The Beta-Adjusted Covariance estimator

Kris Boudt[1,3], Kirill Dragun[1], Steven Vanduffel[1] and Tim Verdonck[2]

[1]Department of Business, Vrije Universiteit Brussel, Belgium
[2]Department of Mathematics, Katholieke Universiteit Leuven, Belgium
[3]School of Business and Economics, Vrije Universiteit Amsterdam, The Netherlands

February 28, 2018

The prices of the traded assets that are comprised in an ETF do not update simultaneously and this asynchronicity creates difficulties when estimating covariances among assets. We propose the Beta Adjusted Covariance estimator as an improvement to a traditional realized covariance estimator by exploiting the information in the realized stock-ETF beta (i.e., the covariance between the ETF's components and the ETF itself) estimated at the highest frequency possible. We find that the proposed estimator efficiently deals with biased approximations by traditional estimators caused by asynchronous trading data and significantly improves accuracy of the estimated covariances.

Contents

1. Introduction

Exchange traded funds keep on attracting record high inflows of capital. In 2017 cumulative amount of ETF has exceeded USD 4.6 trillion [6]. Also growth of the ETFs contributed to the rise of high-frequency trading via ETF arbitrage with for instance daily creations and redemptions amounting to only a small (about 10%) proportion of the ETF transactions and overwhelming majority of trading happening at the secondary market.

We propose the Beta Adjusted Covariance estimator in that we exploit the information from the beta's of the assets comprised in an ETF, i.e., we consider the covariances between the asset and the ETF. Doing so improves precision and robustness of the estimate. There is a great amount of literature on the realized covariance estimation. However the idea of exploiting the information on covariances and correlations of the assets as components contained in the pricing data of the exchange traded funds (ETF) was not examined to a large extent. The precision of covariance estimation depends on returns being synchronised, which is not always the case especially at high-frequency trading. Asynchronous trading causes some estimators to increase estimation error due to a number of reasons, for instance Epps effect [5] when the cross-correlations of high frequency data are significantly smaller than their asymptotic value as observed on longer intervals. Such problems have been investigated in [2], [7] and [3], among others.

The proposed approach based on betas and covariances between individual assets and ETF or an index (market index for instance) allows much more precise estimation of cross-covariances adjusted for individual variances/size of approximation error and trading frequencies. Betas can be estimated at higher precision as a factor in a regression model (such as market portfolio index) are typically very frequently traded and allows exact synchronization. Improved approximation is found by projecting initial estimate to the subset of matrices satisfying more precise constraints defined using high-precision betas. The adjustment for positive semidefiniteness and further estimate improvement are reached using alternating projections and Dykstra's correction [4]. The instruments for fine-tuning of the estimator correction depending on the component volatilities and trading frequencies are also proposed.

The simulation study run for stochastic and constant volatilities and different number of assets confirmed significant reduction in MSE as predicted by theoretical model. The empirical application demonstrated increased robustness of BAC to reduction in frequency of available data comparing to the pairwise estimator. Section 2 of the present paper defines the theoretical setup. In Section 3 we propose our estimation approach. In Sections 4 and 5 we define BAC estimator and discuss some of its properties. Sections 6 and 7 provide a simulation study and an empirical application that show the usefulness of our approach. The most proofs are relegated to an appendix which also discusses some properties, empirical data, etc.

2. Notation, model and some properties

2.1. The securities' prices of interest

We assume p assets with logprices given by the p-dimensional vector $\mathbf{X_t}$ defined as Brownian semimartingale on the probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathcal{P})$,

$$\mathbf{X_t} = \int_0^t \boldsymbol{\mu}_s \,\mathrm{d}s + \int_0^t \boldsymbol{\sigma}_s \,\mathrm{d}\boldsymbol{W}_s, \tag{1}$$

where $\boldsymbol{\sigma}_t \coloneqq \boldsymbol{\sigma}(t)$ and $\boldsymbol{\mu}_t \coloneqq \boldsymbol{\mu}(t)$ are bounded variation vector functions, $\boldsymbol{\mu}_t$ is $p \times 1$-sized and $\boldsymbol{\sigma}_t$ is $p \times m$-sized with $m \geq p$. Furthermore $\boldsymbol{\sigma}_t^{(i)}$ is i-th row of $\boldsymbol{\sigma}_t$ and $\boldsymbol{W}_t$ is m-dimensional vector of independent standard Brownian motions. $\boldsymbol{\Sigma}$ is the integrated covariance matrix:

$$\boldsymbol{\Sigma} = \int_0^1 \boldsymbol{\sigma}_s \boldsymbol{\sigma}_s' \mathrm{d}s. \tag{2}$$

Let $X_t^{(i)}$ be a logprice for ith assets at the moment t with $t_1^{(j)}, t_2^{(j)}, ...t_i^{(j)}, ..., t_{N_j}^{(j)}$ - time points when price of the jth asset is available, N_j is a quantity of the time periods for the asset j. We assume to have initial approximation of the covariance matrix, which we denote $\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}$ or we build it using the approach described in subsection 3.2.

2.2. The ETF

Let Y_t be a logprice of the ETF invested in $1 \ldots p$ assets with given amounts $a_1, \ldots, a_p$ of the components per share of the ETF (the number of shares of the components divided by the number of outstanding shares of the ETF) and $t_1, t_2, ...t_i, ..., t_N$ - time points when price of the ETF is available, N_Y is a quantity of the time periods for the ETF. Pricing information of the ETF is available at reasonably high frequencies. As the ETF itself is traded asset the logprice of the index is also a Brownian semimartingale process:

$$Y_t = \int_0^t \mu_s^{(Y)} \,\mathrm{d}s + \int_0^t \left[\boldsymbol{\sigma}_s^{(Y)}\right]' \mathrm{d}\boldsymbol{W}_s, \tag{3}$$

where $\boldsymbol{\sigma}_s^{(Y)}$ is m-dimensional vector.
We assume no-arbitrage environment, so the value of the ETF is equal to weighted sum of

its components ($a_t^{(i)}$ - amount of the asset i at time t)

$$Y_t = \log\left(\sum_{i=1}^{p} a_i \exp(X_t)\right) \tag{4}$$

Arithmetic return of the index at time t is equal to the weighted sum of arithmetic returns of its components:

$$\exp(\mathrm{d}Y_t) - 1 = \sum_{i=1}^{p} w_t^{(i)} \left[\exp(\mathrm{d}X_t^{(i)}) - 1\right], \tag{5}$$

where

$$w_t^{(i)} = \frac{a_i \exp(X_t^{(i)})}{\exp(Y_t)}$$

We denote expected weights of returns as $\bar{w}$:

$$\bar{w}_j = \mathbb{E}w_t^{(j)}$$

For modelling purposes we assume trading to occur on fixed window of time $[0,1]$, $t_i^{(i)} \in [0,1]$. Number of time points with pricing information available is different across the assets. Some are liquid and others are illiquid. ETF/index is liquid. We assume that $N_Y \geq N_1 \geq N_2 \geq ... \geq N_i \geq ... \geq N_p$.

