Institute for Statistics and Econometrics
Chair in Statistics and Empirical Economics Research
Christian-Albrechts-Universität zu Kiel

Master's Thesis in Quantitative Finance

First assessor: Prof. Dr. Matei Demetrescu
Second assessor: Prof. Dr. Markus Haas

Efficient estimation of predictive models using high-frequency high-dimensional data

Submitted by:
Jan P. Wöltjen
Quantitative Finance
Matr. Nr.: 1110126
j.p.woeltjen@gmail.com
Kiel, Germany
August 2020

Abstract

In this thesis, the data efficiency of linear and nonlinear regression models of asset return panel data is enhanced by accounting for cross-sectional correlations and longitudinal volatility clusters of residuals. The procedure is motivated by the infeasible generalized least squares estimator. In an extension, a generalized least squares loss function is proposed to efficiently fit nonlinear relationships via deep neural networks. Feasibility is achieved by estimating the unobserved covariance matrix of residuals with a nonparametrically eigenvalue-regularized ensembled pairwise integrated covariance (NER_EPIC) matrix estimator applied to high-frequency returns in high dimensions. Monte Carlo evidence confirms efficiency gains for linear and nonlinear conditional expectation models in finite-samples. A study of historical stock market data for the 100 largest US-based stocks shows substantially improved portfolio return characteristics of generalized models compared to their standard counterparts. A trading strategy based on the predictions of a neural network, minimizing the proposed generalized objective function, generates an out-of-sample information ratio of 2.59. Compared to a model with the same hyperparameters but minimizing the conventional MSE loss function, this represents an improvement of close to 150%.

Contents

List of figures

List of tables

List of acronyms

Adam	Adaptive Moment Estimation optimizer
CR	Current-assets-to-current-liabilities ratio
D/E	Debt-to-equity ratio
EBIT	Earnings before interest and taxes
EPIC	Ensembled pairwise integrated covariance
EV	Enterprise Value
FGLS	Feasible generalized least squares
FGMSE	Feasible generalized mean squared error
GARCH	Generalized autoregressive conditional heteroskedasticity
GICS	Global industrial classification standard
GLS	Generalized least squares
GMSE	Generalized mean squared error
GR	Revenue growth rate
HY	Hayashi-Yoshida
IR	Information ratio
KRVM	The kernel realized volatility matrix
MC	Monte Carlo
MOM	Momentum
MRC	Modulated realized covariance
MR	Mean reversion
MSE	Mean squared error
MSRC	Multi-scale realized covariance
MSRV	Multi-scale realized volatility
MV	Minimum variance
NER	Nonparametrically eigenvalue-regularized
NERCOME	Nonparametric eigenvalue-regularized covariance matrix estimator
NERIVE	Nonparametrically eigenvalue-regularized integrated covariance matrix estimator
NM	Net profit margin
OLS	Ordinary least squares
PR	Dividend payout ratio
ROIC	Return on invested capital
ReLU	Rectified Linear Unit
SEC	Security and Exchange Commission

TSCV	Two-scales realized covariance
TSRV	Two-scales realized volatility
TTM	Trailing 12 Months
càdlàg	Continue à droite, limite à gauche ("right continuous with left limits")
e.g.	Exempli gratia ("for the sake of an example")
i.e.	Id est ("that is")
i.i.d.	Independent and identically distributed

List of symbols

a	Scalar
$\boldsymbol{a}$	Vector
$\mathbf{A}$	Matrix
$\mathbf{A}'$	Transpose of $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of $\mathbf{A}$

Operators and standard symbols

$\int_0^T$	The definite integral over the interval $[0, T]$
$\sum_{i=1}^n$	The sum from 1 to n
a^+	Positive part of a
$\bar{a}$	The mean of a
$\mathcal{O}$	Big-O
O_p	Big-O in probability
o_p	Small-O in probability
$\xrightarrow{\mathrm{p}}$	Convergence in probability
$\xrightarrow{\mathbf{a.s}}$	Almost sure convergence
$\sim$	Distributed
$\overset{a}{\sim}$	Asymptotically distributed
E	The expectation operator
Var	The variance operator
std	The standard deviation operator
Avar	The asymptotic variance operator
$\boldsymbol{\gamma}^{(h)}$	The hth order autocovariance operator
min	The minimum operator
abs	The absolute value operator
diag	The diagonal operator
tr	The trace operator
rank	The rank operator
Δ	The difference operator
$\|\mathbf{A}\|_F$	The Frobenius norm of $\mathbf{A}$
$\mathbf{I}_p$	The identity matrix of dimension $p \times p$
$\odot$	The Hadamard product
$\mathbf{J}_p$	A p-dimensional square matrix of ones
Supp	The support of a function

$\mathcal{H}_g$	The Hilbert transform of g
PV	The Cauchy principal value
κ	A kernel estimator
R^2	Coefficient of determination

Sets

$\forall$	For all
$\in$	Element of
$\cap$	Intersection
$\emptyset$	Empty set
$\mathbb{R}$	The real numbers

Sub- and superscripts

j	Asset
t	Time
p	The number of assets
n	The number of time periods
k	The number of predictive features

Eigendecomposition

λ_i	The ith eigenvalue
$\mathbf{\Lambda}$	A diagonal matrix, whose elements are the eigenvalues
$\mathbf{q}_i$	The ith eigenvector
$\mathbf{Q}$	A orthogonal matrix, whose columns are the eigenvectors

Discrete processes

$\Delta y_{j,t}$	The daily open to-close log-return (i.e., the change in log-price) of asset j of day t observable at $t+1$
$\mathbf{\Delta y}_t$	Stacked $\Delta y_{j,t}$
$\mathbf{\Delta y}$	Stacked $\mathbf{\Delta y}_t$
$z_{j,t,i}$	The ith regressor of asset j observable at time t
$\mathbf{z}_{j,t}$	The feature row-vector of k supposedly predictive regressors of asset j observable at time t
$\mathbf{Z}_t$	Stacked $\mathbf{z}_{j,t}$
$\mathbf{Z}$	Stacked $\mathbf{Z}_t$

