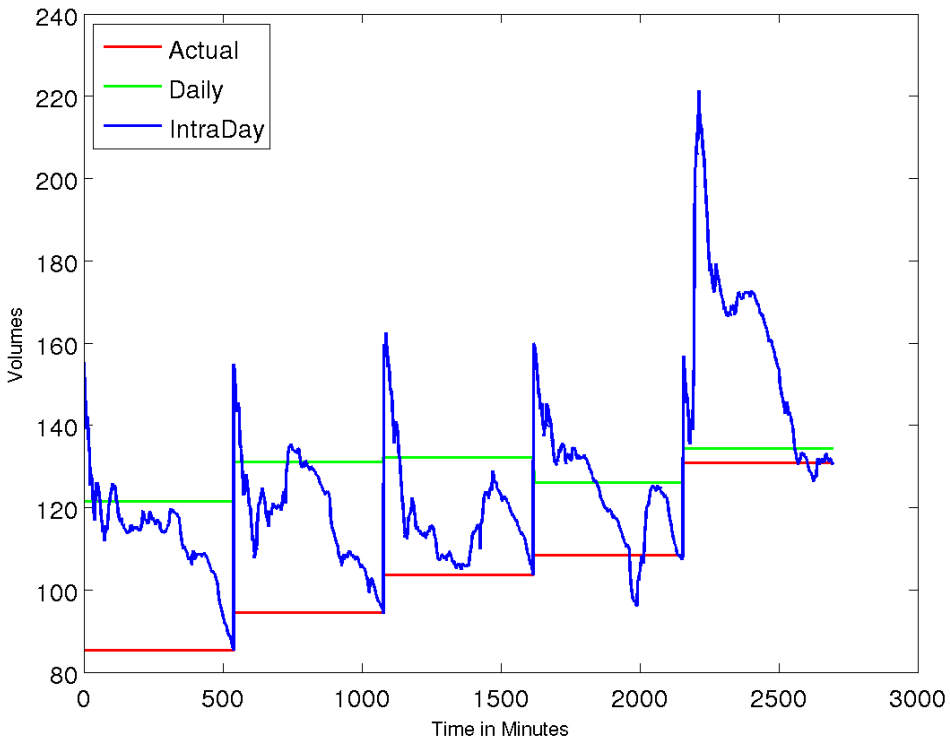


Table 3: PRI for Daily Versus Intraday Model Predictions. Intra/Day reports (16) based on model (12) under the restriction of Hypothesis 1, against the daily prediction (14). IntraNaive/Day reports (16) when we replace MSE_{Intra} with $MSE_{IntraNaive}$, which is the MSE based on (17). Intra/IntraNaive reports (16) when we replace MSE_{daily} with $MSE_{IntraNaive}$ based on (17), while MSE_{intra} refers to the model in (12).

	PRI		
	IntraNaive/Day	Intra/Day	Intra/IntraNaive
6E	105800	118940	13240
6J	42263	82556	40392
6S	2601	3094	594
6B	23555	35196	11741
6A	13701	13901	301
6N	394	467	173
6C	6786	7892	1206

compute the prediction

$$IntraNaive := \max\left\{Pred_{i,daily} - V_i\left(t_j\right), 0\right\} \tag{17}$$

This allows us to quantify the gain from naively using intraday cumulative volumes to improve a daily prediction. Results are in Table 3.

The improvement in using the methodology presented here is huge. However, it is worth noting that making naive use of intraday information as in (17) leads to considerable improvements as expected. However, it is in no way comparable to the gain from using the proposed methodology.

4 Some Finite Sample Analysis via Simulations

We provide some finite sample evidence of the estimator's prediction performance and finite sample distribution using a simulation design that mimics the data from the empirical section. We also focus on the comparison between the estimators in (8) and (9). In both cases, We consider 500 simulations with sample sizes of $n \in \{60, 180\}$. We simulate a population $\left\{\left(V_i\left(t\right)\right)_{t\in\mathcal{T}_N} : i = 1, 2, ..., n_{pop}\right\}$ with $n_{pop} = 10^5$ starting from the mean of the sample data for 6E. This ensures that we capture the same seasonal patterns and overall characteristics of the actual data. We generate data that can have different

volatility and persistency levels. By making the single volumes (before cumulation) more or less volatile we obtain what we call low and high volatility. Persistency is induced by making each individual volume more or less dependent via an autoregressive model. The details of how this population is simulated is given in the Appendix in Section A.2. A few sample trajectories for 6E from such population are shown in Figure 4. This can be compared to the volume curves in Figure 1 for 6E. We consider 500 samples without replacement from the population. Each sample sizes is $n \in \{20, 100\}$ days and within each day we have $N = 539$ once we delete the first and last observation, as in Section 3.

True model. The true model is

$$Y_i(t) = b_0^{(1)}(t) + V_i(t)\, b_0^{(2)}(t) + \varepsilon_i(t). \tag{18}$$

The values of $b_0^{(1)}$ and $b_0^{(2)}$ are set equal to the population OLS estimators.

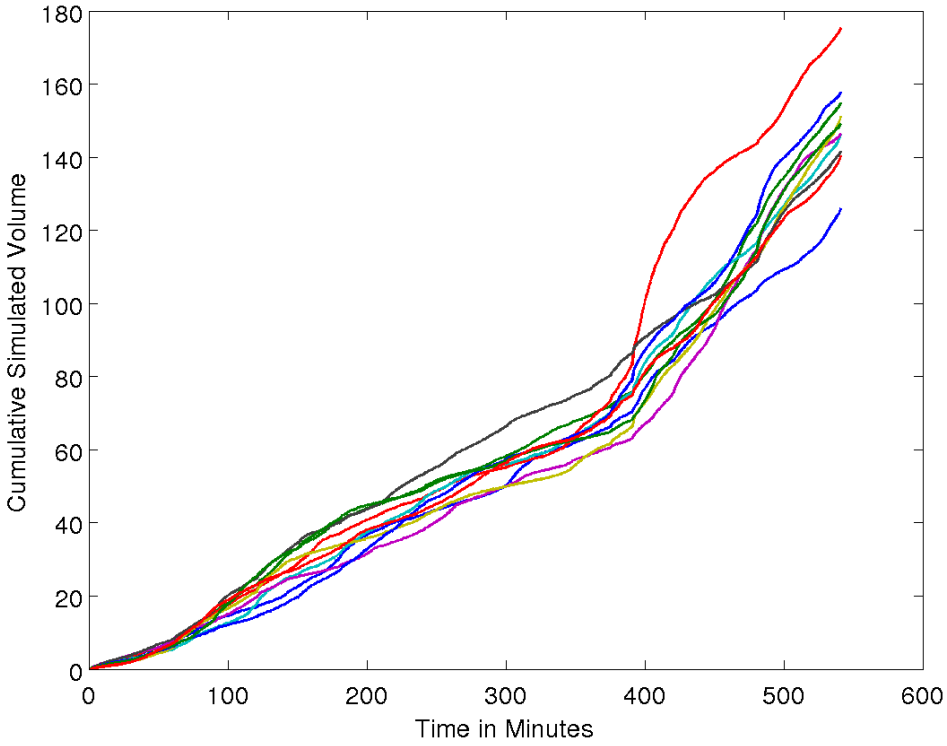
Estimation and evaluation criteria For each sample we estimate the unconstrained and the two constrained estimators from (8) and (9). The constraints are $b_0^{(1)}$decreasing and $b_0^{(2)}$ convex decreasing under the null. This follows the observation from the empirical section. For each sample i, we compute the relative error (RE)

$$RE_i^{(k)} := \frac{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\hat{b}^{(k)}(t_j) - b_0^{(k)}(t_j)\right)^2}}{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(b_0^{(k)}(t_j)\right)^2}},\ k = 1, 2, \tag{19}$$

where a smaller number is preferred to a larger one. A number smaller than 1 means an improvement over a zero function. The results for different values of n, volatility and persistency of $V_i(t)$ are reported in Table 4. These results show that there is a considerable improvement in using a constrained estimator. Moreover, the estimator in (8) fares better than the one in (9) as expected. We also note that when persistency is low, the performance of the estimator for $b_0^{(1)}$ is poor. This is expected: the lower the persistency, the lower is the predictability using past volumes.