2.3. β representation

Comparably lower frequency of the pricing information of the illiquid assets and asynchronous trading creates difficulties of estimating covariances and contributes to the size of errors. The proposed approach is to extract maximum information from the covariances with the asset which price is mostly available and which has high frequency of the trades - the ETF itself. We assume that we are given $\hat{\mathbf{\Sigma}}_\mathbf{0}$ approximation of the integrated covariance matrix $\mathbf{\Sigma}$ (for instance produced by refresh time pair-wise realised covariance estimator). Now we want to express linear relations between elements of rows(or columns) of the covariance matrix and link that to β parameters which we introduce below. We will prove that spot covariance of logreturns of the index and its components can be represented as a weighted sum of spot cross-covariances of the components. This is true for arithmetic returns (5).

$$Z_t^{(i)} = \exp(X_t^{(i)})$$

We use Ito's formula:

$$\mathrm{d}Z_t^{(i)} = Z_t^{(i)}\mathrm{d}X_t^{(i)} + Z_t^{(i)}\frac{1}{2}[\mathrm{d}X_t^{(i)}]$$

$$\frac{\mathrm{d}Z_t^{(i)}}{Z_t^{(i)}} = \mathrm{d}X_t^{(i)} + \frac{1}{2}[\mathrm{d}X_t^{(i)}] = \left(\mu_t^{(i)} + \frac{1}{2}\left[\boldsymbol{\sigma}_t^{(i)2}\right]'\right)\mathrm{d}u + \left[\boldsymbol{\sigma}_t^{(i)}\right]'\mathrm{d}\boldsymbol{W}_t$$

So instantaneous arithmetic return has the same stochastic term and the same variation as instantaneous logarithmic return.

This works for index as well:

$$Z_t^{(Y)} = \exp(Y_t)$$

$$\frac{\mathrm{d}Z_t^{(Y)}}{Z_t^{(Y)}} = \left(\mu_t^{(Y)} + \frac{1}{2}\left[\boldsymbol{\sigma}_t^{(Y)2}\right]'\right)\mathrm{d}u + \left[\boldsymbol{\sigma}_t^{(Y)}\right]'\mathrm{d}\boldsymbol{W}_t \tag{6}$$

From (5):

$$\frac{\mathrm{d}Z_t^{(Y)}}{Z_t^{(Y)}} = \sum_{j=1}^{p} w_t^{(j)}\left(\left(\mu_t^{(i)} + \frac{1}{2}\left[\boldsymbol{\sigma}_t^{(i)2}\right]'\right)\mathrm{d}u + \left[\boldsymbol{\sigma}_t^{(i)}\right]'\mathrm{d}\boldsymbol{W}_t\right) \tag{7}$$

Therefore comparing (6) and (7):

$$\boldsymbol{\sigma}_t^{(Y)} = \sum_{j=1}^{p} w_t^{(j)}\boldsymbol{\sigma}_t^{(i)}$$

It is easy to see that β_i, which we define as integrated covariance between ETF returns and i-th asset returns is equal:

$$\beta_i := \int_0^1 \left[\boldsymbol{\sigma}_t^{(Y)}\right]'\boldsymbol{\sigma}_t^{(i)}dt = \sum_{j=1}^{p}\int_0^1 w_t^{(j)}\left[\boldsymbol{\sigma}_t^{(j)}\right]'\boldsymbol{\sigma}_t^{(i)}dt$$

This immediately follows from linearity of covariance.

To express relation between $\int_0^1 w_t^{(j)}\boldsymbol{\sigma}_t^{(j)\prime}\boldsymbol{\sigma}_t^{(i)}dt$ and $\int_0^1 \boldsymbol{\sigma}_t^{(j)\prime}\boldsymbol{\sigma}_t^{(i)}dt$, let's consider integral from 0 to 1 of spot covariance matrix as expectation of the integrand if t is uniformly distributed in the interval [0;1]

$$\beta_i := \int_0^1 \boldsymbol{\sigma}_t^{(Y)\prime}\boldsymbol{\sigma}_t^{(i)}\mathrm{d}t = \sum_{j=1}^{p}\int_0^1 w_t^{(j)}\boldsymbol{\sigma}_t^{(j)}\boldsymbol{\sigma}_t^{(i)\prime}\mathrm{d}t = \sum_{j=1}^{p}\int_0^1 \bar{w}^{(j)}\boldsymbol{\sigma}_t^{(j)\prime}\boldsymbol{\sigma}_t^{(i)}\mathrm{d}t + u_i$$

$$u_i = \sum_{j=1}^{p}\int_0^1 (w_t^{(j)} - \bar{w}_j)\boldsymbol{\sigma}_t^{(j)\prime}\boldsymbol{\sigma}_t^{(i)}\mathrm{d}t$$

where $w_t^{(j)} = \frac{X_t^{(j)}}{Y_t}$ as we assume no-arbitrage environment $w_t^{(j)}$ is a martingale, $\boldsymbol{\sigma}^{(j)\prime}\boldsymbol{\sigma}^{(i)}$ is bounded variation function. Therefore $\mathbb{E}(w_t^{(j)} - \bar{w}_j)\boldsymbol{\sigma}^{(j)\prime}\boldsymbol{\sigma}^{(i)} = 0$. Therefore $\boldsymbol{u}$ is an error term having zero expectation:

$$\boldsymbol{u} = \int_0^1 (w_t^{(j)} - \bar{w}_j)\boldsymbol{\sigma}^{(j)\prime}\boldsymbol{\sigma}\mathrm{d}t$$

$$\boldsymbol{\beta} \coloneqq \int_0^1 \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t^{(Y)} dt = \overline{w}' \int_0^1 \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t^{(Y)} dt + \boldsymbol{u} \tag{8}$$

The special case to consider is constant $\boldsymbol{\sigma}_t$.
If we assume constant variance $\sigma_t = \sigma$, then

$$\beta_i \coloneqq \sum_{j=1}^{p} \int_0^1 w_t^{(j)} \left[\boldsymbol{\sigma}^{(j)}\right]' \boldsymbol{\sigma}^{(i)} dt = \sum_{j=1}^{p} \left[\boldsymbol{\sigma}^{(j)}\right]' \boldsymbol{\sigma}^{(i)} \int_0^1 w_t^{(j)} dt = \sum_{j=1}^{p} \bar{w}_j \left[\boldsymbol{\sigma}^{(j)}\right] \boldsymbol{\sigma}^{(i)}$$

So we have p linear equations expressing relation between integrated pair-wise covariances of the index components and integrated covariances of index and its components. β is approximated by realized covariance of the index and its components.

3. Synchronization issues when performing pairwise estimation

We use refresh time approach similar to the one from [2].