$u_{j,t}$	The residual with latent factors $u_{j,t} = v_{j,t} + \mathbf{f}_t \boldsymbol{\beta}_{\mathbf{f_t},j,t}$
$\mathbf{u}_t$	Stacked $u_{j,t}$
$\mathbf{u}$	Stacked $\mathbf{u}_t$
$v_{j,t}$	The idiosyncratic residual of asset j at time t
$\mathbf{v}_t$	Stacked $v_{j,t}$
$\mathbf{v}$	Stacked $\mathbf{v}_t$
$\boldsymbol{\beta}$	The parameter vector
$\widehat{\boldsymbol{\beta}}$	An estimate of the parameter vector
$\mathbf{f}_t$	Factor returns at time t
$\boldsymbol{\beta}_{\mathbf{f_t},j,t}$	Factor loading vector of asset j at time t
$\beta_{M,j,t}$	Market factor loading of asset j at time t
$\mathbf{B}_{\mathbf{f}_t,t}$	Factor loading matrix for p assets at time t
$\mathbf{X}_n$	A matrix of n i.i.d. observations with zero-mean for p variables (also denoted as $\mathbf{X}$)
$\boldsymbol{\Sigma}$	The unconditional population covariance matrix of p asset returns over all t
$\boldsymbol{\Sigma}_t$	The conditional population covariance matrix of p asset returns at time t
$\boldsymbol{\Sigma}_{\mathbf{f}_t}$	The conditional population covariance matrix of factor returns at time t
$\boldsymbol{\Sigma}_{\mathbf{v}_t}$	The conditional population covariance matrix of idiosyncratic residuals at time t
$\boldsymbol{\Sigma}_{\mathbf{u}_t}$	The conditional population covariance matrix of residuals at time t
$\boldsymbol{\Phi}$	A block-diagonal matrix with elements $\boldsymbol{\Sigma}_{\mathbf{u}_t}$ for $t = 1 \dots n$
$\boldsymbol{w}_t$	Portfolio weight vector at day t
c_n	The sample concentration ratio $c_n = \frac{p}{n}$
c	The limiting concentration ratio
$[\sigma_0^2, \mu, \alpha, \beta, \omega]$	A GARCH(1, 1) specification

Low-frequency covariance estimators

$\mathbf{S}$	The sample covariance matrix
$\widehat{\delta}_i$	A rotation equivariant estimate of the ith eigenvalues
$\widehat{\boldsymbol{\delta}}$	Rotation equivariant estimates of the p eigenvalues
$\widehat{\boldsymbol{\Delta}}$	A diagonal matrix with elements $\widehat{\boldsymbol{\delta}}$
d_i^*	The infeasible finite-sample optimal rotation equivariant estimate of δ_i

$\widehat{d}_i^{(l,o)}$	The estimate of the ith eigenvalue via the linear oracle shrinkage estimator
$\widehat{d}_i^{(l)}$	The estimate of the ith eigenvalue via the linear shrinkage estimator
$\widehat{d}_i^{(o,nl)}$	The estimate of the ith eigenvalue via the analytic nonlinear oracle shrinkage estimator
$\widehat{d}_i^{(nl)}$	The estimate of the ith eigenvalue via the analytic nonlinear shrinkage estimator
$\widetilde{d}_{n_1,i}^{(\mathbf{NERCOME})}$	The estimate of the ith eigenvalue via the NERCOME estimator based on subsample n_1
$\widehat{\mathbf{S}}$	The linearly shrunk sample covariance matrix
$\widetilde{\mathbf{S}}$	The nonlinearly shrunk sample covariance matrix
$\widetilde{\mathbf{S}}_{n_1,M}^{(\mathbf{NERCOME})}$	The NERCOME estimator based on subsample n_1 and M iterations
ρ	The linear shrinkage intensity

Neural network

f	A deep feedforward artificial neural network
ℓ	Layer
L	Number of layers
N_ℓ	Number of neurons
$\mathbf{A}_\ell$	The weight matrix of layer ℓ of a neural network
$\mathbf{b}_\ell$	The bias vector of layer ℓ of a neural network
$\mathbf{W}_\ell$	An affine linear map of layer ℓ of a neural network
$\sigma(\cdot)$	An activation function
$\mathcal{L}$	A loss function
$p^{(do)}$	Parameter of dropout
$\alpha^{(Adam)}$	The Adam optimizer's learning rate
$\beta_1^{(Adam)}$	Exponential decay factor of first moment estimate of Adam optimizer
$\beta_2^{(Adam)}$	Exponential decay factor of second moment estimate of Adam optimizer

Continuous processes

Ω	The sample space
$\mathcal{F}$	The σ-algebra

$\{\mathcal{F}_t\}_{0\leq t\leq T}$	The filtration i.e., an increasing sequence of sub-σ-algebras of $\mathcal{F}$ over the interval [0,T]
$\mathbb{P}$	The probability measure
$\{\mathbf{X}_t\}_{0\leq t\leq T}$	The unobserved p-dimensional efficient continuous log-price process over the interval [0,T]
$\{\mathbf{Y}_t\}_{0\leq t\leq T}$	The observed p-dimensional contaminated continuous log-price process over the interval [0,T]
$\{\boldsymbol{\epsilon}_t\}_{0\leq t\leq T}$	The p-dimensional continuous market microstructure noise process over the interval [0,T]
$X_t^{(j)}$	The unobserved efficient continuous log-price process of asset j
$Y_t^{(j)}$	The observed contaminated continuous log-price process of asset j
$\epsilon_t^{(j)}$	The continuous market microstructure noise process process of asset j
$\{\mathbf{W}_t\}_{0\leq t\leq T}$	A p-dimensional standard Brownian motion over the interval [0,T]
$\boldsymbol{\sigma}_t$	The volatility $\boldsymbol{\sigma}_t \in \mathbb{R}^{p\times p}$
$\boldsymbol{\mu}_t$	The drift
$\mathcal{V}$	A grid
$t^{(j)}$	The sequence of observation times of asset j's log-price process
$\tau_m^{(j)}$	The mth 'previous tick time' of asset j
$\widetilde{\tau}_l$	The lth sample split point
$\boldsymbol{\Sigma}(a,b)$	The integrated covariance matrix of the log-return processes over the interval $[a,b]$
$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(a,b)$	The integrated covariance matrix of the noise processes over the interval $[a,b]$
$\langle X^{(k)}, X^{(l)}\rangle$	The quadratic (co)variation of asset k and l, $\langle X^{(k)}, X^{(l)}\rangle = \Sigma_{kl}$

High-frequency integrated covariance estimators

$[Y^{(k)}, Y^{(l)}]^{(K)}$	The realized (co)variance of the Kth scale and the kth and lth asset
$\widehat{\Sigma}_{jj}^{(TSRV)}$	The two-scales realized volatility estimator of asset j's log-returns
$\widehat{\Sigma}_{jj}^{(MSRV)}$	The multi-scale realized volatility estimator of asset j's log-returns
$\widehat{\Sigma}_{kl}^{(TSCV)}$	The two-scales realized covariance estimator of log-returns for

	asset k and l
$\widehat{\Sigma}_{kl}^{(HY)}$	The Hayashi-Yoshida estimator of the integrated covariance of log-returns for asset k and l
$\widehat{\mathbf{\Sigma}}^{(MSRC)}$	The multivariate multi-scale realized covariance estimator
$\widehat{\mathbf{\Sigma}}^{(KRVM)}$	The kernel realized volatility matrix estimator
$\widehat{\mathbf{\Sigma}}^{(MRC)}$	The modulated realized covariance estimator
$\widehat{\mathbf{\Sigma}}^{(EPIC)}$	The ensembled pairwise integrated covariance matrix estimator
$\widehat{\mathbf{\Sigma}}^{(NERIVE)}$	The nonparametrically eigenvalue-regularized integrated covariance matrix estimator

1 Introduction

Economic theory states that an asset is worth the sum of its discounted future cash flows. For risky assets, such as stocks, cash flows are uncertain in both magnitude and time of distribution, as is the appropriate discount rate. Trading profits can be made in expectation if the current price observed in the market deviates from the expected net present value given predictive features. The better the conditional expectation model, the higher is the likelihood of such an occurrence. Still, most of the variance of asset returns is due to unpredictable events, such as the broad economic climate, industry-specific, and company-specific shocks, causing the R^2 of predictive models to be very small. Importantly, many of these factors are common to more than one stock. Some of them may be proxied by observable variables while others are latent, evidenced by high correlations of returns. This thesis proposes a method to account for cross-sectional correlations and longitudinal volatility clusters of residuals and thereby increase the efficiency of conditional expectation models. Improved efficiency implies higher accuracy of parameter estimates and thus higher predictive power.