For the unconstrained estimator, we also compute $\left(\hat{b}^{(k)}(t) - b_0^{(k)}(t)\right)/s.e.\left(\hat{b}^{(k)}(t)\right)$, where $s.e.\left(\hat{b}^{(k)}(t)\right)$ is the standard error derived from the simulated data. For each $t \in \mathcal{T}_N$ and each sample, this is approximately standard normal. Hence, from our

Figure 4: Simulated Data. Plot of the realizations when volatility is low and persistency is low (top panel) and when volatility is high and persistency is high (bottom panel).



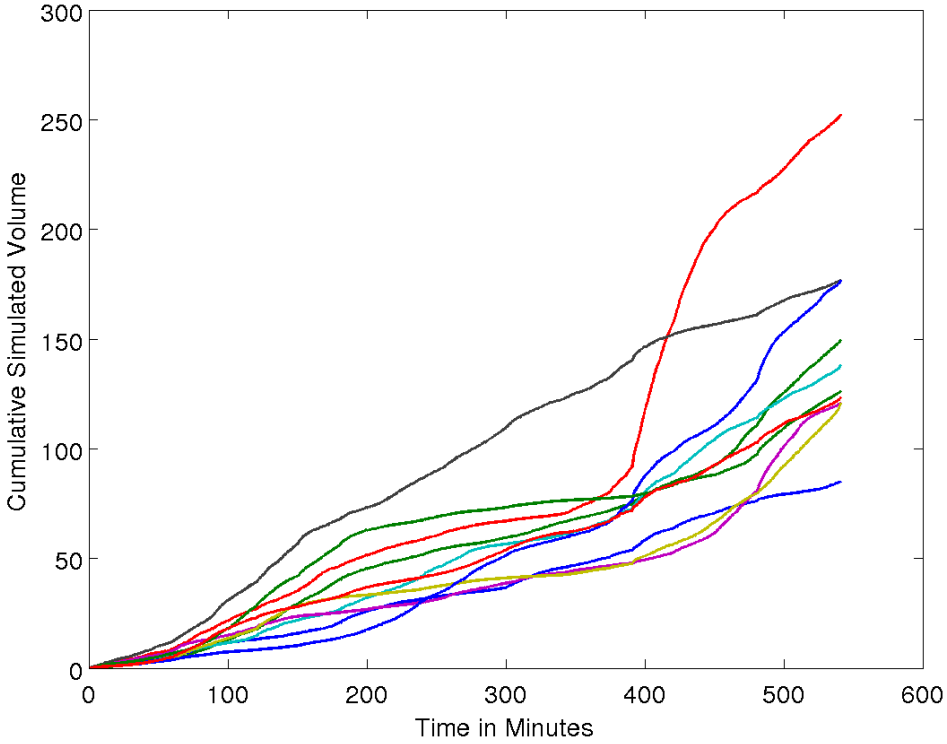


Table 4: Average Relative Error for Different Estimators. The average over 500 simulations is reported for the relative error in (19) for the unconstrained estimator (U) in (3), the constrained simple estimator (CS) in (9) and the constrained estimator (C) in (8). A smaller number is preferred to a large.

	$b^{(1)}$			$b^{(2)}$		
Noise/Persistency	U	CS	C	U	CS	C
			$n = 60$			
low/low	0.12	0.12	0.10	2.62	1.94	1.77
low/high	0.14	0.14	0.13	0.65	0.61	0.60
high/low	0.12	0.12	0.10	2.97	2.21	2.02
high/high	0.16	0.16	0.15	0.76	0.70	0.68
			$n = 180$			
low/low	0.07	0.07	0.06	1.46	1.20	1.13
low/high	0.09	0.09	0.10	0.39	0.38	0.38
high/low	0.07	0.07	0.06	1.62	1.33	1.26
high/high	0.10	0.10	0.10	0.46	0.46	0.45

500 simulations and $N = 539$ points at which we compute $t \in \mathcal{T}_N$, we have a total of $500 \times 539 = 269500$ approximately standard normal random variables. Table 5 reports empirical quantiles for the various simulations designs and compares them to the standard normal ones. The normal approximation appears to be reasonable.

5 Conclusion

This paper considered the problem of estimating the residual volume to the end of day. This quantity was defined in the introduction and it is a major ingredient in optimal trading execution. The estimation problem was cast in the framework of functional data. Tools to conduct inference have been provided in the paper and these include the methodology on how to construct confidence bands (Theorems 1, 2, and 3) and how to compare different restrictions (Theorem 4). Practical implementation details have been given so that the estimation and inference can be conducted using standard software such as MATLAB. The main functions needed for computing the estimators and conducting inference are also publicly available from the GitHub repository URL:https://github.com/asancetta/FDRegression.

Table 5: Quantiles of the Centered and Standardized Unconstrained Estimator. For different simulation designs, the quantiles at different probability levels are reported for the properly centered and scaled estimator. This is asymptotically distributed as a standard normal random variable. The quantiles for the standard normal distribution are reported at the end for comparison.

Probability	0.01	0.025	0.05	0.95	0.975	0.99
Noise/Persistency/n						
			$b^{(1)}$			
low/low/60	-2.34	-1.96	-1.65	1.65	1.95	2.37
low/high/60	-2.30	-1.86	-1.57	1.65	2.00	2.37
high/low/60	-2.39	-1.96	-1.64	1.62	1.99	2.39
high/high/60	-2.44	-1.98	-1.67	1.57	1.91	2.39
low/low/180	-2.44	-1.99	-1.68	1.64	1.96	2.34
low/high/180	-2.38	-1.97	-1.66	1.60	1.92	2.23
high/low/180	-2.51	-2.06	-1.67	1.61	1.94	2.32
high/high/180	-2.54	-2.08	-1.70	1.56	1.84	2.12
			$b^{(2)}$			
low/low/60	-2.28	-1.91	-1.63	1.68	2.00	2.38
low/high/60	-2.26	-1.92	-1.64	1.56	1.92	2.38
high/low/60	-2.24	-1.85	-1.55	1.69	2.07	2.51
high/high/60	-2.11	-1.77	-1.50	1.67	2.13	2.73
low/low/180	-2.35	-1.93	-1.62	1.68	1.98	2.44
low/high/180	-2.18	-1.89	-1.61	1.65	1.98	2.40
high/low/180	-2.28	-1.86	-1.57	1.70	2.06	2.57
high/high/180	-1.98	-1.71	-1.47	1.74	2.16	2.69
Normal Quantile	-2.33	-1.96	-1.64	1.64	1.96	2.33

The empirical section focused on the end of day volume prediction of major CME Fx futures. We found that a simple restricted model that possibly includes information from the mini S&P futures does outperform the unrestricted model over the sample period. We carried out a prediction exercise to estimate a lower bound on the predictive value of including intraday information. As expected, the improvement is huge.

We included a simulation study to show the behaviour of two constrained estimators and provides evidence of the finite sample behaviour of the OLS estimator. These simulations show that, whenever possible, we should use the constrained estimator (8) that uses full information of the cross dependence of the regressors as opposed to (9). Additional simulations, carried out by the author, but using different designs, also confirm this claim.