3.1. Estimating pairwise covariances

The sets of trading points are joined and ordered:

$$T^i = \{t_1^{(i)}, ..., t_{N_i}^{(i)}\}$$

$$T^j = \{t_1^{(j)}, ..., t_{N_i}^{(j)}\}$$

$$T = T^j \cup T^i$$

$T_1, ...T_m$ are points in T such that each of the intervals in $\{(0, T_1], (T_1, T_2], ...\}$ contains a least one point from each asset trades. Returns of the index are grouped according to the interval between asset trades they belong to.

$$T_i = \max(t \in T \mid t \in (T_{i-1}, T_i])$$

$$\min(|T^Y \cap (T_{i-1}, T_i]|, |T^i \cap (T_{i-1}, T_i]|) = 1$$

We denote as $\hat{\beta}_i$ estimated covariance of the index (as discussed above) and ith asset and u_i as an error term of approximation

$$X_k^{(i)} := X^{(i)}_{max(T^{(i)} \cap (T_{k-1}, T_k])}$$

$$X_k^{(j)} := X^{(j)}_{max(T^{(i)} \cap (T_{k-1}, T_k])}$$

$$\tilde{N} = |X_j|$$

$$Cov(X^{(i)}, X^{(j)}) = \sum_{k=1}^{\tilde{N}} (X_k^{(i)} - X_{k-1}^{(i)})(X_k^{(j)} - X_{k-1}^{(j)}) \tag{9}$$

3.2. Estimating β

Due to high frequency of the ETF returns we are able to estimate its covariance with all p assets with less synchronization difficulties as not overlapping periods will be smaller thus resulting in higher precision.

The sets of trading points are joined and ordered:

$$T^Y = \{t_1^{(Y)}, ..., t_N^{(Y)}\}$$

$$T^i = \{t_1^{(i)}, ..., t_{N_i}^{(i)}\}$$

$$T = T^Y \cup T^i$$

$T_1, ...T_m$ are points in T such that each of the intervals in $\{(0, T_1], (T_1, T_2], ...\}$ contains a least one point from asset trades and one points from ETF trades. Returns of the index are grouped according to the interval between asset trades they belong to.

$$T_i = \max(t \in T \mid t \in (T_{i-1}, T_i])$$

$$\min(|T^Y \cap (T_{i-1}, T_i]|, |T^i \cap (T_{i-1}, T_i]|) = 1$$

We denote as $\hat{\beta}_i$ estimated covariance of the index (as discussed above) and ith asset and u_i as an error term of approximation

$$X_j^{(i)} := X_{max(T^{(i)} \cap (T_{j-1}, T_j])}$$

$$Y_j := Y_{max(T^Y \cap (T_{j-1}, T_j])}$$

$$\tilde{N} = |X_j|$$

$$\hat{\beta}_i = \sum_{j=1}^{\tilde{N}} (X_j^{(i)} - X_{j-1}^{(i)})(Y_j - Y_{j-1}) \tag{10}$$

$$\hat{\beta}_i = \beta_i + u_i$$

4. BAC Estimator

We propose an approach to estimate covariance matrix based on some less-accurate initial approximation produced for instance by some of the other estimators performing non-optimally in the presence of asynchronity. Pairwise covariances of the components are grouped by rows and their row-wise sums are equal to betas as shown above. Using these relations we define a subset of all symmetric matrices of the appropriate dimension and we are looking for the minimal correction (orthogonal projection to the subset) to adjust initial estimator in such a way that it fits set constraints (the set S_β defined in detail in 4.1). Improved estimate may not be positive semidefinite (the closed set of positive semidefinite matrices S_+) and requires further processing. So we need to find the point in the intersection of S_β and S_+, $\hat{\mathbf{\Sigma}}$ such that

$$\underset{\hat{\mathbf{\Sigma}} \in S_\beta \cap S_+}{\operatorname{argmin}} \left\| \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_F .$$

For that purpose we are using Dykstra algorithm [4] a special case of the alternating projection method. In 9.1 and 4.2 we construct projections to S_β and S_+ respectively.

4.1. Nearest correction

We start with some approximation of the covariance matrix given by symmetric matrix $\hat{\Sigma}_0$. Common choices for $\hat{\Sigma}_0$ are synchronized or pair-wise covariance estimators as for example the one used in 3.2 for estimation

Now we adjust estimated values using addition information from differences between covariance estimated above (10) and their analytical representation (8). We estimate covariance matrix

$$\hat{\mathbf{\Sigma}}_{\mathbf{1}} = \hat{\mathbf{\Sigma}}_{\mathbf{0}} + \mathbf{\Delta} \tag{11}$$

$$\mathbf{\Delta} = \begin{bmatrix} 0 & \delta_{12} & \dots & \delta_{1p} \\ \delta_{21} & 0 & \dots & \delta_{2p} \\ \dots & \dots & \dots & \dots \\ \delta_{p1} & \delta_{p2} & \dots & 0 \end{bmatrix},$$

$$\Delta_i = \beta_i - \sum_{j=1, j\neq i}^{p} \bar{w}_j \hat{\Sigma}_{ij}, \tag{12}$$

where $\delta_{ij} = \delta_{ji}$. Then

$$\hat{\mathbf{\Sigma}}_{\mathbf{1}} \in S_\beta \coloneqq \{Z \in R^{p\times p} : Z_{ii} = \sigma_{ii}, Z_{ij} = Z_{ji} \forall i, j = 1 \dots p, Z \times \boldsymbol{\iota}_p = \boldsymbol{\beta}\},$$

where $\boldsymbol{\iota}_p$ is p-sized vector of 1, $\boldsymbol{\iota}_p = (1, ..., 1)'$. So $\hat{\mathbf{\Sigma}}_{\mathbf{1}}$ belongs to the flat(subset of $p \times p$ dimensional Euclidean space congruent to a subspace of that space) formed by intersection of p hyperplanes given by p equations $X \times \mathbf{1} = \boldsymbol{\beta}$ and subspace of symmetric matrices with given diagonal values.
We are looking for such $\hat{\mathbf{\Sigma}}_{\mathbf{1}} \in S_\beta$ that $\left\|\hat{\mathbf{\Sigma}}_{\mathbf{1}} - \hat{\mathbf{\Sigma}}_{\mathbf{0}}\right\|_F$ is minimal, which is equivalent for corrections δ_{ij} such that sum of their squares is minimized given constraints (12):

$$\underset{\hat{\mathbf{\Sigma}}_{\mathbf{1}} \in S_\beta}{\operatorname{argmin}} \left\|\hat{\mathbf{\Sigma}}_{\mathbf{1}} - \hat{\mathbf{\Sigma}}_{\mathbf{0}}\right\|_F = \underset{\hat{\mathbf{\Sigma}}_{\mathbf{1}} \in S_\beta}{\operatorname{argmin}} \sum_{i=1}^{p} \sum_{j=1}^{i-1} \delta_{ij}^2. \tag{13}$$