1.1 Approximate factor structure of asset returns

Consider the regression equation for asset $j = 1, \ldots, p$ at time t

$$\Delta y_{j,t} = \mathbf{z}_{j,t}\boldsymbol{\beta} + v_{j,t} + \mathbf{f}_t\boldsymbol{\beta}_{\mathbf{f_t},j,t}, \tag{1}$$

where $\Delta y_{j,t}$ is the daily open-to-close log-return of day t observable at $t+1$, $\mathbf{z}_{j,t}$ is the feature row-vector of k supposedly predictive regressors (including an intercept) observable at t, $\mathbf{f}_t$ is the row-vector of factor returns with the corresponding loading vector $\boldsymbol{\beta}_{\mathbf{f_t},j,t}$, and $v_{j,t}$ is the asset-specific residual for $j = 1, \ldots, p$ of day t. $\boldsymbol{\beta}$ denotes the parameter vector of interest. Since $\mathbf{f}_t$ is assumed latent, it is accommodated by the error term $u_{j,t} = v_{j,t} + \mathbf{f}_t\boldsymbol{\beta}_{\mathbf{f_t},j,t}$. The covariance matrix of the residual $\mathbf{u}_t$ for all p assets at time t is

$$\boldsymbol{\Sigma}_{\mathbf{u}_t} = \mathbf{B}'_{\mathbf{f}_t,t}\boldsymbol{\Sigma}_{\mathbf{f}_t}\mathbf{B}_{\mathbf{f}_t,t} + \boldsymbol{\Sigma}_{\mathbf{v}_t}, \tag{2}$$

where $\mathbf{B} = \left(\boldsymbol{\beta}_{\mathbf{f_t},1,t}, \ldots, \boldsymbol{\beta}_{\mathbf{f_t},p,t}\right)$ and $\boldsymbol{\Sigma}_{\mathbf{f}_t}$ is the covariance matrix of factor returns.

If the cross-sectional correlation of asset returns due, in part, to the factor structure is not accounted for, then using ordinary least squares (OLS) to estimate $\boldsymbol{\beta}$ in Equation (1) is inefficient since the Gauss-Markov theorem is violated. In particular, the covariance matrix of the residual is not equal to the scaled identity matrix. Hence, there exists a more efficient estimator that linearly transforms the regression equation

at time t with $\mathbf{\Sigma}_{\mathbf{u}_t}^{-1/2}$ for $t = 1, \ldots, n$. This estimator is the generalized least squares (GLS) estimator of Aitken (1936) with a specific structure imposed on the covariance matrix. The major challenge in implementing the GLS estimator in practice is that $\mathbf{\Sigma}_{\mathbf{u}_t}$ is not known and finding a good estimate is exceedingly difficult since $p(p+1)/2$ unique parameters have to be estimated with p observations. The way to overcome this seemingly impossible challenge is to utilize intraday price data that are observed between the daily open and close prices and regularize the covariance matrix estimate by imposing structure.

To capture the correlation due to the low-rank factor structure, factor models can be used. Static factor models assume that (i) the loadings, (ii) the factor covariance matrix, and (iii) the residual covariance matrix are constant through time. Such models are studied by Bai (2003) in high dimensions, i.e., when p goes to infinity, who establishes an inferential theory in this setting. This author finds that estimated common components are asymptotically normal with convergence rate $\min(\sqrt{n}, \sqrt{p})$. As usual, factor loadings can be consistently estimated when n grows large. In contrast to the classical setting, however, where p is fixed, the common factors themselves can now be consistently estimated as well.

In dynamic factor models, introduced by Gourieroux and Jasiak (2001), at least one of the three conditions of a static factor model is violated. In particular, the conditional factor model, studied by Avramov and Chordia (2006), allows for dynamic factor loadings. Unconditional dynamic factor models, in contrast, assume constant loadings but time-varying covariance matrices of factors and residuals. Engle (2002) proposes a method to account for time-varying covariance matrices of modest dimensions, which can be used to model the dynamic covariance matrix of factor returns.

Empirical evidence suggests, however, that asset returns typically exhibit covariances not fully explained by an exact factor model, which assumes the asset-specific covariance matrix $\mathbf{\Sigma}_{\mathbf{v}_t}$ to be diagonal. In this case, one speaks of an approximate factor model, introduced by Chamberlain and Rothschild (1983). As detailed in Section 4.1, estimating the elements of $\mathbf{\Sigma}_{\mathbf{v}_t}$ is then problematic in high dimensions. Regularization must be employed to improve the conditioning of its estimate $\widehat{\mathbf{\Sigma}}_{\mathbf{v}_t}$.

A successful attempt at this in a static setting is made by Fan et al. (2013). These authors assume that, after a low-rank approximation has been conditioned on, $\mathbf{\Sigma}_{\mathbf{v}}$ is sparse. This assumption is justified by the fact that stocks usually belong to one or very few industries, within which cross-sectional correlation is elevated. They propose a thresholding estimator to regularize $\widehat{\mathbf{\Sigma}}_{\mathbf{v}}$. Other approaches along those lines aim at incorporating additional data sources to improve accuracy. Ait-Sahalia and Xiu (2017), for example, order the rows and columns of $\widehat{\mathbf{\Sigma}}_{\mathbf{v}_t}$ according to their corresponding stocks' global industrial classification standard (GICS) code and then promote a block-

diagonal structure. These authors also extend the analysis to a dynamic setting with high-frequency data.

Another approach to consistently estimate covariance matrices in high dimension is the class of rotation equivariant estimators originally due to Stein (1975). These estimators do not assume any specific structure of the covariance matrix. Instead, they improve the conditioning of its estimate by reducing eigenvalue dispersion. An estimator of this type is the linear shrinkage estimator of Ledoit and Wolf (2004), which has been widely adopted by the financial econometrics literature. Building on their success, Ledoit and Wolf (2012) propose a method that allows for a more nuanced shrinkage function. An approach that combines an approximate factor model with shrinkage estimation is proposed by De Nard et al. (2018). These authors build on a dynamic shrinkage estimator introduced in Engle et al. (2019). They find that a single-factor model with subsequent nonlinear shrinkage of the time-varying residual covariance matrix performs well for a portfolio allocation problem. Lam (2016) proposes a nonparametrically eigenvalue-regularized covariance estimator that nonlinearly shrinks the covariance matrix estimate optimally without having to condition on a factor model first. Lam and Feng (2018) study this estimator in a high-frequency setting and confirm favorable performance relative to competing approaches. Since the covariance matrix, and not the factors themselves, is of primary interest, an estimator of this form shall be the focus of the thesis.