In order to focus on the main idea, this paper left out important aspects of prediction. Decomposition of volumes into various components at different frequencies as done in Brownlees et al. (2010) is one of such important topics that have not been included. In applications, this should be incorporated into the present functional data framework in order to improve the forecast. In the empirical application we did not find a significant benefit in the inclusion of the volumes of the e-mini S&P500 futures. This is surprising. A more in depth analysis would be required in order to assess the extent to which the e-mini futures could drive the volumes of the other instruments. Granger Causality for functional data (Saumard, 2017) could be employed to address this question. Such analysis is left to future research.

References

[1] Almgren, R., and N. Chriss (2001) Optimal Execution of Portfolio Transactions. Journal of Risk 3. 5-40.

[2] Bosq, D. (2000) Linear Processes in Function Spaces. New York: Springer.

[3] Brémaud, P. (1988) Point Processes and Queues: Martingale Dynamics. New York: Springer.

[4] Brownlees, C.T., F. Cipollini and G.M. Gallo (2011) Intra-Day Volume Modelling and Prediction for Algorithmic Trading. Journal of Financial Econometrics 9, 489-518.

[5] Cox, D.R. (1975) A Note on Data-Splitting for the Evaluation of Significance Levels. Biometrika 62, 441-444.

[6] Dawid, A. P. (1997) Prequential Analysis. In S. Kotz, C.B. Read and D.L. Banks (eds.), Encyclopedia of Statistical Sciences 1, 464-470. New York: Wiley-Interscience.

[7] De Vito, E., V. Umanitá, S. Villa (2013) An Extension of Mercer theorem to matrix-valued measurable kernels. Applied Computational Harmonic Analysis 34, 339-351.

[8] Diebold, F.X. and R.S. Mariano (1995) Comparing Predictive Accuracy. Journal of Business and Economic Statistics 13, 253-263.

[9] Engle, R.F. and J.R. Russell (1998) Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66, 1127-1162.

[10] Engle, R.F. and M.E. Sokalska (2012) Forecasting Intraday Volatility in the US Equity Market. Multiplicative Component GARCH. Journal of Financial Econometrics 10, 54-83.

[11] Fan, Y., and Q. Li (1996) Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms. Econometrica 64, 865-890.

[12] Gatheral, J. (2010) No-Dynamic-Arbitrage and Market Impact. Quantitative Finance 10, 749-759.

[13] Geyer, C.J. (1994) On the Asymptotics of Constrained M-Estimation. Annals of Statistics 22, 1993-2010.

[14] Horváth, L. and P. Kokoszka (2012) Inference for Functional Data with Applications. New York: Springer.

[15] Kokoszka, P., H. Miao and X. Zhang (2015) Functional Dynamic Factor Model for Intraday Price Curves. Journal of Financial Econometrics 13, 456-477.

[16] McLeish, D.L. (1974) Dependent Central Limit Theorems and Invariance Principles. Annals of Probability 2, 620-628.

[17] Sancetta, A. (2015) A Nonparametric Estimator for the Covariance Function of Functional Data. Econometric Theory 31, 1359-1381.

[18] Sancetta, A. (2018) Estimation for the Prediction of Point Processes with Many Covariates. Econometric Theory 34, 598-627.

[19] Saumard, M. (2017) Linear Causality in the Sense of Granger with Stationary Functional Time Series. In G. Aneiros, E.G. Bongiorno, R. Cao and P. Vieu (eds.), Functional Statistics and Related Fields, 225-231. Berlin: Springer.

[20] Seillier-Moiseiwitsch, F. and A.P. Dawid (1993). On Testing the Validity of Sequential Probability Forecasts. Journal of the American Statistical Association 88, 355-359.

[21] Szucs, J.M. (1985) Ergodic Theorems for Tensor Products. Journal of Functional Analysis 64, 125-133.

[22] Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016) Review of Functional Data Analysis. Annual Review of Statistics and Its Application 3, 257-295.

[23] Yatchew, A.J. (1992) Nonparametric Regression Tests Based on Least Squares. Econometric Theory 8, 435-451.

[24] Yatchew, A.J., and W. Härdle (2006) Nonparametric State Price Density Estimation Using Constrained Least Squares and the Bootstrap. Journal of Econometrics 133, 579-599.

[25] Zheng, J., (1996) A Consistent Test of Functional Form via Nonparametric Estimation Techniques. Journal of Econometrics 75, 263-289.

Appendix

A.1 Proofs

As mentioned in Section 2.1, we shall use the same symbol for an operator and its kernel. We shall call covariance any self adjoint operator. Hence, we shall call C_{XX} covariance function even when $\mathbb{E}X$ is not zero. For a $K \times K$ matrix valued covariance function $C(s,t)$ $s,t \in \mathcal{T}$, the Hilbert-Schmidt norm of C can be written as $|C|_{\mathcal{S}} = \sqrt{\int_{\mathcal{T}} \int_{\mathcal{T}} |C(s,t)|_F^2 \, dsdt}$ where $|\cdot|_F$ is the Frobenius norm of the matrix $C(s,t)$, i.e. $|C(s,t)|_F^2 = \text{Trace}\left(C(s,t)' C(s,t)\right)$.

We also use the symbol $\lesssim$ when the l.h.s. is bounded by a constant times the r.h.s.

A.1.1 Preliminary Lemmas

The following establishes the convergence of $\hat{C}_{XX}$.

Lemma 2 *Under Condition 1,* $\left|\hat{C}_{XX} - C_{XX}\right|_{\mathcal{S}} \to 0$ *and* $\sup_{t \in \mathcal{T}} \left|\hat{C}_{XX}(t,t) - C_{XX}(t,t)\right|_F^2 \to 0$ *in probability, in both cases.*

Proof. Note that $\hat{C}_{XX}(s,t) = \frac{1}{n} \sum_{i=1}^n X_i(s) X_i(t)'$. Hence we first use an ergodic theorem for the product $W_i(s,t) := X_i(s) X_i(t)'$ $s,t \in \mathcal{T}$. The variables (X_i) are stationary and ergodic random variables in the Hilbert space $\mathcal{H}^K$ equipped with norm $|\cdot|_{\mathcal{H}^K}$ .The variables (W_i) are random variables with values in $\mathcal{G} := \mathcal{H}^K \otimes \mathcal{H}^K$ equipped with the norm $|\cdot|_{\mathcal{H}^K \otimes \mathcal{H}^K} = |\cdot|_{\mathcal{S}}$. Here, $\mathcal{H}^K \otimes \mathcal{H}^K$ is the tensor product of $\mathcal{H}^K$ and $\mathcal{H}^K$. Hence, by definition of the Hilbert-Schmidt norm,

$$|W_i|_{\mathcal{S}} = \sqrt{\int_{\mathcal{T}} \int_{\mathcal{T}} \sum_{k,l=1}^K \left|X_i^{(k)}(s) X_i^{(l)}(t)\right|^2 dsdt} = |X_i|_{\mathcal{H}^K} |X_i|_{\mathcal{H}^K} . \tag{A.1}$$