The solution is obtained using the method of Lagrange multipliers (please see 9.1). $\hat{\mathbf{\Sigma}}_{\mathbf{1}}$ is by construction the orthogonal projection of the initial approximation $\hat{\mathbf{\Sigma}}_{\mathbf{0}}$ on the subspace defined by the given constraints. The distance between $\hat{\mathbf{\Sigma}}_{\mathbf{1}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{0}}$ is an initial approximation error and the distance between $\hat{\mathbf{\Sigma}}_{\mathbf{1}}$ and $\mathbf{\Sigma}$ is corrected approximation error. If given constraints are correct then $\mathbf{\Sigma}$ belongs to the same subspace S and the distance between $\hat{\mathbf{\Sigma}}_{\mathbf{1}}$ and $\mathbf{\Sigma}$ is the orthogonal projection of the distance between $\hat{\mathbf{\Sigma}}_0$ and $\mathbf{\Sigma}$.

$$\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_{\mathbf{1}}\right\|_F^2 = \left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_{\mathbf{0}}\right\|_F^2 - \|\mathbf{\Delta}\|_F^2$$

Therefore estimation error in the beta adjusted covariance estimate is always less than the initial error given correctly estimated $\boldsymbol{\beta}$/constraints.

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}_1\|_F \leq \left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_{\mathbf{0}}\right\|_F$$

4.2. Regularization

The generated corrected matrix $\mathbf{\Sigma}_\beta = P_\beta(\hat{\mathbf{\Sigma}}_0)$ and the original approximation $\hat{\mathbf{\Sigma}}_\beta$ are not necessarily positive semi-definite as we noted above. As final part of each iteration providing positive semidefinite estimate we use positive semi-definite projection of $\hat{\mathbf{\Sigma}}_\beta$ (nearest PSD matrix) and denote it as $\hat{\mathbf{\Sigma}}_+$. We use spectral decomposition $\mathbf{\Sigma}_\beta = \boldsymbol{QEQ'}$, where $\boldsymbol{Q}$ is orthogonal matrix and $\boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I}$, to find $\hat{\boldsymbol{\Sigma}}_+$:

$$\boldsymbol{E}_+ = \operatorname{diag}(\max(\operatorname{diag}(\boldsymbol{E}), 0))$$

$$\hat{\boldsymbol{\Sigma}}_+ = \boldsymbol{Q}\boldsymbol{E}_+\boldsymbol{Q}'$$

where $\hat{\boldsymbol{\Sigma}}_+$ is diagonal matrix consisting of only positive eigenvalues of $\hat{\boldsymbol{\Sigma}}_\beta$.

Property 1 *The resulting matrix $\hat{\boldsymbol{\Sigma}}_+$ is positive semidefinite and it is more precise approximation of $\hat{\boldsymbol{\Sigma}}$ than $\hat{\boldsymbol{\Sigma}}_\beta$*

$$\left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_\beta\right\|_F \geq \left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_+\right\|_F$$

and its becomes strong inequality if $\hat{\boldsymbol{\Sigma}}_\beta$ is not positive semi-definite. For the proof please see

This also improves accuracy of the estimate as distance from $\hat{\boldsymbol{\Sigma}}_+$ to $\boldsymbol{\Sigma}$ is less than $\left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_\beta\right\|_F$. Let's represent $\hat{\Sigma}$ as a sum of positive and negative definite matrices.

$$\hat{\boldsymbol{\Sigma}}_\beta = \hat{\boldsymbol{\Sigma}}_+ + \hat{\boldsymbol{\Sigma}}_-$$

$$\left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_\beta\right\|_F^2 = \left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_+ - \hat{\boldsymbol{\Sigma}}_-\right\|_F^2 = \left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_+\right\|_F^2 + \left\|\hat{\boldsymbol{\Sigma}}_-\right\|_F^2 - < \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_+, \hat{\boldsymbol{\Sigma}}_- >$$

From definition of the inner product in Euclidian space it follows that:

$$< \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_+, \hat{\Sigma}_- >= \operatorname{tr}(\boldsymbol{\Sigma}'\hat{\boldsymbol{\Sigma}}_-) - \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_+, \hat{\boldsymbol{\Sigma}}_-)$$

It is easy to see that $\hat{\boldsymbol{\Sigma}}_+\hat{\boldsymbol{\Sigma}}_- = 0$ as $\hat{\boldsymbol{\Sigma}}_+\hat{\boldsymbol{\Sigma}}_- = \boldsymbol{Q}\boldsymbol{E}_+\boldsymbol{E}_-\boldsymbol{Q}' = 0$. And $\boldsymbol{\Sigma}'\hat{\boldsymbol{\Sigma}}_-$ is a product of positive semidefinite and negative semidefinte matrices and is negative semidefinite, so the sum of its eigenvalues is not positive and so $\operatorname{tr}(\boldsymbol{\Sigma}'\hat{\boldsymbol{\Sigma}}_-) \leq 0$.

$$\left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_\beta\right\|_F^2 \geq \left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_+\right\|_F^2 + \left\|\hat{\boldsymbol{\Sigma}}_-\right\|_F^2 \geq \left\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_+\right\|_F^2$$

4.3. Projection to the intersection of the sets

However now projection to the set of positive semi-definite matrices $\hat{\Sigma}_+$ doesn't necessarily belong to the subspace defined by β-constraints (12). Ideally we have to find projection to the intercession of β-constrained flat and the convex set of positive semidefinite matrices. We do it using iterative projections:

do

$\hat{\Sigma}_i = P_{PSD}(P_\beta(\hat{\Sigma}_{i-1}))$

while $\left\|\hat{\Sigma}_i - \hat{\Sigma}_{i-1}\right\|_F > \epsilon$

Where P_β and P_{PSD} operators of projection as defined in 9.1 and 4.2, $\hat{\Sigma}_0$ is initial approximation and ϵ is precision limit.
The sequence $\hat{\Sigma}_1, ...\hat{\Sigma}_i, ...$ converges to the limit belonging to the intersection of β-subspace and positive semidefinite matrices set

$$\left\|\Sigma - \hat{\Sigma}_i\right\|_F < \left\|\Sigma - \hat{\Sigma}_{i-1}\right\|_F < ... < \left\|\Sigma - \hat{\Sigma}_1\right\|_F .$$

As in 9.1 and 4.2 $\|\Delta\|_F>0$ and $\|\Sigma_-\|_F>0$ if projections do not belong to the intersection of the sets. So we have a bounded and decreasing sequence, therefore the sequence of iterations has a limit

$$\hat{\Sigma} = \lim_{i\to\infty} \hat{\Sigma}_i$$

and it is easy to see it belongs to the intersection of the sets.
We repeat iterations until the improvement of the estimate is below the some limit of precision. However as we project to the intersection with convex set of positive semidefinite matrices the limit of the iterations is not necessarily optimal point (nearest to initial estimate) [4, 8]. We have to use Alternating Projections Method with Dykstra correction for P_{PSD}.