1.2 The structure of the thesis

The thesis is structured as follows. Section 2 details a linear conditional expectation model and establishes the possible efficiency gains. This motivates the feasible linear model. In Section 3, the model is extended to a nonlinear specification via a deep neural network with a linearly transformed loss function. Section 4 describes the estimation procedure of the covariance matrix to make the large leap from GLS to FGLS. The difficulties associated with high-dimensional covariance matrix estimation are described in Section 4.1 and a solution via eigenvalue-regularization is proposed in Section 4.2. The covariance estimator is based on high-frequency price observations, which pose two major challenges. First, they are observed non-synchronously across assets. Non-synchronicity introduces bias into the covariance estimates of the realized covariance estimator. Second, they are observed with microstructure noise, which biases the realized variance estimates. Several high-frequency estimators that overcome those issues are discussed in Section 4.3 and an ensemble estimator is proposed to improve finite-sample properties. Monte Carlo evidence in Section 5 confirms the efficiency gains for the proposed generalized estimators in finite-samples. An em-

pirical study is performed with historic stock market data in Section 6. This study demonstrates superior performance of trading strategies based on predictions of the generalized versions of both the linear regression model and the neural network. Section 7 concludes.

2 A linear conditional expectation model

2.1 The generalized least squares estimator

Consider the regression equation

$$\mathbf{\Delta y}_t = \mathbf{Z}_t\boldsymbol{\beta} + \mathbf{u}_t, \tag{3}$$

where $\mathbf{\Delta y}_t$ is the daily log-return, $\mathbf{Z}_t$ is the matrix of k supposedly predictive regressors (including a column of ones for the intercept), and $\mathbf{u}_t$ is the residual with approximate factor structure for p financial assets at time t given by the following objects:

$$\mathbf{\Delta y}_t = \begin{pmatrix} \Delta y_{t1} \\ \vdots \\ \Delta y_{tp} \end{pmatrix}, \quad \mathbf{Z}_t = \begin{pmatrix} \mathbf{z}_{t1} \\ \vdots \\ \mathbf{z}_{tp} \end{pmatrix} \quad \mathbf{u}_t = \begin{pmatrix} \mathbf{u}_{t1} \\ \vdots \\ \mathbf{u}_{tp} \end{pmatrix}. \tag{4}$$

$\boldsymbol{\beta}$ denotes the parameter vector of interest. Stacking these objects over all time points, the system of equations can be written as

$$\begin{array}{ccccccc} \mathbf{\Delta y} & = & \mathbf{Z} & \boldsymbol{\beta} & + & \mathbf{u} \\ np \times 1 & & np \times k & k \times 1 & & np \times 1 \end{array} \tag{5}$$

with

$$\mathbf{\Delta y} = \begin{pmatrix} \mathbf{\Delta y}_1 \\ \vdots \\ \mathbf{\Delta y}_n \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_n \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{pmatrix}. \tag{6}$$

The covariance matrix of the stacked residual is given by

$$\mathbf{\Phi} = \mathrm{E}\left[\mathbf{u}\mathbf{u}'\right] = \mathrm{E}\left[\begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1' & \dots & \mathbf{u}_n' \end{pmatrix}\right] = \begin{pmatrix} \mathrm{E}\left[\mathbf{u}_1\mathbf{u}_1'\right] & \dots & \mathrm{E}\left[\mathbf{u}_1\mathbf{u}_n'\right] \\ \vdots & \dots & \vdots \\ \mathrm{E}\left[\mathbf{u}_n\mathbf{u}_1'\right] & \dots & \mathrm{E}\left[\mathbf{u}_n\mathbf{u}_n'\right] \end{pmatrix}. \tag{7}$$

Assuming that the autocovariances are zero, the covariance matrix can be written in the block-diagonal form

$$\mathbf{\Phi} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{u}_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{\Sigma}_{\mathbf{u}_n} \end{pmatrix}. \tag{8}$$

In the econometrics literature, it has been widely established that stock returns exhibit significant correlations in the cross-section. Further, stock returns exhibit no or very weak autocorrelations but their variances are dynamically evolving through time. They are well described by autoregressive stochastic variance models, going back to Taylor (1982), where shocks occur randomly, amplifying variance, which will then persist through time. Since $\mathbf{\Phi} \neq \sigma^2 \mathbf{I}_{np}$, OLS is inefficient. To establish efficiency, the model can be linearly transformed such that the covariance matrix of the transformed residuals $\mathrm{E}[\mathbf{u}^*\mathbf{u}^{*\prime}] = \mathbf{I}_{np}$. To achieve this, define the square-root inverse matrix $\mathbf{\Phi}^{-1/2}$ such that

$$\mathbf{\Phi}^{-1/2}\mathbf{\Phi}(\mathbf{\Phi}^{-1/2})' = \mathbf{I}_{np}. \tag{9}$$

Multiplying Equation (5) with $\mathbf{\Phi}^{-1/2}$ from the left, the following transformed equation is obtained:

$$\begin{aligned} \mathbf{\Phi}^{-1/2}\mathbf{y} &= \mathbf{\Phi}^{-1/2}\mathbf{Z}\boldsymbol{\beta} + \mathbf{\Phi}^{-1/2}\mathbf{u} \\ \mathbf{\Delta}\mathbf{y}^* &= \mathbf{Z}^*\boldsymbol{\beta} + \mathbf{u}^*. \end{aligned} \tag{10}$$

It is easy to see that the covariance matrix of the transformed residual is of the desired form such that OLS of the transformed system is efficient according to the Gauss-Markov theorem

$$\mathrm{E}[\mathbf{u}^*\mathbf{u}^{*\prime}|\mathbf{Z}^*] = \mathbf{\Phi}^{-1/2}\mathrm{E}[\mathbf{u}\mathbf{u}'|\mathbf{Z}](\mathbf{\Phi}^{-1/2})' = \mathbf{\Phi}^{-1/2}\mathbf{\Phi}(\mathbf{\Phi}^{-1/2})' = \mathbf{I}_{pn}. \tag{11}$$

The GLS estimator of Aitken (1936) can be obtained by starting from the standard OLS definition applied to the transformed system as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GLS} &= \left(\mathbf{Z}^{*\prime}\mathbf{Z}^*\right)^{-1}\mathbf{Z}^{*\prime}\mathbf{\Delta}\mathbf{y}^* \\ &= \left(\mathbf{Z}'\mathbf{\Phi}^{-1/2}(\mathbf{\Phi}^{-1/2})'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{\Phi}^{-1/2}(\mathbf{\Phi}^{-1/2})'\mathbf{\Delta}\mathbf{y} \\ &= \left(\mathbf{Z}'\mathbf{\Phi}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{\Phi}^{-1}\mathbf{\Delta}\mathbf{y}. \end{aligned} \tag{12}$$