In consequence, Theorem 1 in Szucs (1982) says that (W_i) is ergodic in $\mathcal{G}$. Then, by the von Neumann Ergodic Theorem, and stationarity, $\left|\frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E}) W_i\right|_{\mathcal{S}} \to 0$ in probability. This proves the first convergence. We want to turn this convergence into uniform. This is possible if

$$\lim_{\rho \to 0} \mathbb{E} \sup_{|t_1 - t_2| \leq \rho, |s_1 - s_2| \leq \rho} \left|W_i^{(k,l)}(s_1,t_1) - W_i^{(k,l)}(s_2,t_2)\right| = 0, \tag{A.2}$$

i.e. showing stochastic equicontinuity. To this end, use the inequality $|ab-cd| \leq |a-c|\,|b| + |c|\,|b-d|$ for any real valued a, b, c, d, to deduce that

$$\begin{aligned}\left|W_i^{(k,l)}(s_1,t_1) - W_i^{(k,l)}(s_2,t_2)\right| &= \left|X_i^{(k)}(s_1)\,X_i^{(l)}(t_1) - X_i^{(k)}(s_2)\,X_i^{(l)}(t_2)\right| \\ &\leq \left|X_i^{(k)}(s_1) - X_i^{(k)}(s_2)\right|\left|X_i^{(l)}(t_1)\right| \\ &\quad + \left|X_i^{(k)}(s_2)\right|\left|X_i^{(l)}(t_1) - X_i^{(l)}(t_2)\right|.\end{aligned}$$

Given that the $\left|X_i^{(l)}(s) - X_i^{(l)}(t)\right| \lesssim |s-t|^{\alpha}$ we deduce that (A.2) holds and the lemma is proved. ■

The following is a bound on the minimal eigenvalue of $\hat{C}_{XX}(t,t)$ uniformly in $t \in \mathcal{T}$.

Lemma 3 *Under Condition 1,* $\sup_{t\in\mathcal{T}} \sup_{|x|_2\leq 1} \left|\left[\hat{C}_{XX}(t,t)\right]^{-1} x\right|_2 \to \sup_{t\in\mathcal{T}} [\lambda_{\min}(t)]^{-1} \leq \underline{\lambda}^{-1} < \infty$ *in probability as* $n \to \infty$.

Proof. Let D be a $K \times K$ symmetric matrix. Then, for $x \in \mathbb{R}^K$, $|D^{-1}x|_2^2 \leq [\sigma_K^2(D)]^{-1}|x|_2^2$ where $\sigma_K(D)$ is the smallest singular value of D. In this proof $\sigma_k(\cdot)$ will denote the the k^{th} singular value of its argument, and the singular values are ordered in decreasing order, i.e. $\sigma_k(D) \geq \sigma_{k+1}(D)$. If all eigenvalues of D are positive, then $\sigma_k(D)$ is the k^{th} eigenvalue of D, ordered in decreasing order. Let A and B be two positive definite matrices. Then, by Weyl's inequality deduce that $\sigma_K(A+B) > \sigma_K(A) - \sigma_1(B)$. Clearly, $\sigma_1^2(B) \leq \sum_{k=1}^K \sigma_k^2(B) = |B|_F^2$ by the properties of the Frobenius norm. Write $\hat{C}_{XX}(t,t) = C_{XX}(t,t) + \left[\hat{C}_{XX}(t,t) - C_{XX}(t,t)\right]$. By Condition 1, $\sigma_K(C_{XX}(t,t)) \geq \underline{\lambda} > 0$. By Lemma 2, $\sup_{t\in\mathcal{T}} \left|\hat{C}_{XX}(t,t) - C_{XX}(t,t)\right|_F^2 \to 0$ in probability. By these remarks letting $D = D(t) = \hat{C}(t,t)$ at the start of the proof, and setting $A = C_{XX}(t,t)$ and $B = D - A$, deduce that the lemma holds. ■

The Lemma also means that $\inf_{t\in\mathcal{T}} \inf_{|x|_2=1} \left|\hat{C}_{XX}(t,t)\,x\right|_2 \to \underline{\lambda} > 0$ in probability.

Next we state a classical central limit theorem for random variables with values in a Hilbert space (Bosq, 2000), but slightly simplified.

Lemma 4 *Let* $(W_i)_{i\in\mathbb{Z}}$ *be a martingale difference sequence with values in a Hilbert space* $\mathcal{W}$ *of vector valued functions on* $\mathcal{T}$ *with inner product* $\langle\cdot,\cdot\rangle_{\mathcal{W}}$. *Suppose that* $C_{WW}(s,t) := \mathbb{E}W_i(s)\,W_i(t)'$ *for* $i \in \mathbb{Z}$. *Let* $(e_l)_{l\geq 1}$ *be an orthonormal basis for* $\mathcal{W}$. *Suppose that the following hold:*

1. $\lim_n \mathbb{E}\max_{i\leq n} |W_i|_{\mathcal{W}}/\sqrt{n} = 0$;

2. $\lim_n \frac{1}{n}\sum_{i=1}^n \langle W_i, e_k\rangle_{\mathcal{W}}\langle W_i, e_l\rangle_{\mathcal{W}} = \psi_{k,l}$ *a.s. for real numbers* $(\psi_{k,l})_{k,l\geq 1}$*;*

3. $\lim_{L\to\infty} \limsup_n \Pr\left(\sum_{i=1}^n r_L^2\left(n^{-1/2}W_i\right) > \epsilon\right) = 0$ *for any* $\epsilon > 0$*, where* $r_L^2(w) = \sum_{l\geq L}\langle w, e_k\rangle_{\mathcal{W}}^2$ *for* $w \in \mathcal{W}$*.*

Then, $\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \to G$ *weakly in* $\mathcal{W}$*, where* $G = (G(t))_{t\in\mathcal{T}}$ *is a mean zero Gaussian process with values in* $\mathcal{W}$ *and matrix valued covariance function* $\mathbb{E}G(s)G(t) = C_{WW}(s,t)$ *such that* $\int_{\mathcal{T}}\int_{\mathcal{T}} e_k(s)' C_{WW}(s,t) e_k(t)\, dsdt = \psi_{k,l}$*,* $k, l \geq 1$*.*

Lemma 5 *Under Condition 1,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i(t)\varepsilon_i(t) \to G(t)\ t \in \mathcal{T}$$

weakly in $\mathcal{H}$*, where* G *is a mean zero Gaussian process with matrix covariance function* $\mathbb{E}G(s)G(t) = C_\sigma(s,t)$*.*

Proof. We use Lemma 4 with $W_i(t) := X_i(t)\varepsilon_i(t)$ and inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}^K}$. Then, $(W_i)_{i\in\mathbb{Z}}$ is a sequence of stationary martingale differences adapted to $(\mathcal{F}_i)_{i\in\mathbb{Z}}$. By the bound

$$\mathbb{E}\max_{i\leq n}\frac{|W_i|_{\mathcal{H}^K}}{\sqrt{n}} \leq \left[\mathbb{E}\max_{i\leq n}\frac{|W_i|_{\mathcal{H}^K}^2}{n}\right]^{1/2}$$

we just need to show that the r.h.s. goes to zero. The quantity $|W_i|_{\mathcal{H}^K}^2$ can be split as $|W_i|_{\mathcal{H}^K}^2 1\{|W_i|_{\mathcal{H}^K} \leq \epsilon\sqrt{n}\}$ plus $\mathbb{E}|W_i|_{\mathcal{H}^K}^2 1\{|W_i|_{\mathcal{H}^K} > \epsilon\sqrt{n}\}$; here $1\{\cdot\}$ is the indicator function. The first of such terms is bounded above by $\epsilon^2 n$. Hence, inserting in the r.h.s. of the above display we have the upper bound

$$\left[\epsilon^2 + \mathbb{E}\max_{i\leq n}\frac{|W_i|_{\mathcal{H}^K}^2 1\{|W_i|_{\mathcal{H}^K} > \epsilon\sqrt{n}\}}{n}\right]^{1/2}.$$

Bounding the maximum by the sum we deduce that the expectation of the maximum is less than

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}|W_i|_{\mathcal{H}^K}^2 1\left\{|W_i|_{\mathcal{H}^K} > \epsilon\sqrt{n}\right\}.$$

Using stationarity this display is equal to $\mathbb{E}|W_1|_{\mathcal{H}^K}^2 1\{|W_1|_{\mathcal{H}^K} > \epsilon\sqrt{n}\}$ which goes to zero as n goes to infinity for any $\epsilon > 0$, because $\mathbb{E}|W_1|_{\mathcal{H}^K}^2 < \infty$ by Condition 1. Letting $\epsilon \to 0$ slowly enough, verifies Point 1 in Lemma 4.