$$\hat{\Sigma}_i = P_{PSD}(P_\beta(\hat{\Sigma}_{i-1}) - C_{i-1})),$$

where C_i is Dykstra's correction given by:

$$C_i = \hat{\Sigma}_i - P_\beta(\hat{\Sigma}_{i-1}) + C_{i-1}, \forall i \in \mathbb{N} \mid i > 0, C_0 = 0$$

4.4. Weighted BAC

While assigning individual corrections naive BAC considers all elements of the initial estimate of the integrated covariance matrix equally significant as a potential source of an error. A more efficient way is to adjust correction accordingly to the size of a potential approximation error of the pairwise covariance depending on the variances of the ETF components and of course on frequency of their trades. More advanced alternative Weighted BAC is minimizing the sum of squares of corrections normalized according to corresponding variances and even adjusted to reflect frequencies of trade.

$$\sum_{i=1}^{p}\left[v_i\frac{\epsilon_i^2}{\sigma_i^2\sigma_Y^2} + \sum_{j=1,j\neq i}^{p}\frac{\delta_{ij}^2}{\sigma_i^2\sigma_j^2}\right],$$

where σ_i and σ_Y are observed standard deviations of the ETF components, such that they reflect initial model σ and frequencies of trades and serve as confidence weights for the

projection. Also we add error factors ϵ_i to make the model more robust to errors in approximations of Δ_i. It relaxes the link between estimated δ_{ij} and Δ_i as the later incorporates estimation error. ϵ should have heavy weight multiplier v_i as Δ_i are estimated with significantly higher precision. We apply similar Lagrange multipliers optimization method as in previous BAC case, please see 9.2.
$\hat{\mathbf{\Sigma}}_{\mathbf{1}} = \hat{\mathbf{\Sigma}}_{\mathbf{0}} + \mathbf{\Delta}$ is the nearest to $\hat{\mathbf{\Sigma}}_{\mathbf{0}}$ point in the flat defined by β-constraints if norm is defined as

$$\|\boldsymbol{A}\|_W = \sqrt{\operatorname{tr}(\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}\boldsymbol{W})}$$

$$\boldsymbol{W} = \operatorname{diag}(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_p})$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{1}} = P_\beta(\hat{\mathbf{\Sigma}}_{\mathbf{0}})$$

We assigned confidence weights to the projection to β-flat. Now we have to do the same with projection to the positive semidefinite matrices convex set. Using the results from [8]

$$P_{PSD}(\boldsymbol{A}) = \boldsymbol{W}^{1\frac{1}{2}} \left(\left(\boldsymbol{W}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{W}^{\frac{1}{2}} \right)_+ \right) \boldsymbol{W}^{-\frac{1}{2}}$$

5. Properties of the estimator

BAC estimator is computed using initial approximation. So its properties will heavily depend on properties of the initial estimator. We have seen above that it improves MSE of the initial estimator.

- **Consistency**
  Given that initial estimator $\hat{\mathbf{\Sigma}}_0$ is consistent, BAC $\hat{\mathbf{\Sigma}}$ is consistent as well.

  $$\lim_{n\to\infty} \left\{ \left\| \hat{\mathbf{\Sigma}}_{\boldsymbol{n}} - \mathbf{\Sigma} \right\|_F < \epsilon \right\} = 1, \forall \epsilon > 0$$

  this follows from

  $$\lim_{n\to\infty} \left\{ \left\| \hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma} \right\|_F < \epsilon \right\} = 1, \forall \epsilon > 0$$

  $$\left\| \hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma} \right\|_F \geq \left\| \hat{\mathbf{\Sigma}}_{\boldsymbol{n}} - \mathbf{\Sigma} \right\|_F$$

  It can be easily demonstrated that consistency is also preserved on per element basic.

- **Asymptotic normality**
  We need to prove that if initial estimator is asymptotically normal

  $$\sqrt{n}(\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}) \xrightarrow{D} \mathbb{N}(0, \mathbf{\Lambda}_0),$$

where $\mathbf{\Lambda}_0$ is $p \times p$-sized vector/matrix of individual variances of the elements of the initial estimate then $\hat{\mathbf{\Sigma}}$ is also asymptotically normal. In [9] the case of differentiability properties of the projection to the intersection of the shifted linear subspace of the space of $p \times p$ symmetric matrices and the cone of positive semidefinite matrices are studied. And such projection is proven to be directionally differentiable and so asymptotic normality property is preserved. In order to assess the variance of the adjusted estimate we can apply Delta method to our projection of the initial estimator

$$\hat{\mathbf{\Sigma}} = P_S(\hat{\mathbf{\Sigma}}_0),$$

$$\sqrt{n}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) \xrightarrow{D} \mathbb{N}(0, \nabla\hat{\mathbf{\Sigma}}^2 \circ \mathbf{\Lambda}),$$

where $\nabla\hat{\mathbf{\Sigma}}$ is gradient of the projection function. As we don't have analytical form for such projection we have to use numerical methods to compute $\nabla\hat{\mathbf{\Sigma}}$.

6. Simulation

We compare performance of the proposed estimator to conventional pair-wise synchronization estimator (PS), where covariance matrix is built by element using covariance of two synchronized series per element.

One day / 7.5 hours of shares trading is simulated. Minimum trading interval for the ETF is set at 50 milliseconds, making up to $7.5 \times 60 \times 60 \times 20 = 540000$ intervals per day.

$$N = 540000$$

$$T^{(Y)} = \{(M-1)/N : M \in \mathbb{Z} \mid M \in [0; N]\}$$

Asset prices are calculated using generated returns for all $t_i \in T^{(Y)}$, index price is calculated as sum of assets. ETF logreturns are built using asset prices

$$Y_t = \log \left(\sum_{j=1}^{p} \exp(X_t^{(j)}) \right).$$

Precise covariance matrix needed for assessment of the accuracy of the estimators is computed from perfectly synchronized simulated returns

$$\boldsymbol{\Sigma} = \sum_{t=1}^{N} (\boldsymbol{X}_{\frac{t}{N}} - \boldsymbol{X}_{\frac{t-1}{N}})(\boldsymbol{X}_{\frac{t}{N}} - \boldsymbol{X}_{\frac{t-1}{N}})'.$$

After that some of the time points are randomly eliminated using provided frequency of trades, such that T_i set of time points for asset i includes only points with trading times generated by sequential addition of generated exponentially distributed random numbers corresponding to the intervals of the observed trades:

$$T^{(i)} = \{t_i : t_i \in T^{(Y)} \mid i \in E\}$$

$$E = \{t_1 \sim \text{Exp}(1), \dots, t_i = t_{i-1} + \Delta t_i \mid \Delta t_i \sim \text{Exp}(\frac{p-i+1}{p})\}$$

Finally logreturns are recalculated using price data for remaining time points.

