A 2-step functional principal component analysis of intra-day volatility trajectories

Jacopo Staccioli†

Matteo Barigozzi‡

†Institute of Economics, Scuola Superiore Sant’Anna, Pisa
‡Department of Statistics, London School of Economics and Political Science, London, UK

IDPE Workshop 2018
Pisa, 2nd October 2018
Outline

1. Context and motivation
2. Model
3. Empirical analysis
4. Prediction performance
5. Concluding remarks
Context and motivation

**volatility**
captures the degree of variation of a trading price series over time

volatility is a top concern in the decision making of any financial actor

risk managers: purchase a financial instrument
banks: issue/price a new derivative
policymakers: maintain orderly operations within the financial system

volatility is also a commodity that can be priced and exchanged as such

- VIX, IVX, etc…
the value attached to understanding the emergence and evolution of volatility over time is ‘tautologically’ justified by the enormous amount of wealth at stake

**our contribution**

we study intra-day volatility trajectories from two major financial markets along the lines of recent developments in functional data analysis (FDA)
Context and motivation

- Volatility is not an observable phenomenon and as such it cannot be directly inferred from discretely observed tick data.
- Theoretically, it could be determined from the quadratic variation of the log price process, if this process were to be observed continuously.
- It is generally assumed that the volatility process is functionally linked to some observed state variable, such as prices or returns (Bandi and Phillips, 2003; Florens-Zmirou, 1993; Rennò, 2008; ...).
- These models target a single realisation of the volatility process whereas we aim at modelling repeated intra-day realisations.
functional data analysis

- provides rigorous statistical tools that are naturally adapted to the analysis of collections of smooth continuous curves defined over the same support
- in our framework these curves consist of individual intra-day volatility trajectories observed over multiple trading days

Müller et al. (2011, *Journal of Econometrics*)

- devise a general diffusion model with drift for log-returns
- the model is then discretised and applied to real data on the S&P500 index
- intra-day time series are transformed into continuous functional objects by means of a smoothing algorithm
- resulting curves are analysed by *functional* PCA (FPCA)
our extension (i)

- we generalise their framework, focussed on a single market index, in order to allow for the joint investigation of an arbitrarily large collection of assets, e.g. all the constituents of the underlying index (baseline model)

- this dramatically complicates the analysis since it brings an additional dimension to the collection of curves

- we devise a 2-step procedure based on FPCA
  1. reduces dimensionality across days
  2. reduces dimensionality across assets

- we focus on the relative contribution of the mean and of the first functional principal component in predicting the original volatility trajectories of the individual assets
our extension (ii)

- we fit a CAPM-inspired factor model to the data in order to separate the market component that correlates with the index (Müller et al. (2011)) and the idiosyncratic component residual part.
- (extended model)
- our 2-step procedure allows the analysis of the common idiosyncratic volatility (CIV), i.e. the part of idiosyncratic returns volatilities that is common across assets after CAPM betas are filtered out (Herskovic et al., 2016; Barigozzi and Hallin, 2016)
- we apply our extended theoretical framework to all the constituents of the S&P500 and the Euronext 100 indices

contributions close to ours in spirit include

- Kokoszka et al. (2014), Hays et al. (2012), Aït-Sahalia and Xiu (2018)
Outline

1. Context and motivation
2. Model
3. Empirical analysis
4. Prediction performance
5. Concluding remarks

A 2-step functional principal component analysis of intra-day volatility
Jacopo Staccioli
2nd October 2018
Model of log-returns

Müller et al. (2011) stochastic diffusion model with drift for log-returns

\[ d \log \bar{X}_j(t, \omega) = \bar{\mu}_j(t, \omega) \, dt + \bar{\sigma}_j(t, \omega) \, dW_j(t, \omega) \]

\( \bar{X}_j(t, \omega) \) denotes the intra-day price of the S&P500 index at time \( t \in [0, T] \) on day \( j = 1, \ldots, J \)

\( \omega \in \Omega \) highlight the stochastic nature of the underlying processes

\( \bar{\mu}_j(t, \omega), \bar{\sigma}_j(t, \omega) \) i.i.d. copies of the usual drift and variance processes