Now define the real variable $\bar{W}_{i,l} = \int_{\mathcal{T}} e_l(t)' X_i(t)\varepsilon_i(t)\, dt$ where $\{e_l : l \geq 1\}$ is an orthonormal basis of $\mathcal{H}^K$. By ergodicity of $\{(X_i, \varepsilon_i) : i \in \mathbb{Z}\}$, $\frac{1}{n}\sum_{i=1}^n \bar{W}_{i,k}\bar{W}_{i,l} \to \mathbb{E}\bar{W}_{i,k}\bar{W}_{i,l}$

a.s., where

$$\begin{aligned}\mathbb{E}\bar{W}_{i,k}\bar{W}_{i,l} &= \mathbb{E}\int_{\mathcal{T}}\int_{\mathcal{T}} e_l(s)'\left[\varepsilon_i(s)\,\varepsilon_i(t)\,X_i(s)\,X_i(t)'\right]e_l(t)\,dsdt \\ &= \int_{\mathcal{T}}\int_{\mathcal{T}} e_l(s)'\,C_\sigma(s,t)\,e_l(s)\,dsdt = \psi_{k,l}\end{aligned}$$

where the r.h.s. is as in Lemma 4. We need to show that this is a finite quantity in absolute value. Note that

$$\left|\bar{W}_{i,l}\right| = \left|\langle e_l, \varepsilon_i X_i\rangle_{\mathcal{H}^K}\right| \leq |e_l|_{\mathcal{H}^K}\,|\varepsilon_i X_i|_{\mathcal{H}^K} = |\varepsilon_i X_i|_{\mathcal{H}^K}$$

where the last equality follows because e_l has unit norm. In consequence, $\left|\mathbb{E}\bar{W}_{i,k}\bar{W}_{i,l}\right| \leq \mathbb{E}\,|\varepsilon_i X_i|^2_{\mathcal{H}^K} < \infty$. This proves Point 2 in Lemma 4.

To prove Point 3 in Lemma 4, from the previous display, deduce that $W_i(t)$ has covariance $C_\sigma(s,t)$, $s,t \in \mathcal{T}$. The operator associated to this covariance has finite Hilbert-Schmidt norm:

$$|C_\sigma|^2_{\mathcal{S}} = \int_{\mathcal{T}}\int_{\mathcal{T}}\sum_{k,l=1}^{K}\left|\mathbb{E}\varepsilon_i(s)\,\varepsilon_i(t)\,X_i^{(k)}(s)\,X_i^{(l)}(t)\right|^2 dsdt \leq \mathbb{E}\,|\varepsilon_i X_i|^2_{\mathcal{H}^K} < \infty$$

where the inequality uses Jensen inequality and the same argument used to derive (A.1). Hence, the operator associated to $C_\sigma(s,t)$ is compact so that $C_\sigma(s,t)$ admits the expansion $\sum_{l=1}^{\infty}\rho_l\Psi_l(s)\,\Psi_l(t)'$ for non negative coefficients ρ_l such that $\sum_{l=1}^{\infty}\rho_l^2 < \infty$, and a basis $\{\Psi_l : l = 1,2,...\}$ orthonormal under $\langle\cdot,\cdot\rangle_{\mathcal{H}^K}$. The expansion holds under the Hilbert-Schmidt norm by the the spectral theorem. (It is of interest to note that it also holds under the uniform norm by the matrix valued version of Mercer theorem (de Vito et al., 2013).) Also note that

$$\int_{\mathcal{T}}\mathrm{Trace}\left(C_\sigma(t,t)\right)dt = \int_{\mathcal{T}}\sum_{k=1}^{K}\mathbb{E}\varepsilon_i^2(t)\left|X_i^{(k)}(t)\right|^2 dt = \mathbb{E}\,|\varepsilon_i X_i|^2_{\mathcal{H}^K} < \infty.$$

This means that $\sum_{l=1}^{\infty} \rho_l < \infty$ because

$$\begin{aligned}\int_{\mathcal{T}} \text{Trace}\left(C_\sigma\left(t,t\right)\right) dt &= \int_{\mathcal{T}} \text{Trace}\left(\sum_{l=1}^{\infty} \rho_l \Psi_l\left(t\right) \Psi_l\left(t\right)'\right) dt \\ &= \sum_{l=1}^{\infty} \rho_l \int_{\mathcal{T}} \text{Trace}\left(\Psi_l\left(t\right)' \Psi_l\left(t\right)\right) dt = \sum_{l=1}^{\infty} \rho_l\end{aligned}$$

using the properties of the trace. In the statement of the lemma, we choose $e_l = \Psi_l$ and let $W_i\left(t\right) = \sum_{l=1}^{\infty} \rho_l^{1/2} \eta_{i,l} \Psi_l\left(t\right)$ for uncorrelated real valued mean zero variance one random variables $\eta_{i,l}$; the equality holds under the norm $\left|\cdot\right|_{\mathcal{H}^K}$. (It is again of interest to note that it holds under the uniform norm by the Karhunen-Loeve Theorem, (Bosq, 2000), whose vector valued version follows directly from the matrix valued Mercer theorem because of continuity of C_σ.) This allows us to . In consequence, $\left\langle W_i, \Psi_l \right\rangle_{\mathcal{H}^K} = \rho_l^{1/2} \eta_{i,l}$ so that by Chebyshev's inequality,

$$\Pr\left(\sum_{i=1}^{n} r_L^2\left(n^{-1/2} W_i\right) > \epsilon\right) \leq \sum_{l \geq L} \rho_l \to 0$$

using the notation in Lemma 4, stationarity, and the summability of the eigenvalues ρ_l. Hence, Lemma 4 applies and the proof is completed. ■

A.1.2 Proof of Theorems and Results in the Text

Proof. [Theorem 1]

By standard argument for ordinary least square,

$$\sqrt{n}\left(\hat{b}\left(t\right) - b_0\left(t\right)\right) = \left[\hat{C}_{XX}\left(t,t\right)\right]^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i\left(t\right) \varepsilon_i\left(t\right). \tag{A.3}$$

By Lemma 5,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\left(t\right) \varepsilon_i\left(t\right) \to G\left(t\right), \, t \in \mathcal{T}$$

weakly in $\mathcal{H}^K$, where G is a mean zero Gaussian process with matrix covariance function

$$\mathbb{E} G\left(s\right) G\left(t\right) = C_\sigma\left(s,t\right).$$

For later reference, this also means that $\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\varepsilon_i\right|_{\mathcal{H}^K} = O_p(1)$. By Lemma 3,

$$\left|\left[\hat{C}_{XX}(t,t)\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i(t)\,\varepsilon_i(t)\right|_2^2 \leq \left(\underline{\lambda} - o_p(1)\right)^{-1}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i(t)\,\varepsilon_i(t)\right|_2^2$$

so that $\left|\sqrt{n}\left(\hat{b}-b_0\right)\right|_{\mathcal{H}^K} = O_p(1)$. Note that the $o_p(1)$ term in the above display is uniform in $t \in \mathcal{T}$. This establishes tightness of (A.3). By Lemmas 2 and 3, we can replace $\left[\hat{C}_{XX}(t,t)\right]^{-1}$ with $[C_{XX}(t,t)]^{-1}$. Hence, Defining $G_X(t) := [C_{XX}(t,t)]^{-1} G(t)$ We can now deduce that

$$\sqrt{n}\left(\hat{b}(t) - b_0(t)\right) \to G_X(t),\ t \in \mathcal{T},$$

weakly in $\mathcal{H}^K$, as stated in the theorem. ■

Proof. [Theorem 2] Define

$$\hat{Q}(b) := -\int_{\mathcal{T}} b(t)'\,\hat{b}(t)\,dt + \frac{1}{2}\int_{\mathcal{T}} b(t)'\,b(t)\,dt.$$