If Equation (8) holds, the estimator can be written in the more computationally stable

and efficient form of Zellner (1962)

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \mathbf{Z}_t \right)^{-1} \left(\sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \boldsymbol{\Delta}\mathbf{y}_t \right) \tag{13}$$

by noting that

$$\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1' & \dots & \mathbf{Z}_n' \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{u}_1}^{-1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{\mathbf{u}_n}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_n \end{pmatrix} = \sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \mathbf{Z}_t \tag{14}$$

and

$$\mathbf{Z}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Delta}\mathbf{y} = \begin{pmatrix} \mathbf{Z}_1' & \dots & \mathbf{Z}_n' \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{u}_1}^{-1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{\mathbf{u}_n}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta}\mathbf{y}_1 \\ \vdots \\ \boldsymbol{\Delta}\mathbf{y}_n \end{pmatrix} = \sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \boldsymbol{\Delta}\mathbf{y}_t. \tag{15}$$

One important difference, however, is that this estimator, unlike Zellner (1962), does not assume a constant covariance matrix through time. For proving consistency, asymptotic normality, and efficiency the following assumptions will be made:

Assumption GLS.1. $\mathrm{E}\left[\mathbf{u}_t | \mathbf{y}_1, \dots, \mathbf{y}_t, \mathbf{Z}_1, \dots, \mathbf{Z}_t, \boldsymbol{\Sigma}_{\mathbf{u}_1}, \dots, \boldsymbol{\Sigma}_{\mathbf{u}_t}\right] = 0.$

Assumption GLS.2. $\mathrm{Var}[\mathbf{u}|\mathbf{Z}] = \mathrm{E}\left[\mathbf{u}\mathbf{u}'|\mathbf{Z}\right] = \boldsymbol{\Phi}$ *is a positive definite block-diagonal matrix with elements* $\boldsymbol{\Sigma}_{\mathbf{u}_t}$ *for* $t = 1, \dots, n$ *along the diagonal.*

Assumption GLS.3. $\operatorname{rank} \mathrm{E}[\mathbf{Z}\mathbf{Z}'] = kp.$

Assumption GLS.4. $\mathbf{y_t}$, $\mathbf{Z_t}$, $\boldsymbol{\Sigma}_{\mathbf{u}_t}$ *are stationary and ergodic processes.*

Assumption GLS.5. $\mathbf{y_t}$, $\mathbf{Z_t}$, $\boldsymbol{\Sigma}_{\mathbf{u}_t}$ *have finite 6th moments.*

Theorem 2.1. *Under the assumptions GLS.1, GLS.2, GLS.3, GLS.4, and GLS.5, the GLS estimator is consistent, i.e.,* $\hat{\boldsymbol{\beta}}_{GLS} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}$.

Proof. Beginning by substituting Equation (3) for $\boldsymbol{\Delta}\mathbf{y}$, the following standard form is obtained:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GLS} &= \left(n^{-1} \sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \mathbf{Z}_t \right)^{-1} \left(n^{-1} \sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \boldsymbol{\Delta}\mathbf{y}_t \right) \\ &= \boldsymbol{\beta} + \left(n^{-1} \sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \mathbf{Z}_t \right)^{-1} \left(n^{-1} \sum_{t=1}^{n} \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1} \mathbf{u}_t \right) \end{aligned} \tag{16}$$

Consistency is given if the second term converges in probability to zero. By a weak law of large numbers (WLLN)

$$n^{-1}\sum_{t=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t \xrightarrow{p} \mathrm{E}\left(\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t\right) =: \mathbf{A}, \tag{17}$$

where $\mathbf{A}$ is finite due to GLS.4 and GLS.5 and has full rank due to GLS.2 and GLS.3. Similarly by a WLLN

$$n^{-1}\sum_{t=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t \xrightarrow{\mathrm{p}} \mathrm{E}\left(\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t\right). \tag{18}$$

By the law of iterated expectations and GLS.1, GLS.4, and GLS.5

$$\mathrm{E}\left[\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t\right] = \mathrm{E}\left[\mathrm{E}\left(\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t|\mathbf{Z}_t\right)\right] = 0. \tag{19}$$

This gives the desired result

$$\hat{\boldsymbol{\beta}}_{GLS} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}.$$

□

Theorem 2.2. *Under the assumptions GLS.1, GLS.2, GLS.3, GLS.4, and GLS.5, the GLS estimator is asymptotically normally distributed, i.e.,* $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}\right) \xrightarrow{\mathrm{d}}$ *Normal* $(\mathbf{0}, \mathbf{A}^{-1})$ *with* $\mathbf{A} = \mathrm{E}\left(\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t\right)$.

Proof.

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}\right) = \left(n^{-1}\sum_{i=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t\right)^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t\right) \tag{20}$$

By the martingale central limit theorem of Ibragimov (1963)

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t \xrightarrow{d} \mathrm{Normal}(\mathbf{0}, \mathbf{B}), \tag{21}$$

where

$$\mathbf{B} = \mathrm{E}\left(\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t\mathbf{u}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t\right). \tag{22}$$

By the law of iterated expectations

$$\mathbf{B} = \mathrm{E}\left(\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t\mathbf{u}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t\right) = \mathrm{E}\left(\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t\right) = \mathbf{A}. \tag{23}$$

Since $n^{-1/2}\sum_{i=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t = \mathrm{O}_p(1)$ and $\left(n^{-1}\sum_{i=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{Z}_t\right)^{-1} - \mathbf{A}^{-1} = \mathrm{o}_p(1)$, it can

be written $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}\right)=\mathbf{A}^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\mathbf{Z}_t'\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\mathbf{u}_t\right)+\mathrm{o}_p(1)$. From the asymptotic equivalence lemma it then follows that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}\right)\overset{a}{\sim}\text{Normal}\left(\mathbf{0},\mathbf{A}^{-1}\right).$$

□

Theorem 2.3. *Under the assumptions GLS.1, GLS.2, GLS.3, GLS.4, and GLS.5, the GLS estimator is more efficient than OLS if* $\boldsymbol{\Phi}\neq\sigma^2\mathbf{I}_{np}$.