\( W_j(t, \omega) \) independent standard Wiener processes

We will drop dependence on \( \omega \) hereafter for notational convenience.
Model of log-returns

for every constituent $i$ of the index we estimate the following factor model

$$ d \log X_{ij}(t) = \beta_i \cdot M_j(t) + e_{ij}(t) $$

$M_j(t) := d \log \bar{X}_j(t)$ is the market return component

$e_{ij}(t)$ is the idiosyncratic return component

we then apply the same stochastic diffusion model with drift

$$ d \log X_{ij}(t) = \beta_i \cdot \left[ \mu_j(t) \, dt + \sigma_j(t) \, dW_j(t) \right] + \mu_{ij}(t) \, dt + \sigma_{ij}(t) \, dW_{ij}(t) $$

market return

idiosyncratic return
as customary in the literature, we assume that market data is observable at short regular time intervals $\Delta \to 0$ and we define

$$X_{ijt}^\Delta = \frac{1}{\sqrt{\Delta}} \log \left( \frac{X_{ij}(t+\Delta)}{X_{ij}(t)} \right)$$

$$W_{ijt}^\Delta = \frac{W_{ij}(t+\Delta) - W_{ij}(t)}{\sqrt{\Delta}}$$

- subscript $t = 1, \ldots, T/\Delta$ denotes regularly $\Delta$-spaced intra-day times

the discretised diffusion model with drift becomes

$$X_{ijt}^\Delta = \frac{\beta_i}{\sqrt{\Delta}} \cdot \left[ \int_t^{t+\Delta} \bar{\mu}_j(v) \, dv + \int_t^{t+\Delta} \bar{\sigma}_j(v) \, dW_j(v) \right]$$

$$+ \frac{1}{\sqrt{\Delta}} \cdot \left[ \int_t^{t+\Delta} \mu_{ij}(v) \, dv + \int_t^{t+\Delta} \sigma_{ij}(v) \, dW_{ij}(v) \right]$$
Discretisation

Müller et al. (2011, Lemma 1): under suitable regularity assumptions\(^1\)

\[ X_{ijt}^\Delta \approx \beta_i \cdot M_{jt}^\Delta + e_{ijt}^\Delta \]

the smooth generic log-volatility process can be approximated as

\[ V_{ijt}^\Delta = \log \left( (X_{ijt}^\Delta)^2 \right) \]  
\[ \overline{V}_{jt}^\Delta = \log \left( (M_{jt}^\Delta)^2 \right) \]  
\[ V_{ijt}^{\Delta*} = \log \left( (e_{ijt}^\Delta)^2 \right) \]

\(-\)

\(^1\)uniform Lipschitz continuity of order 1 and boundedness of the processes \(\mu(t)\) and \(\sigma(t)\), and smoothness and boundedness of derivatives of \(\sigma(t)\)
a continuous measure of the underlying volatility processes can then be approximated by a $\Lambda$th-order Fourier series over the interval $[0, 1]$.

\[
\overline{V}_j(t) \approx \overline{\alpha}_j + \sum_{\lambda=1}^{\Lambda} \overline{\gamma}_{j\lambda} \cdot \cos\left(\frac{2\pi \lambda t}{T}\right) + \sum_{\lambda=1}^{\Lambda} \overline{\delta}_{j\lambda} \cdot \sin\left(\frac{2\pi \lambda t}{T}\right)
\]

\[
V_{ij}^*(t) \approx \alpha_{ij}^* + \sum_{\lambda=1}^{\Lambda} \gamma_{ij\lambda}^* \cdot \cos\left(\frac{2\pi \lambda t}{T}\right) + \sum_{\lambda=1}^{\Lambda} \delta_{ij\lambda}^* \cdot \sin\left(\frac{2\pi \lambda t}{T}\right)
\]

$\alpha$’s, $\gamma$’s and $\delta$’s are chosen to minimise the distance with the original discrete series by penalised least squares.

- we use the fda R package (Ramsay et al., 2009)
- we set $\Lambda = 7$
FPCA amounts to performing the usual PCA eigen-decomposition on the collection of vectors of Fourier coefficients:

- $\{\alpha_j, \gamma_{j1}, \ldots, \gamma_{j\Lambda}, \delta_{j1}, \ldots, \delta_{j\Lambda}\}$ for market volatility
- $\{\alpha^*_{ij}, \gamma^*_{ij1}, \ldots, \gamma^*_{ij\Lambda}, \delta^*_{ij1}, \ldots, \delta^*_{ij\Lambda}\}$ for idiosyncratic volatilities