Note that the minimizer of $\hat{Q}(b)$ and $\hat{Q}(b) - \hat{Q}(b_0)$ are the same. Hence, adding and subtracting $\int_{\mathcal{T}} (b(t) - b_0(t))'\,b_0(t)\,dt$, deduce that

$$\hat{Q}(b) - \hat{Q}(b_0) = -\int_{\mathcal{T}} (b(t) - b_0(t))'\left(\hat{b}(t) - b_0(t)\right)dt + \frac{1}{2}\int_{\mathcal{T}} (b(t) - b_0(t))'\,(b(t) - b_0(t))\,dt.$$

Define $\hat{\delta} = \sqrt{n}\left(\hat{b} - b_0\right)$ and $\delta = \sqrt{n}(b - b_0)$ and minimize

$$n\hat{Q}\left(\delta, \hat{\delta}\right) := -\int_{\mathcal{T}} \delta(t)'\,\hat{\delta}(t)\,dt + \frac{1}{2}\int_{\mathcal{T}} \delta(t)'\,\delta(t)\,dt. \tag{A.4}$$

w.r.t. $\delta \in \mathcal{C}_0$. For any δ contained in a ball of finite radius under the norm $|\cdot|_{\mathcal{H}^K}$, $\int_{\mathcal{T}} \delta(t)'\,\hat{\delta}(t)\,dt \to \int_{\mathcal{T}} \delta(t)'\,G_X(t)\,dt$ in distribution using Theorem 1 and the continuous mapping theorem. By the continuous mapping theorem we also have the convergence of $|\delta|_{\mathcal{H}^K}\left|\hat{\delta}\right|_{\mathcal{H}^K} \to |\delta|_{\mathcal{H}^K}|G_X|_{\mathcal{H}^K}$ in distribution. We need to verify that the $|\cdot|_{\mathcal{H}^K}$ norm of the minimizer does not "escape" to infinity (i.e. we need to establish tightness). Let $\delta_B \in \mathcal{C}_0$ be such that $|\delta_B|_{\mathcal{H}^K} = B$ for arbitrary, but fixed and finite constant B. Note

that

$$\begin{aligned} n\hat{Q}\left(\delta_B, \hat{\delta}\right) &= -\int_{\mathcal{T}} \delta_B(t)' \hat{\delta}(t)\, dt + \frac{1}{2}\int_{\mathcal{T}} \delta_B(t)' \delta_B(t)\, dt \\ &\geq -|\delta_B|_{\mathcal{H}^K} \left|\hat{\delta}\right|_{\mathcal{H}^K} + \frac{1}{2}|\delta_B|^2_{\mathcal{H}^K} = -B\left|\hat{\delta}\right|_{\mathcal{H}^K} + \frac{B^2}{2}. \end{aligned} \tag{A.5}$$

Let $\underline{c}$ be a constant to be defined in due course. By the Portmanteau theorem for convergence of measures,

$$\liminf_n \Pr\left(-B\left|\hat{\delta}\right|_{\mathcal{H}^K} + \frac{B^2}{2} > \underline{c}\right) \geq \Pr\left(-B|G_X|_{\mathcal{H}^K} + \frac{B^2}{2} > \underline{c}\right).$$

By tightness of the Gaussian process G_X, for any $\tau > 0$, there is a finite absolute constant c_τ such that $\Pr\left(|G_X|_{\mathcal{H}^K} \leq c_\tau\right) \geq 1-\tau$, . Choose $\underline{c} = -Bc_\tau + 2^{-1}B^2$. Deduce that asymptotically, with probability at least $1-\tau$, $n\hat{Q}\left(\delta_B, \hat{\delta}\right) \geq -Bc_\tau + 2^{-1}B^2$. This quantity is eventually increasing in B. Hence, for any $\tau > 0$, there is a finite B_τ such that with probability at least $1-\tau$, the minimizer of $n\hat{Q}\left(\delta, \hat{\delta}\right)$ w.r.t. $\delta \in \mathcal{C}_0$ needs to have $|\cdot|_{\mathcal{H}^K}$ norm less than B_τ. This means that $\tilde{\delta} := \sqrt{n}\left(\tilde{b} - b_0\right)$ is tight and does not "escape" to infinity. In consequence, by the continuous mapping theorem,

$$\tilde{\delta} := \arg\inf_{\delta \in \mathcal{C}_0} n\hat{Q}\left(\delta, \hat{\delta}\right) \to \arg\inf_{\delta \in \mathcal{C}_0}\left(-\int_{\mathcal{T}} \delta(t)' G_X(t) + \frac{1}{2}\int_{\mathcal{T}} \delta(t)' \delta(t)\, dt\right)$$

weakly, as stated in the theorem. ■

Proof. [Theorem 3] Minimization of

$$\hat{Q}(b) := -\int_{\mathcal{T}} b(t)' \hat{C}_{XY}(t,t)\, dt + \frac{1}{2}\int_{\mathcal{T}} b(t)' \hat{C}_{XX}(t,t)\, b(t)\, dt$$

is equivalent to minimization of $\hat{Q}(b) - \hat{Q}(b_0)$ so that, after basic algebra using (1),

$$\begin{aligned} & n\left(\hat{Q}(b) - \hat{Q}(b_0)\right) \\ = \; & -2\int_{\mathcal{T}} \sqrt{n}\left(b(t) - b_0(t)\right)'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i(t)\,\varepsilon_i(t)\right) dt \\ & + \int_{\mathcal{T}} \sqrt{n}\left(b(t) - b_0(t)\right)' \hat{C}_{XX}(t,t)\,\sqrt{n}\left(b(t) - b_0(t)\right) dt \end{aligned} \tag{A.6}$$

By Lemma 5, $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\varepsilon_i$ converges weakly to the Gaussian process G. We then

proceed as in the proof of Theorem 2. Mutatis mutandis, we only need to find a bound similar to (A.5). By Lemma 3,

$$\int_{\mathcal{T}} \delta(t)' \hat{C}_{XX}(t,t) \delta(t) dt \geq (\underline{\lambda} - o_p(1)) \int_{\mathcal{T}} \delta(t)' \delta(t) dt. \tag{A.7}$$

Hence, we now have all the ingredients to find a bound similar to (A.5) and proceed as in the proof of Theorem 2. ■

Proof. [Theorem 4]Under the conditions of the theorem,

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} D_i = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (D_i - \mathbb{E}_{i-1}[D_i|B]) + o_p(1). \tag{A.8}$$