6.1. Constant volatility and zero drift returns setup

Under constant volatility and constant drift assumption the model (1) reduces to the Geometric Brownian Motion. Logreturns for p assets are simulated using random normal numbers generated with given standard deviations as defined in (1), initial $\sigma_0 = 0.01$ multiplied by uniformly distributed random number $U(0,1)$, mean $\mu = 0$ and pairwise correlations are distributed in the interval $U(0.5,1)$

$$X_t^{(i)} = \sigma_i[W_t + \phi_i Z_t^{(i)}]$$

W_t and $Z_t^{(i)}$ are independent Brownian motions.

$$\phi_i \sim U(0,1)$$

$$\sigma_0 = 0.01$$

$$\sigma_i = \sigma_0 \times U(0,1)$$

The table below shows the error as the norm of the difference between covariance matrices estimated using pair-wise and BAC estimators and precise covariance matrix computed from original high frequency simulated data without artificially created gaps.

Number of assets	Simulations	Pair-wise estimator	BAC estimator	BAC/PWE %
100	20	32.82	11.07	33.73%
200	20	60.18	15.08	25.06%
300	20	93.39	22.26	23.83%

6.2. Stochastic volatility

Now we use the same simulation setup as defined in [2]. Simulation interval is set to $[0,1]$.

$$\mathrm{d}X^{(i)} = \mu^{(i)}\mathrm{d}t + \mathrm{d}V^{(i)} + \mathrm{d}F,$$

$$\mathrm{d}V^{(i)} = \rho^{(i)}\sigma^{(i)}\mathrm{d}B^{(i)},$$

$$\mathrm{d}F^{(i)} = \sqrt{1-(\rho^{(i)})^2}\sigma^{(i)}\mathrm{d}W,$$

where $F^{(i)}$ is common factor,$W \perp\!\!\!\perp B^{(i)}$, $\mu^{(i)}$ is constant drift and $\sigma^{(i)}$ is stochastic volatility

$$\sigma^{(i)} = \exp(\beta_0^{(i)} + \beta_1^{(i)}\varrho^{(i)}),$$

$$\mathrm{d}\varrho^{(i)} = \alpha^{(i)}\varrho^{(i)}\mathrm{d}t + \mathrm{d}B^{(i)}.$$

The parameters are set as in [2] $(\mu^{(i)}, \beta_0^{(i)}, \beta_1^{(i)}, \alpha^{(i)}, \rho^{(i)}) = (0.03, 5/16, 1/8, 1/40, 0.3)$. The ETF price is modelled as weighted sum of the prices of the assets

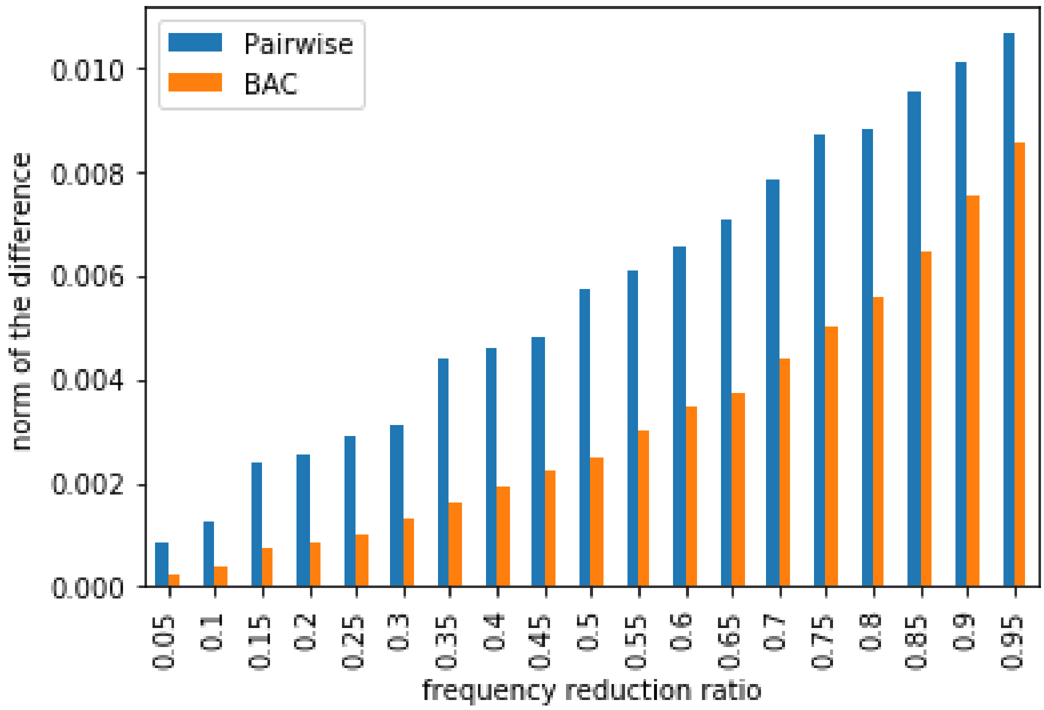
$$Y = \log\left(\sum_{j=1}^{p} \exp(X^{(j)})\right).$$

The availability of prices is adjusted according to the frequency of trading is modelled in a same way as in 6.1.

Number of assets	Simulations	Pair-wise estimator	BAC estimator	BAC/PWE %
100	20	5.84	1.69	28.93%
200	20	11.55	2.43	21.04%
300	20	17.72	3.75	21.16%

BAC estimator is found to be much more (on average 4 times more) accurate than the pair-wise covariance estimator used for initial estimate generation and as a comparable estimator. The results of the simulation confirm our theoretical expectations and prove that there is potential for practical usage of the ETF quotes data for more precise estimation of the components cross-covariance.

Figure 1. Robustness to the data set(frequency of trade) reduction



7. Empirical application

For the real-life data we don't have fully precise covariance matrix to compare with, however we can adjust available frequencies of trading by removing some of the data and analyse the impact of the trading frequency on the quality of the estimate. We use high frequency data for the SPDR Dow Jones Industrial Average ETF (DIA) imitating Dow Jones Industrial Average and trades data for its components for the period 7/01/16-14/01/16 and clean trading data using principles similar to that of [1]. For the initial frequencies please see 9.5. First we estimate covariance matrix using 100% of the available data then at following steps we randomly reduce quantity of the trades at rates $5\%, \ldots, 95\%$ so that at the end only 5% of initial data for the ETF components remains. At each step we also compute the Frobenius norm of the difference (the distance) between initial estimates and their versions based on reduced data set. For the results please see Figure 1.

BAC estimator is very robust to small dilution of the trading data and in general more robust than the pairwise estimator. In the table 7 the norms of the differences between the full data set estimates (most accurate) and the estimated based on reduced data sets (less accurate) both for BAC and the pairwise estimator are provided.