Proof. The conditional variance of the GLS estimator is

$$\begin{aligned}\text{Var}\left[\hat{\boldsymbol{\beta}}_{GLS}|\mathbf{Z}\right]&=\text{E}\left[\left(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{GLS}-\boldsymbol{\beta}\right)'|\mathbf{Z}\right]\\&=\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}\text{E}\left[\mathbf{u}\mathbf{u}'|\mathbf{Z}\right]\boldsymbol{\Phi}^{-1}\mathbf{Z}\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\\&=\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{-1}\mathbf{Z}\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\\&=\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\\&=\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\end{aligned}$$

The conditional variance of the OLS estimator is

$$\begin{aligned}\text{Var}\left[\hat{\boldsymbol{\beta}}_{OLS}|\mathbf{Z}\right]&=\text{E}\left[\left(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{\beta}\right)'|\mathbf{Z}\right]\\&=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\left[\mathbf{u}\mathbf{u}'|\mathbf{Z}\right]\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\\&=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Phi}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\end{aligned}$$

The difference between the variances of the OLS and the GLS estimators is

$$\begin{aligned}\mathbf{D}&=\text{Var}\left[\hat{\boldsymbol{\beta}}_{OLS}|\mathbf{Z}\right]-\text{Var}\left[\hat{\boldsymbol{\beta}}_{GLS}|\mathbf{Z}\right]\\&=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Phi}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}-\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\\&=\mathbf{F}\boldsymbol{\Phi}\mathbf{F}',\end{aligned}$$

where $\mathbf{F}=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'-\left(\mathbf{Z}'\boldsymbol{\Phi}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Phi}^{-1}$. By assumption $\boldsymbol{\Phi}\neq\sigma^2\mathbf{I}_{np}$. Further, $\mathbf{F}\boldsymbol{\Phi}\mathbf{F}'$ is of symmetric quadratic form. Hence, $\mathbf{D}$ is positive definite, which implies that all its diagonal elements are greater than zero. This proves the claim. □

2.2 Making GLS feasible

In practice, the population covariance matrix $\boldsymbol{\Phi}$ is unknown and has to be estimated in order to compute the feasible GLS estimator. If $\boldsymbol{\Phi}$ is replaced by a consistent estimator, the difference between GLS and feasible GLS (FGLS) vanishes asymptotically. This

means that FGLS is consistent and has the same asymptotic distribution as GLS. However, it is impossible to estimate a covariance matrix of dimensions $np \times np$ from np observations consistently. So restrictions have to be imposed and more price data are required. One restriction was already mentioned. $\mathbf{\Phi}$ is assumed to be a block-diagonal matrix with elements $\mathbf{\Sigma}_{\mathbf{u}_t}$ for $t = 1, \ldots, n$ along the diagonal. This reduces the number of unique elements to $np(p+1)/2$ parameters. Luckily, recent advances in automation and data management make high-frequency intraday data more available. This increases the number of price observations several orders of magnitude. The approach in this thesis is to estimate the stochastic $\mathbf{\Sigma}_{\mathbf{u}_t}$ with an integrated covariance matrix estimator applied to intraday returns of day t, $\widehat{\mathbf{\Sigma}}_t$. While this is the major step in making the GLS approach feasible, some problems remain. First, even based on high-frequency data, the estimate is unstable if the dimension is high. The shrinkage approach of Section 4.2 deals with this challenge. Second, high-frequency data are observed irregularly and with noise. Section 4.3 describes several state of the art estimators that are consistent in this setting. Note that even though every available tick may be used to compute $\widehat{\mathbf{\Sigma}}_t$, the resulting estimate is not the population covariance matrix, since traded prices are just random observations of the underlying continuous price process.

Estimators of the integrated covariance matrix of intraday log-returns, described in Section 4.3, consistently estimate $\mathbf{\Sigma}_{\mathbf{u}_t}$ even when applied to uncentered log-returns. In finite-samples, however, it has been found much preferable to center intraday returns by their unconditional mean over the day. This shall be done throughout the thesis. Finite-sample properties of such an estimator are studied by Laurent and Shi (2020). Accuracy of the proposed method may further be improved if the FGLS procedure was applied iteratively, using the previous step's prediction as the conditional mean estimate to center log-returns in the next step. Whether the FGLS estimator is more efficient than OLS in finite-samples depends on how much $\mathbf{\Sigma}_{\mathbf{u}_t}$ deviates from the scaled identity matrix, and how good the estimate of that matrix is. The efficacy of the procedure is demonstrated via Monte Carlo simulations of sample-paths exhibiting some of the stylized facts observed in financial markets in Section 5.

2.3 Estimating the asymptotic variance of the FGLS estimator

To estimate confidence intervals, an estimate of the asymptotic variance of the FGLS estimator is needed. It can be obtained as follows.

1. Compute the daily covariance matrix estimates

$$\widehat{\mathbf{\Sigma}}_t \quad \text{for } t = 1, \ldots, n.$$

2. Estimate $\mathbf{A}$ as its sample analogue

$$\hat{\mathbf{A}} = n^{-1} \sum_{t=1}^{n} \mathbf{Z}_t' \widehat{\mathbf{\Sigma}}_t^{-1} \mathbf{Z}_t.$$

3. With a slight abuse of notation (since, as n goes to infinity, the term will obviously converge to zero), the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is estimated as

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{A}}^{-1}/n.$$

3 Extension to a nonlinear conditional expectation model

3.1 The deep neural network

Let $\mathbf{Z} \in \mathbb{R}^{np \times k}$ denote a sequence of feature vectors $\mathbf{z}_i \in \mathbb{R}^k$ for $i = 1, \ldots, np$. A deep feedforward neural network, tracing back to Rosenblatt (1961), is a function $f : \mathbb{R}^k \to \mathbb{R}^{N_L}$ defined by the composition of L layers as follows:

$$f(\mathbf{z}_i) = \mathbf{W}_L \sigma \left(\mathbf{W}_{L-1} \sigma \left(\ldots \sigma \left(\mathbf{W}_1(\mathbf{z}_i) \right) \right) \right), \quad \mathbf{z}_i \in \mathbb{R}^k \,. \tag{24}$$

$\mathbf{W}_\ell$ is an affine linear map defined by a matrix $\mathbf{A}_\ell \in \mathbb{R}^{N_{\ell-1} \times N_\ell}$, called the weights, and an affine part $\mathbf{b}_\ell \in \mathbb{R}^{N_\ell}$, called the biases, via

$$\mathbf{W}_\ell(\mathbf{x}) = \mathbf{A}_\ell \mathbf{x} + \mathbf{b}_\ell. \tag{25}$$

The number of neurons per layer N_ℓ for $\ell = 1, \ldots, L$ determines the dimensionality of $\mathbf{W}_\ell$. For the regression problem considered here, $N_L = 1$. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ must be nonlinear, otherwise the model reduces to a linear regression model. Throughout the thesis, the Rectified Linear Unit (ReLU) of Nair and Hinton (2010) is chosen. The layers $\ell = 2, \ldots, L-1$, i.e., the layers between the input and the output layers, are called hidden layers.