If we show that the first term on the r.h.s. converges to a normal random variable, then the theorem holds. Given that we are conditioning on B, we can redefine the sequence $\{(X_i, \varepsilon_i) : i = 1, 2, 3, ...\}$ on a filtered probability space $\left(B, \mathcal{G}, (\mathcal{G}_i)_{i\geq 0}, P\right)$ so that the sample space is now B. The filtration is such that $\mathcal{G}_i = \{A \cap B : A \in \mathcal{F}_i\}$, where $\mathcal{F}_i$, as defined in the text, is the data history until day i. Letting $\mathbb{E}^P$ be expectation in this new probability space, it follows that $\mathbb{E}_{i-1}[D_i|B] = \mathbb{E}^P[D_i|\mathcal{G}_{i-1}]$ and we shall write $\mathbb{E}^P_{i-1}$ for $\mathbb{E}^P[\cdot|\mathcal{G}_{i-1}]$. Note that (D_i) is not stationary because it depends on the estimated coefficients. The convergence in distribution of the first term on the r.h.s. of (A.8) will follow by an application of Theorem 2.3 in McLeish (1974). That result requires that (i) $\left|\frac{1}{m}\sum_{i=1}^{m}\left(1 - \mathbb{E}^P_{i-1}\right) D_i^2\right| \to c > 0$ in probability, (ii) $\lim_{m\to\infty} \mathbb{E}^P\left[\max_{1\leq i\leq m}\left[\left(1 - \mathbb{E}^P_{i-1}\right) D_i\right]^2 / m\right] < \infty$ and (iii) $\max_{1\leq i\leq m}\left|\left(1 - \mathbb{E}^P_{i-1}\right) D_i/\sqrt{m}\right| \to 0$ in probability.

The condition of the theorem implies that there is a $t_0 \in \mathcal{T}$, such that $\max_{k\leq K} \mathbb{E}^P\left|X^{(k)}(t_0)\right|^{4\eta} < \infty$, for some $\eta > 1$. The Holder continuity from Condition 1 means that $\left|X_i^{(k)}(t)\right| \leq \left|X_i^{(k)}(t_0)\right| + \left|X_i^{(k)}(t) - X_i^{(k)}(t_0)\right| \leq \left|X_i^{(k)}(t_0)\right| + \kappa|\mathcal{T}|$, where $|\mathcal{T}|$ is the Lebesgue measure of $\mathcal{T}$. To ease notation, define $\xi_i := \max_{k\leq K}\left|X_i^{(k)}(t_0)\right| + \kappa|\mathcal{T}|$ and note that $\mathbb{E}^P|\xi_i|^{4\eta} < \infty$ because $\mathbb{E}^P|\xi_i|^{4\eta} \lesssim \sum_{k=1}^{K} \mathbb{E}^P\left|X_i^{(k)}(t_0)\right|^{4\eta} + |\kappa|\mathcal{T}||^{4\eta} < \infty$. This will be used with no further mention.

Point (i) is satisfied by the conditions of the theorem. Points (ii) and (iii) are implied by

$$\mathbb{E}^P\left[\frac{1}{m}\sum_{i=1}^{m}\left[\left(1 - \mathbb{E}_{i-1}\right) D_i\right]^{2\eta}\right] < \infty$$

in probability. Rewrite

$$\begin{aligned} D_i &= -2\int_{\mathcal{T}} Y_i(t) X_i(t)' \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right) dt + \int_{\mathcal{T}} \tilde{b}_{est}(t)' X_i(t) X_i(t)' \tilde{b}_{est}(t) dt \\ &\quad - \int_{\mathcal{T}} \hat{b}_{est}(t)' X_i(t) X_i(t)' \hat{b}_{est}(t) dt. \end{aligned}$$

Use (1) and add and subtract $2\int_{\mathcal{T}} \tilde{b}_{est}(t)' X_i(t) X_i(t)' \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right) dt$ to find that the above is equal to

$$\begin{aligned} &-2\int_{\mathcal{T}} \varepsilon_i(t) X_i(t)' \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right) dt \\ &-2\int_{\mathcal{T}} \left(b_0(t) - \tilde{b}_{est}(t)\right)' X_i(t) X_i(t)' \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right) dt \\ &-\int_{\mathcal{T}} \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right)' X_i(t) X_i(t)' \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right) dt. \end{aligned}$$

Denote the three terms by $U_{1,i}$, $U_{2,i}$, and $U_{3,i}$ respectively. Note that

$$\begin{aligned} \mathbb{E}^P U_{3,i}^{2\eta} &= \mathbb{E}^P \left[\int_{\mathcal{T}} \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right)' X_i(t) X_i(t)' \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right) dt\right]^{2\eta} \\ &\leq \left[\int_{\mathcal{T}} \left|\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right|_2^2 dt\right]^{2\eta} \mathbb{E}^P |\xi_i|^{4\eta} \leq \left|\tilde{b}_{est} - \hat{b}_{est}\right|_{\mathcal{H}^K}^{4\eta} \mathbb{E}^P |\xi_i|^{4\eta}. \end{aligned}$$

By similar arguments,

$$\mathbb{E}^P U_{2,i}^{2\eta} \lesssim \left|b_0 - \hat{b}_{est}\right|_{\mathcal{H}^K}^{2\eta} \left|\tilde{b}_{est} - \hat{b}_{est}\right|_{\mathcal{H}^K}^{2\eta} \mathbb{E}^P |\xi_i|^{4\eta}.$$

Moreover,

$$\begin{aligned} \mathbb{E}^P U_{1,i}^2 &= \mathbb{E}^P \left[\int_{\mathcal{T}} \varepsilon_i(t) X_i(t)' \left(\tilde{b}_{est}(t) - \hat{b}_{est}(t)\right) dt\right]^{2\eta} \\ &\leq \left|\tilde{b}_{est} - \hat{b}_{est}\right|_{\mathcal{H}^K}^{2\eta} \mathbb{E}^P |\varepsilon_i X_i|_{\mathcal{H}^K}^{2\eta}. \end{aligned}$$

Hence,

$$\mathbb{E}^P \frac{1}{m}\sum_{i=1}^{m}\left[\left(1-\mathbb{E}_{i-1}\right)D_i\right]^{2\eta} \lesssim \left|\tilde{b}_{est}-\hat{b}_{est}\right|_{\mathcal{H}^K}^{2\eta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}^P\left|\varepsilon_i X_i\right|_{\mathcal{H}^K}^{2\eta} + \left|\tilde{b}_{est}-\hat{b}_{est}\right|_{\mathcal{H}^K}^{2\eta}\left[\left|\tilde{b}_{est}-\hat{b}_{est}\right|_{\mathcal{H}^K}^{2\eta}+\left|b_0-\hat{b}_{est}\right|_{\mathcal{H}^K}^{2\eta}\right]\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}^P\left|\xi_i\right|^{4\eta},$$

which is finite by the conditions of the theorem. Note that $b_0 \in \mathcal{H}^K$ so that by the triangle inequality, on B,

$$\left|\hat{b}_{est}-b_0\right|_{\mathcal{H}^K} \leq \left|\tilde{b}_{est}-\hat{b}_{est}\right|_{\mathcal{H}^K}+\left|b_0\right|_{\mathcal{H}^K}<\infty.$$