**Karhunen-Loève theorem**

A centred continuous stochastic process $V(t) : [a, b] \rightarrow \mathbb{R}$ admits the infinite representation

$$V(t) = \sum_{k=1}^{\infty} \xi_k \cdot \phi_k(t)$$

where $\xi_k$ are pairwise uncorrelated random variables and $\phi_k(t)$ are continuous real-valued functions on $[a, b]$ and pairwise orthogonal in $L^2([a, b])$.
we set $K = 1$

baseline model

$$\hat{V}_{ij}(t) = \mu_i(t) + d_{ij} \cdot \phi_i(t)$$

$$\hat{\phi}_i(t) = h_i \cdot \psi(t)$$

$$\hat{\hat{V}}_{ij}(t) = \mu_i(t) + (d_{ij} \cdot h_i) \cdot \psi(t)$$
Extended FPCA

**market volatility**

\[ \hat{V}_j(t) = \bar{\mu}(t) + a_j \cdot \bar{\phi}(t) \]

**idiosyncratic volatility**

\[ \hat{V}_{ij}^*(t) = \mu_i^*(t) + b_{ij} \cdot \phi_i^*(t) \]

\[ \hat{\phi}_i^*(t) = c_i \cdot \psi^*(t) \]

\[ \hat{V}_{ij}^*(t) = \mu_i^*(t) + (b_{ij} \cdot c_i) \cdot \psi^*(t) \]
1. Context and motivation

2. Model

3. Empirical analysis

4. Prediction performance

5. Concluding remarks
Data description

Tick data about trades between 24th May 2017 and 8th December 2017

**S&P500** 1,559,379,085 data points for 505 constituents across 139 days

**Euronext 100** 109,946,543 data points for 100 constituents across 143 days

We clean the data (Gallo and Brownlees, 2006) and exclude:

- non-regular trades
- trades occurred outside the open-market session
- early closing days
- days with temporary market-wide suspension
- constituents that experienced at least one suspension

We construct regular time series of $\Delta = 5$-minute log-returns

| & | \( I \) & \( J \) & \( T/\Delta \) |
|---|---|---|
| **S&P500** & 502 & 137 & 78 |
| **Euronext 100** & 93 & 142 & 102 |
Figure: 5-minute log-returns (left) and intra-day volatility trajectories (right); S&P500 (top) and Euronext 100 (bottom).
Figure: Mean functions $\mu_i(t)$ (left), eigenfunctions $\phi_i(t)$ and principal eigenfunction $\psi(t)$ (centre), and loading coefficients $d_{ij}$ (right); S&P500 (top) and Euronext 100 (bottom).
Figure: Mean functions $\mu(t)$ (left), principal eigenfunction $\phi(t)$ (centre), and predicted volatility trajectories $\hat{V}_j(t)$ (right); S&P500 (top) and EURONEXT 100 (bottom).
**Figure:** Market volatility loading coefficients $a_j$; S&P500 (blue) and Euronext 100 (green).
Idiosyncratic volatility

Figure: Mean functions $\mu_i^*(t)$ (left), eigenfunctions $\phi_i^*(t)$ and principal eigenfunction $\psi^*(t)$ (centre), and loading coefficients $b_{ij}$ (right); S&P500 (top) and Euronext 100 (bottom).
Outline

1 Context and motivation
2 Model
3 Empirical analysis
4 Prediction performance
5 Concluding remarks
we estimate the following linear models for every constituent asset $i$ and for every available trading day $j$

**baseline model**

$$V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \mu_i(t) + \epsilon_{ij}(t)$$

$$V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \mu_i(t) + \nu_{2ij} \cdot \phi_i(t) + \epsilon_{ij}(t)$$

$$V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \mu_i(t) + \nu_{2ij} \cdot \psi(t) + \epsilon_{ij}(t)$$

- from each experiment we obtain $i \times j$ adjusted $R^2$ coefficients (68,774 for the S&P500 and 13,206 for the Euronext 100)
- we compute averages across days and compare the resulting empirical distribution
Prediction performance

**market volatility**

\[
V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \bar{\mu}(t) + \varepsilon_{ij}(t)
\]

\[
V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \bar{\mu}(t) + \nu_{2ij} \cdot \bar{\phi}(t) + \varepsilon_{ij}(t)
\]