3.2 The generalized mean squared error loss function

The regression equation is of the familiar error form

$$\boldsymbol{\Delta y} = f(\boldsymbol{Z}) + \mathbf{u}, \tag{26}$$

but in contrast to Equation (5), f is now a nonlinear function. The error $\mathbf{u}$ has, as previously discussed, an approximate factor structure. By applying the same linear transformation as in the linear model, the transformed regression equation can analogously be written as

$$\begin{aligned} \boldsymbol{\Phi}^{-1/2}\boldsymbol{\Delta y} &= \boldsymbol{\Phi}^{-1/2} f(\boldsymbol{Z}) + \boldsymbol{\Phi}^{-1/2}\mathbf{u} \\ \boldsymbol{\Delta y}^* &= f^*(\boldsymbol{Z}) + \mathbf{u}^*. \end{aligned} \tag{27}$$

The parameters of $\mathbf{W}_\ell$ for $\ell = 1, \ldots, L$ are tuned such that the loss function $\mathcal{L} : \mathbb{R}^{np \times N_L} \to \mathbb{R}$ is minimized. An often-used loss function for regression problems is the MSE loss function

$$\mathcal{L}^{(MSE)} = n^{-1}\mathbf{u}'\mathbf{u}. \tag{28}$$

In contrast to the linear case, however, an analytic solution is not available. Instead, some version of gradient descent is used to find an approximate optimum. The hope is that by minimizing the loss of the transformed error $\mathcal{L}(\mathbf{u}^*)$, instead of $\mathcal{L}(\mathbf{u})$, the model can be fitted more data efficiently and thus a larger number of predictive signals can be discovered with more accuracy, given a finite training sample. Simulation results in Section 5 show that this is indeed the case under the assumptions made there. The transformed loss function can be easily and efficiently implemented with the PyTorch package of Paszke et al. (2017) and its implementation of the Einstein summation convention as shown in Appendix A.1.

4 Covariance matrix estimation

4.1 Estimator instability in high dimensions

The sample covariance matrix and especially its inverse, the precision matrix, have bad properties in high dimensions, i.e., when the concentration ratio $c_n = \frac{p}{n}$ is not a small number. Then (i) the sample covariance matrix is estimated with a lot of noise since $\mathcal{O}(p^2)$ parameters have to be estimated with $pn = \mathcal{O}(p^2)$ observations (if n is of the same order of magnitude as p). And (ii), if the first principal component of returns, e.g., the overall market factor, explains a large part of their variance, the condition

number of the population covariance matrix $\mathbf{\Sigma}$ is already high. A high concentration ratio increases the dispersion of sample eigenvalues above and beyond the dispersion of population eigenvalues, increasing the condition number further and leading to an ill-conditioned sample covariance matrix.

Mathematically, as illustrated by Engle et al. (2019), the last point can be seen as follows. Define the population and sample spectral distribution, i.e., the cross-section cumulative distribution function that returns the proportion of population eigenvalues $\tilde{\lambda}_{i,n}$ and sample eigenvalues $\lambda_{i,n}$ smaller than x, respectively, as

$$H_n(x) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}_{\left\{x \leq \tilde{\lambda}_{i,n}\right\}} \quad \forall x \in \mathbb{R},$$
$$F_n(x) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}_{\{x \leq \lambda_{i,n}\}} \quad \forall x \in \mathbb{R}.$$

Under general asymptotics and its standard assumptions, p is a function of n and both p and n – not just n – go to infinity. Then, according to Silverstein (1995),

$$F_n(x) \xrightarrow{\text{a.s}} F(x),$$

where F denotes the nonrandom limiting spectral distribution. From the equality

$$\int_{-\infty}^{+\infty} x^2 dF(x) = \int_{-\infty}^{+\infty} x^2 dH(x) + c\left[\int_{-\infty}^{+\infty} x dH(x)\right]^2 \tag{29}$$

it can be seen that if the limiting concentration ratio c is greater than zero, the sample eigenvalue dispersion is inflated. The mean of the sample eigenvalues, however, is unbiased

$$\int_{-\infty}^{+\infty} x dF(x) = \int_{-\infty}^{+\infty} x dH(x). \tag{30}$$

The distortion of extreme eigenvalues is very large for high concentrations. But even for relatively small c, the eigenvalue dispersion can be high enough such that regularization is necessary to ameliorate instability. The mathematical intuition behind this can be seen from the Marchenko-Pastur law of Marchenko and Pastur (1967), which states that the limiting spectrum of the sample covariance matrix $\mathbf{S} = \mathbf{X}\mathbf{X}'/n$ of independent and identically distributed p-dimensional random vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)'$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, has density

$$f_c(x) = \begin{cases} \frac{1}{2\pi x c\sigma^2}\sqrt{(b-x)(x-a)}, & a \leq x \leq b \\ 0, & \text{otherwise,} \end{cases} \tag{31}$$

where the smallest and the largest eigenvalues are given by $a = \sigma^2(1 - \sqrt{c})^2$ and $b = \sigma^2(1+\sqrt{c})^2$, respectively, as $p, n \to \infty$ with $p/n \to c > 0$. To illustrate, say the interest lies in estimating a covariance matrix for 1000 stocks on daily data and suppose finite-sample behavior of eigenvalues is reasonably well approximated by the Marchenko-Pastur law. Using a rolling window of anything less than approximately 4 years of daily returns results in a singular covariance matrix. Widening the window to include 8 years of data, the concentration ratio would be approximately $c_n = 1/2$. Plugging this number into the equations for the smallest and largest eigenvalues, it can be seen that they are, respectively, 91% smaller and 191% larger than the population eigenvalues, which are equal to σ^2. Even for an extremely wide window of 40 years, the eigenvalues would still be underestimated by 53% for the smallest, and overestimated by 73% for the largest. To reduce the overdispersion enough such that the covariance matrix becomes well-conditioned, such a wide window would be necessary that inaccuracies due to non-stationarity would dominate, if the data are even available.

Empirically, the cross-section of stock returns exhibits correlations. Hence $\mathbf{\Sigma} \neq \sigma^2\mathbf{I}_p$. Ait-Sahalia and Xiu (2017), for example, find a low-rank factor structure plus a sparse industry-clustered residual covariance matrix, which may exacerbate the instability of the precision matrix since the eigenvalues of the population covariance matrix are already highly dispersed.

One approach to combat the eigenvalue overdispersion is to shrink the covariance matrix towards a shrinkage target which has a stable structure, such as the appropriately scaled identity matrix. This has the effect of pulling eigenvalues of the sample covariance matrix towards their grand mean, which is unbiased due to Equation (30), while keeping the sample eigenvectors unaltered. Estimators that retain the original eigenvectors while seeking better properties through modification of the eigenvalues are called rotation equivariant. The result is a reduction of dispersion of sample eigenvalues and thus the condition number. Since overfitting is known to increase the dispersion of sample eigenvalues, shrinking them has the effect of regularization.