Table 1: The norm of the difference with full data set estimates

Deleted share of time points	Pairwise	BAC
0.05	0.000846	0.000227
0.10	0.001262	0.000383
0.15	0.002409	0.000720
0.20	0.002522	0.000853
0.25	0.002912	0.001005
0.30	0.003107	0.001318
0.35	0.004370	0.001614
0.40	0.004604	0.001900
0.45	0.004821	0.002258
0.50	0.005734	0.002480
0.55	0.006086	0.003018
0.60	0.006561	0.003476
0.65	0.007047	0.003717
0.70	0.007813	0.004367
0.75	0.008710	0.005001
0.80	0.008821	0.005568
0.85	0.009550	0.006438
0.90	0.010120	0.007521
0.95	0.010648	0.008533

8. Conclusion

We propose the Beta Adjusted Covariance (BAC) estimator designed to use information from ETF stock prices to improve efficiency of any given original estimator based on stock prices only. The BAC effectively utilizes information contained in covariance of ETF proces and its components preserving and improving most of the original estimator properties such as consistency and asymptotic normality while reducing MSE. The simulations for both constant and stochastic volatility models have confirmed significant reduction of the error (4-5 times in terms of Frobenius norm of the difference of the matrices in our setup - square root of the sum of the squared error per element). Real high frequency data oriented empirical application demonstrated increased robustness of BAC to dilution of the trading observations (artificial trade frequency decrease) comparing to the original estimator.

9. Appendix

9.1. Solving for projection

We solve (13) by the method of Lagrange multipliers looking for a maximum of the Lagrange function:

$$\mathcal{L} = -\sum_{i=1}^{p}\sum_{j=1}^{i-1}\delta_{ij}^2 - \sum_{i=1}^{p}\lambda_i\left[\sum_{j=1}^{i-1}\bar{w}_j\delta_{ji} + \sum_{j=i+1}^{p}\bar{w}_j\delta_{ij} - \Delta_i\right]$$

under the constraint that

$$\begin{cases} \nabla \sum_{i=1}^{p}\sum_{j=1}^{i-1}\delta_{ij}^2 + \sum_{i=1}^{p}\nabla\lambda_i\left[\sum_{j=1}^{i-1}\bar{w}_j\delta_{ji} + \sum_{j=i+1}^{p}\bar{w}_j\delta_{ij} - \Delta_i\right] = 0 \\ \sum_{j=2}^{p}\bar{w}_j\delta_{j1} - \Delta_1 = 0 \\ \ldots \\ \sum_{j=1}^{i-1}\bar{w}_j\delta_{ji} + \sum_{j=i+1}^{p}\bar{w}_j\delta_{ij} - \Delta_i = 0 \\ \ldots \\ \sum_{j=1}^{p-1}\bar{w}_j\delta_{pj} - \Delta_p = 0 \end{cases},$$

where ∇ is gradient with respect to variables $\sigma_{21}, \ldots, \sigma_{p,p-1}, \lambda_1, \ldots, \lambda_p$. Taking partial derivatives with respect to δ_{ij} and λ_i and solving for $p(p+1)/2$ variables gives us correction matrix:

$$\begin{cases} 2\delta_{21} + \lambda_1\bar{w}_2 + \lambda_2\bar{w}_1 = 0 \\ \ldots \\ 2\delta_{ij} + \lambda_i\bar{w}_i + \lambda_j\bar{w}_j = 0 \\ \ldots \\ 2\delta_{p,p-1} + \lambda_p\bar{w}_p + \lambda_{p-1}\bar{w}_{p-1} = 0 \\ \sum_{j=2}^{p}\bar{w}_j\delta_{j1} = \Delta_1 \\ \ldots \\ \sum_{j=1}^{i-1}\bar{w}_j\delta_{ji} + \sum_{j=i+1}^{p}\bar{w}_j\delta_{ij} = \Delta_i \\ \ldots \\ \sum_{j=1}^{p-1}\bar{w}_j\delta_{pj} = \Delta_p \end{cases}$$

Efficient way to solve the system of the equations is to express δ_{ij} in terms of λ_i and substitute the former in last p equations of the system. This way we get p equations of p

unknown λ_i. Solving for λ_i and computing δ_{ij} gives the resulting correction.

$$P_\beta(\hat{\mathbf{\Sigma}}_0) = \hat{\mathbf{\Sigma}}_0 + \mathbf{\Delta}$$

9.2. Solving for projection - Weighted BAC case

We apply similar Lagrange multipliers optimization method as in previous BAC case. Lagrange function:

$$\mathcal{L} = -\sum_{i=1}^{p}\left[v_i\frac{\epsilon_i^2}{\sigma_i^2\sigma_Y^2} + \sum_{j=1,j\neq i}^{p}\frac{\delta_{ij}^2}{\sigma_i^2\sigma_j^2}\right] - \sum_{i=1}^{p}\lambda_i\left[\sum_{j=1}^{i-1}\bar{w}_j\delta_{ji} + \sum_{j=i+1}^{p}\bar{w}_j\delta_{ij} + \epsilon_i - \Delta_i\right]$$

σ_i and σ_Y are observed standard deviations of the ETF components, such that they reflect initial model σ and frequencies of trades and serve as confidence weights for the projection.

$$\begin{cases} \nabla\sum_{i=1}^{p}\left[v_i\frac{\epsilon_i^2}{\sigma_i\sigma_Y} + \sum_{j=1,j\neq i}^{p}\frac{\delta_{ij}^2}{\sigma_i\sigma_j}\right] + \sum_{i=1}^{p}\nabla\lambda_i\left[\sum_{j=1}^{i-1}\bar{w}_j\delta_{ji} + \sum_{j=i+1}^{p}\bar{w}_j\delta_{ij} + \epsilon_i - \Delta_i\right] = 0 \\ \sum_{j=2}^{p}\bar{w}_j\delta_{j1} + \epsilon_1 - \Delta_1 = 0 \\ ... \\ \sum_{j=1}^{i-1}\bar{w}_j\delta_{ji} + \sum_{j=i+1}^{p}\bar{w}_j\delta_{ij} + \epsilon_i - \Delta_i = 0 \\ ... \\ \sum_{j=1}^{p-1}\bar{w}_j\delta_{pj} + \epsilon_p - \Delta_p = 0 \end{cases}$$

Taking partial derivatives with respect to δ_{ij} and ϵ_i and solving for $p(p+3)/2$ variables gives us correction matrix:

$$\begin{cases} 2\frac{\delta_{21}}{\sigma_2^2\sigma_1^2} + \lambda_1\bar{w}_2 + \lambda_2\bar{w}_1 = 0 \\ ... \\ 2\frac{\delta_{ij}}{\sigma_i^2\sigma_j^2} + \lambda_i\bar{w}_i + \lambda_j\bar{w}_j = 0 \\ ... \\ 2\frac{\delta_{p,p-1}}{\sigma_p^2\sigma_{p-1}^2} + \lambda_p\bar{w}_p + \lambda_{p-1}\bar{w}_{p-1} = 0 \\ 2v_1\frac{\epsilon_1}{\sigma_1^2\sigma_Y^2} + \lambda_1\bar{w}_1 = 0 \\ ... \\ 2v_i\frac{\epsilon_i}{\sigma_i^2\sigma_Y^2} + \lambda_i\bar{w}_i = 0 \\ ... \\ 2v_p\frac{\epsilon_p}{\sigma_p^2\sigma_Y^2} + \lambda_p\bar{w}_p = 0 \\ \sum_{j=2}^p \bar{w}_j\delta_{j1} + \epsilon_1 = \Delta_1 \\ ... \\ \sum_{j=1}^{i-1} \bar{w}_j\delta_{ji} + \sum_{j=i+1}^p \bar{w}_j\delta_{ij} + \epsilon_i = \Delta_i \\ ... \\ \sum_{j=1}^{p-1} \bar{w}_j\delta_{pj} + \epsilon_p = \Delta_p \end{cases}$$