This establishes Points (ii) and (iii) and concludes the proof. ■

Proof. [Lemma 1] Using the same arguments as in the proof of Lemma 2

$$\int_{\mathcal{T}}\int_{\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(s)\,\varepsilon_i(t)\,X_i(s)\,X_i(t)'-C_\sigma(s,t)\right|_F^2 dsdt \to 0$$

in probability. Hence, it is sufficient to bound

$$\begin{aligned}&\sqrt{\int_{\mathcal{T}}\int_{\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i(s)\,\hat{\varepsilon}_i(t)-\varepsilon_i(s)\,\varepsilon_i(t)\right)X_i(s)\,X_i(t)'\right|_F^2 dsdt}\\ &\leq \int_{\mathcal{T}}\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\varepsilon}_i^2(t)-\varepsilon_i^2(t)\right|\left|X_i(t)\right|_2^2 dt, \qquad (A.9)\end{aligned}$$

where the r.h.s. is the nuclear norm, which is stronger than the Hilbert-Schmidt norm (Bosq, 2000). The display can be rewritten as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i(t)-\varepsilon_i(t)\right)\hat{\varepsilon}_i(t)\left|X_i(t)\right|_2^2+\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i(t)-\varepsilon_i(t)\right)\varepsilon_i(t)\left|X_i(t)\right|_2^2\\ =\;&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i(t)-\varepsilon_i(t)\right)^2\left|X_i(t)\right|_2^2+\frac{2}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i(t)-\varepsilon_i(t)\right)\varepsilon_i(t)\left|X_i(t)\right|_2^2 \qquad (A.10)\end{aligned}$$

using $\hat{\varepsilon}_i^2(t)-\varepsilon_i^2(t)=\left(\hat{\varepsilon}_i(t)-\varepsilon_i(t)\right)\left(\hat{\varepsilon}_i(t)+\varepsilon_i(t)\right)$ in the first equality and adding and

subtracting $\varepsilon_i(t)$ in the second equality. Note that

$$\begin{aligned}|\hat{\varepsilon}_i(t) - \varepsilon_i(t)| &= \left| X_i(t)' \left[\hat{C}_{XX}(t,t)\right]^{-1} \frac{1}{n}\sum_{j=1}^n X_j(t)\varepsilon_j(t) \right| \\ &\leq |X_i(t)|_2 \left| \left[\hat{C}_{XX}(t,t)\right]^{-1} \right|_F \left| \frac{1}{n}\sum_{j=1}^n X_j(t)\varepsilon_j(t) \right|_2,\end{aligned}$$

where the equality follows by standard properties of OLS, and the inequality follows by the properties of the Frobenius norm. For any symmetric matrix A, $|A|_F^2$ is the sum of the square eigenvalues. Given that the minimum eigenvalue of $\hat{C}_{XX}(t,t)$ is bounded below by $\underline{\lambda} - o_p(1)$, we have that $\sup_{t\in\mathcal{T}} \left| \left[\hat{C}_{XX}(t,t)\right]^{-1} \right|_F \leq \sqrt{K}/(\underline{\lambda} - o_p(1))$ by Lemma 2. This is finite with probability going to one because K is fixed. From these remarks, we deduce that, with probability going to one,

$$|\hat{\varepsilon}_i(t) - \varepsilon_i(t)| \lesssim |X_i(t)|_2 \left| \frac{1}{n}\sum_{j=1}^n X_j(t)\varepsilon_j(t) \right|_2.$$

Substituting this equality in (A.10), we find that (A.9) is bounded above, with probability going to one, by a constant multiple of

$$\frac{1}{n}\sum_{i=1}^n |X_i(t)|_2^4 \left| \frac{1}{n}\sum_{j=1}^n X_j(t)\varepsilon_j(t) \right|_2^2 + \frac{1}{n}\sum_{i=1}^n |X_i(t)|_2^3 |\varepsilon_i(t)| \left| \frac{1}{n}\sum_{j=1}^n X_j(t)\varepsilon_j(t) \right|_2.$$

By the same arguments as in the proof of Theorem 4, use the bound $\left|X_i^{(k)}(s)\right| \leq \xi_i$ with ξ_i as defined in that proof. Then the above is further bounded above by

$$\frac{K^2}{n}\sum_{i=1}^n \xi_i^4 \left| \frac{1}{n}\sum_{j=1}^n X_j(t)\varepsilon_j(t) \right|_2^2 + \frac{K^{3/2}}{n}\sum_{i=1}^n \xi_i^3 |\varepsilon_i(t)| \left| \frac{1}{n}\sum_{j=1}^n X_j(t)\varepsilon_j(t) \right|_2.$$

Integrating w.r.t. $t \in \mathcal{T}$ and using Holder inequality in the last term, deduce the upper bound

$$\frac{K^2}{n}\sum_{i=1}^n \xi_i^4 \left| \frac{1}{n}\sum_{j=1}^n X_j\varepsilon_j \right|_{\mathcal{H}^K}^2 + \frac{K^{3/2}}{n}\sum_{i=1}^n \xi_i^3 |\varepsilon_i|_{\mathcal{H}} \left| \frac{1}{n}\sum_{j=1}^n X_j\varepsilon_j \right|_{\mathcal{H}^K}. \tag{A.11}$$

The expectation of the right hand most term goes to zero:

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}X_i\varepsilon_i\right|_{\mathcal{H}^K}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left|X_i\varepsilon_i\right|_{\mathcal{H}^K}^2 \leq \frac{1}{n}\max_{i\leq n}\mathbb{E}\left|X_i\varepsilon_i\right|_{\mathcal{H}^K}^2 \to 0$$

because $\max_{i\leq n}\mathbb{E}\left|X_i\varepsilon_i\right|_{\mathcal{H}^K}^2 < \infty$ by assumption. Moreover, by Holder inequality, $\mathbb{E}\xi_i^3\left|\varepsilon_i\right|_{\mathcal{H}} \leq \left(\mathbb{E}\xi_i^4\right)^{3/4}\left(\mathbb{E}\left|\varepsilon_i\right|_{\mathcal{H}}^4\right)^{1/4} < \infty$ because by assumption, $\mathbb{E}\xi_i^4 < \infty$ and $\mathbb{E}\left|\varepsilon_i\right|_{\mathcal{H}}^4 < \infty$. Hence (A.11) goes to zero in probability implying that (A.9) goes to zero in probability. This proves the lemma. ■

A.2 Simulation Using Empirical Data

We consider the volume data from 6E and compute the sample mean volume $\bar{V}(t) = \frac{1}{n}\sum_{i=1}^{n}V_i(t), t \in \mathcal{T}_N$. We observed that the sample variance of $\left\{V_i(t)/\bar{V}(t) : i = 1, 2, ..., n\right\}$ tends to be nearly constant across $t \in \mathcal{T}_N$ except for the first few values of $t \in \mathcal{T}_N$. Hence, we simulate

$$\hat{V}_i(t_j) = \sum_{s=1}^{j}\exp\left\{\eta_{s+(i-1)N}\right\}\left(\bar{V}(t_s) - \bar{V}(t_{s-1})\right),$$

where $\bar{V}(t_0) := 0$, $\eta_j = \omega + \alpha\eta_{j-1} + \sigma_u u_j$ and $(u_j)_{j\in\mathbb{Z}}$ is a sequence of i.i.d. standard normal random variables, while ω, α and σ_u are parameters. Hence, $(\eta_j)_{j\in\mathbb{Z}}$ is an autoregresssive process of order one (AR(1)). In particular, we choose $\omega = -(1-\alpha)\sigma_u^2/2$ and $\sigma_u^2/(1-\alpha^2) \in \{0.15, 0.5\}$ with $\alpha \in \{.95, .99\}$. Using the properties of a Gaussian AR(1), $\exp\left\{\eta_{s+(i-1)N}\right\}$ is lognormal with mean zero and variance equal to $\exp\left\{\sigma_u^2/(1-\alpha^2)\right\} - 1 \simeq \sigma_u^2/(1-\alpha^2)$. This method allows us to simulate one-minute volumes that are autocorrelated with intraday seasonality similar to the original 6E data. We generate a population of $n_{pop} = 10^5$ observations. We sample without replacement from this population in our simulations. In the simulation, we regard as true parameters, the OLS estimate of the model in (18) from the population of size n_{pop}.