**idiosyncratic volatility**

\[
V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \bar{\mu}_i^*(t) + \varepsilon_{ij}(t)
\]

\[
V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \bar{\mu}_i^*(t) + \nu_{2ij} \cdot \bar{\phi}_i^*(t) + \varepsilon_{ij}(t)
\]

\[
V_{ij}(t) = \nu_{0ij} + \nu_{1ij} \cdot \bar{\mu}_i^*(t) + \nu_{2ij} \cdot \bar{\psi}_i^*(t) + \varepsilon_{ij}(t)
\]
Figure: Prediction performance of the baseline (top), market (middle), and idiosyncratic volatility model (bottom); S&P500 (left) and Euronext 100 (right).
we also compare the ECDF of the last regression with that of
Prediction performance

Figure: Prediction performance of the extended model (top) and ECDF of the complete regressions (bottom) featuring either $\psi^*(t)$ (blue) or $\phi_i^*(t)$ (green); S&P500 (left) and EURONEXT 100 (right).
1 Context and motivation

2 Model

3 Empirical analysis

4 Prediction performance

5 Concluding remarks
Recap

- we investigate the intra-day volatility trajectories of the single constituents of two major international financial markets
- the incumbent literature has so far been addressing one security (or market index) at a time
- we extend the theoretical framework and develop the necessary methodology in order to study an arbitrarily large collection of assets in a joint fashion
- we devise a 2-step dimension-reduction procedure based on FPCA that aims at singling out a small set of curves, consisting of mean functions and eigenfunctions, that attempt to incorporate most of the information contained in the original large database of individual intra-day trajectories
- we propose an alternative model based on a CAPM-inspired distinction between market volatility and idiosyncratic volatility
Our findings

- The loss of information linked to the consecutive application of principal component analysis is negligible.
- The *baseline* model outperforms Müller et al. (2011)’s model in predicting individual volatility trajectories.
- The *extended* model outperforms the *baseline* model in predicting individual volatility trajectories.
- The contribution of the common idiosyncratic volatility curve is not negligible.
- Thereof, the mean function alone is not an optimal predictor and that adding the first functional principal component to the equation, even discarding higher orders, brings a substantial improvement.
Conceivable extensions

1. the second step of our 2-step procedure only targets the eigenfunctions applying a similar procedure to the set of mean functions enables the prediction of an *arbitrarily large* set of volatility trajectories using a *constant* number of curves

2. we neglect the possibility of serial dependence in the data does not prevent the consistent estimation of the eigen-decomposition (Hörmann and Kokoszka, 2010) but fails to take into account the potentially very valuable information that is carried by the past values of the functional observations using *dynamic* FPCA (Hörmann et al., 2014) in our 2-step procedure is not straightforward although it would likely improve the prediction performance
Thank you very much!

j.staccioli<at>santannapisa.it
Appendix
Figure: 5-minute log-returns of the S&P500 (left) and EURONEXT 100 (right) indices.
Intra-day volatility trajectories

Figure: Intra-day volatility curves $\bar{V}_j(t)$ of the index for all available trading days $j$; S&P500 (left) and EURONEXT 100 (right).
Figure: Mean functions $\mu_i(t)$ obtained by estimating the baseline model; S&P500 (left) and EURONEXT 100 (right).
Baseline model (eigenfunctions)

Figure: Eigenfunctions $\phi_i(t)$ obtained by estimating the baseline model and their principal eigenfunction $\psi(t)$ (thick curve); S&P500 (left) and EURONEXT 100 (right).
Figure: Heatmap of the loading coefficients $d_{ij}$ within the baseline model; S&P500 (left) and EURONEXT 100 (right). Darker shades denote larger magnitudes.
Figure: Mean function $\bar{\mu}(t)$ obtained by estimating the market volatility model; S&P500 (left) and EURONEXT 100 (right).
Figure: Eigenfunction $\phi(t)$ obtained by estimating the market volatility model; S&P500 (left) and EURONEXT 100 (right).
Figure: Predicted market volatility trajectories $\hat{V}_j(t)$ by the market volatility model; S&P500 (left) and EURONEXT 100 (right).
Figure: Mean functions $\mu_i(t)$ obtained by estimating the idiosyncratic volatility model; S&P500 (left) and EURONEXT 100 (right).
Idiosyncratic volatility (eigenfunctions)

**Figure:** Eigenfunctions $\phi_i^*(t)$ obtained by estimating the idiosyncratic volatility model and their principal eigenfunction $\psi^*(t)$ (thick curve); S&P500 (left) and EURONEXT 100 (right).
Figure: Heatmap of the loading coefficients $b_{ij}$ within the idiosyncratic volatility model; S&P500 (left) and EURONEXT 100 (right). Darker shades denote larger magnitudes.