Efficient way to solve the system of the equations is to express δ_{ij} and ϵ_i in terms of λ_i and substitute the former in last p equations of the system. This way we get p equations of p unknown λ_i. Solving for λ_i and computing δ_{ij} gives the resulting correction.

9.3. Orthogonality of projection to the β-constraints subspace

In we use the fact that projection of the distance between initial approximation $\hat{\Sigma}_0$ and exact value $\mathbf{\Sigma}$ to the subset defined by β-constraints is less or equal that original distance:

$$\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_0\right\|_F \leq \left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_1\right\|_F$$

$$\left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_\mathbf{1}\right\|^2 = \left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_1 - \mathbf{\Delta} + \mathbf{\Delta}\right\|^2 = \left\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_\mathbf{0}\right\|^2 + \|\mathbf{\Delta}\|^2 + < \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_0 - \mathbf{\Delta}, \mathbf{\Delta} >$$

We need to prove that $< \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_0 - \mathbf{\Delta}, \mathbf{\Delta} >= 0$. Where elements of $\mathbf{\Delta}$ are equal to:

$$\delta_{21} = \frac{\lambda_1\bar{w}_2 + \lambda_2\bar{w}_1}{2}$$

and s_{ij} and $\tilde{s}_{ij}$ belong to subspace defined by constraints (12)

$$< \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_0 - \mathbf{\Delta}, \mathbf{\Delta} >= \sum_{i=1}^{p} \sum_{j=1, j\neq i}^{i} (s_{ij} - \tilde{s}_{ij}) * \frac{\lambda_i \bar{w}_j + \lambda_j \bar{w}_i}{2} =$$

$$\sum_{i=1}^{p} \lambda_i \sum_{j=1, j\neq i}^{i} (s_{ij} - \tilde{s}_{ij}) * \frac{\bar{w}_j}{2} + \sum_{j=1}^{p} \lambda_j \sum_{i=1, i\neq j}^{i} (s_{ij} - \tilde{s}_{ij}) * \frac{\bar{w}_i}{2} =$$

$$\sum_{i=1}^{p} \lambda_i \sum_{j=1, j\neq i}^{i} \frac{\Delta_i - \Delta_i}{2} + \sum_{j=1}^{p} \lambda_j \sum_{i=1, i\neq j}^{i} \frac{\Delta_j - \Delta_j}{2} = 0$$

9.4. Properties of the projection to the positive semidefinite set

Let's represent $\hat{\Sigma}$ as a sum of positive and negative definite matrices.

$$\hat{\mathbf{\Sigma}}_\beta = \hat{\mathbf{\Sigma}}_+ + \hat{\mathbf{\Sigma}}_-$$

$$\left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_\beta \right\|^2 = \left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_+ - \hat{\mathbf{\Sigma}}_- \right\|^2 = \left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_+ \right\|^2 + \left\| \hat{\mathbf{\Sigma}}_- \right\|^2 - < \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_+, \hat{\mathbf{\Sigma}}_- >$$

From the definition of the inner product in Euclidian space it follows that:

$$< \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_+, \hat{\Sigma}_- >= \mathrm{tr}(\mathbf{\Sigma}' \hat{\mathbf{\Sigma}}_-) - \mathrm{tr}(\hat{\mathbf{\Sigma}}_+, \hat{\mathbf{\Sigma}}_-)$$

It is easy to see that $\hat{\mathbf{\Sigma}}_+ \hat{\mathbf{\Sigma}}_- = 0$ as $\hat{\mathbf{\Sigma}}_+ \hat{\mathbf{\Sigma}}_- = \boldsymbol{Q} \boldsymbol{E}_+ \boldsymbol{E}_- \boldsymbol{Q}' = 0$. And $\mathbf{\Sigma}' \hat{\mathbf{\Sigma}}_-$ is a product of positive semidefinite and negative semidefinte matrices and is negative semidefinite, so the sum of its eigenvalues is not positive and so $\mathrm{tr}(\mathbf{\Sigma}' \hat{\mathbf{\Sigma}}_-) \leq 0$.

$$\left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_\beta \right\|^2 \geq \left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_+ \right\|^2 + \left\| \hat{\mathbf{\Sigma}}_- \right\|^2 \geq \left\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_+ \right\|^2$$

9.5. Trading frequencies of stocks used in empirical application

Ticker	Number of cleaned trades during the observation period
DIA	394454
GE	297100
PFE	229496
XOM	211304
AAPL	194469
JPM	190271
VZ	148996
CVX	127747
MRK	123085
KO	122298
WMT	121489
DIS	118406
V	98584
PG	95128
HD	94126
NKE	91871
JNJ	78631
CAT	75404
UTX	72679
AXP	70750
BA	67239
MCD	62476
IBM	60357
GS	56165
DD	55032
UNH	46475
MSFT	45332
MMM	36368
TRV	26028
CSCO	21171
INTC	21134

References

[1] Ole E Barndorff-Nielsen, P Reinhard Hansen, Asger Lunde, and Neil Shephard. Realized kernels in practice: Trades and quotes. *The Econometrics Journal*, 12(3), 2009.

[2] Ole E Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162(2):149–169, 2011.

[3] Kris Boudt, Sébastien Laurent, Asger Lunde, Rogier Quaedvlieg, and Orimar Sauri. Positive semidefinite integrated covariance estimation, factorizations and asynchronicity. *Journal of Econometrics*, 196(2):347–367, 2017.

[4] James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer, 1986.

[5] Thomas W Epps. Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a):291–298, 1979.

[6] Forbes. Global etf industry swells to $4.6 trillion, blackrock remains the leader by far, 2017. Online; accessed 19-February-2018.

[7] Nikolaus Hautsch, Lada M Kyj, and Roel CA Oomen. A blocking and regularization approach to high-dimensional realized covariance estimation. *Journal of Applied Econometrics*, 27(4):625–645, 2012.

[8] Nicholas J Higham. Computing the nearest correlation matrix - a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002.

[9] Alexander Shapiro. Differentiability properties of metric projections onto convex sets. *Journal of Optimization Theory and Applications*, 169(3):953–964, 2016.