4.2 Regularizing the eigenstructure

4.2.1 Rotation equivariance

The literature of eigenstructure regularized covariance estimation, especially for portfolio allocation problems, is predominantly focused on rotation equivariant estimators. To define the estimator, consider the sample covariance matrix $\mathbf{S}_n = \mathbf{X}_n\mathbf{X}_n'/n$ based on a sample of n i.i.d. observations $\mathbf{X}_n$ with zero-mean. For the sake of readability, the sample subscript is suppressed in the following text unless it is not clear from the context that the quantity depends on the sample. According to the spectral theo-

rem, the sample covariance matrix can be decomposed into $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$, where $\mathbf{\Lambda}$ is a diagonal matrix, whose elements are the eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ and $\mathbf{Q}$ is an orthogonal matrix, whose columns $[\mathbf{q}_1 \ldots \mathbf{q}_p]$ are the corresponding eigenvectors. A rotation equivariant estimator of the population covariance matrix $\mathbf{\Sigma}$ is of the form

$$\widehat{\mathbf{\Sigma}} = \mathbf{Q}\widehat{\mathbf{\Delta}}\mathbf{Q}' = \sum_{i=1}^{p} \widehat{\delta}_i \mathbf{q}_i \mathbf{q}_i', \tag{32}$$

where $\widehat{\mathbf{\Delta}}$ is a diagonal matrix with elements $\widehat{\boldsymbol{\delta}}$. The infeasible finite-sample optimal rotation equivariant estimator choses the elements of the diagonal matrix in Equation (32) as

$$\mathbf{d}^* = \left(d_1^*, \ldots, d_p^*\right) = \left(\mathbf{q}_1' \mathbf{\Sigma} \mathbf{q}_1, \ldots, \mathbf{q}_p' \mathbf{\Sigma} \mathbf{q}_p\right). \tag{33}$$

This estimator is an oracle since it depends on the unobservable population covariance matrix. In order to be feasible, $\widehat{\mathbf{\Delta}}$ has to be estimated from data. The linear and the nonlinear shrinkage estimator introduced in Section 4.2.3 and Section 4.2.4, respectively, estimate the elements of $\widehat{\mathbf{\Delta}}$, i.e., $\widehat{\boldsymbol{\delta}} = \left(\widehat{\delta}_1, \ldots, \widehat{\delta}_p\right) \in (0, +\infty)^p$, as a function of $\boldsymbol{\lambda}_n$. These two approaches differ in how many parameters this function takes. They both are elements of the rotation equivariant class, though. Within this framework, due to Stein (1975) and Stein (1986), lecture 4, rotations of the original data are deemed irrelevant for covariance estimation. Hence, the rotation equivariant covariance estimate based on the rotated data equals the rotation of the covariance estimate based on the original data. This characteristic distinguishes the class of rotation equivariant estimators from regularization schemes based on sparsity since a change of basis does generally not preserve the sparse structure of the matrix.

4.2.2 Loss function

To derive estimators that minimize a certain error, it must be established what this error should look like. The literature specifies several functions to describe the error that are tailored to the use case of the covariance matrix. One of the most universal loss functions of a covariance estimator is the squared Frobenius loss defined as

$$\mathcal{L}^{\mathrm{FR}}\left(\widehat{\mathbf{\Sigma}}, \mathbf{\Sigma}\right) = \frac{1}{p} \operatorname{tr}\left[\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)^2\right]. \tag{34}$$

The minimum variance (MV)-loss function is proposed by Engle et al. (2019) as a loss function that is appropriate for covariance matrix estimator evaluation for the problem of minimum variance portfolio allocations under a linear constraint and high-

dimensional asymptotic theory. It is defined as

$$\mathcal{L}^{\text{MV}}(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \frac{\text{tr}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\right)/p}{\left[\text{tr}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)/p\right]^2} - \frac{1}{\text{tr}\left(\boldsymbol{\Sigma}^{-1}\right)/p}. \tag{35}$$

This loss function can be interpreted as the true variance of the minimum variance portfolio constructed from the estimated covariance matrix.

4.2.3 Linear shrinkage

The most ubiquitous estimator of the rotation equivariant type is perhaps the linear shrinkage estimator of Ledoit and Wolf (2004). These authors propose a weighted average of the sample covariance matrix and the identity (or some other highly structured) matrix that is scaled such that the trace remains the same. The weight, or shrinkage intensity, ρ, is chosen such that the squared Frobenius loss is minimized by inducing bias but reducing variance. In Theorem 1, the authors show that the optimal ρ is given by

$$\rho = \beta^2 / \left(\alpha^2 + \beta^2\right) = \beta^2/\delta^2, \tag{36}$$

where $\mu = \text{tr}(\boldsymbol{\Sigma})/p$, $\alpha^2 = \|\boldsymbol{\Sigma} - \mu\mathbf{I}_p\|_F^2$, $\beta^2 = E\left[\|\mathbf{S} - \boldsymbol{\Sigma}\|_F^2\right]$, and $\delta^2 = E\left[\|\mathbf{S} - \mu\mathbf{I}_p\|_F^2\right]$. β^2/δ^2 can be interpreted as a normalized measure of the error of the sample covariance matrix. Shrinking the covariance matrix towards the μ-scaled identity matrix has the effect of pulling the eigenvalues towards μ. This reduces eigenvalue dispersion. In other words, the elements of the diagonal matrix in Equation (32) are chosen as

$$\widehat{\boldsymbol{\delta}}^{(l,o)} = \left(\widehat{d}_1^{(l,o)}, \ldots, \widehat{d}_p^{(l,o)}\right) = \left(\rho\mu + (1-\rho)\lambda_1, \ldots, \rho\mu + (1-\rho)\lambda_p\right). \tag{37}$$

In the current form, it is still an oracle estimator since it depends on unobservable quantities. To make it a feasible estimator, these quantities have to be estimated. To this end, define the grand mean as

$$\bar{\lambda} = \frac{1}{p}\sum_{i=1}^{p}\lambda_i. \tag{38}$$

The estimator for ρ is given by

$$\widehat{\rho} = \frac{b^2}{d^2}, \tag{39}$$

where $d^2 = \left\|\mathbf{S} - \bar{\lambda}\mathbf{I}_p\right\|_F^2$ and $b^2 = \min\left(\bar{b}^2, d^2\right)$ with $\bar{b}^2 = \frac{1}{n^2}\sum_{t=1}^{n}\left\|\mathbf{x}_t\mathbf{x}_t' - \mathbf{S}\right\|_F^2$, where $\mathbf{x}_t$ denotes the tth column of the observations matrix $\mathbf{X}$ for $t = 1, \ldots, n$. In order for this estimator to be consistent, the assumption that $\mathbf{X}$ is i.i.d with finite fourth moments must be satisfied. The feasible linear shrinkage estimator is then of form Equation (32) with the elements of the diagonal chosen as

$$\widehat{\boldsymbol{\delta}}^{(l)} = \left(\widehat{d}_1^{(l)}, \ldots, \widehat{d}_p^{(l)}\right) = \left(\widehat{\rho}\bar{\lambda} + (1-\widehat{\rho})\lambda_1, \ldots, \widehat{\rho}\bar{\lambda} + (1-\widehat{\rho})\lambda_p\right). \tag{40}$$

Hence, the linear shrinkage estimator[1] is given by

$$\widehat{\mathbf{S}} = \sum_{i=1}^{p} \widehat{d}_i^{(l)}\mathbf{q}_i\mathbf{q}_i'. \tag{41}$$

The result is a biased but well-conditioned covariance matrix estimate.

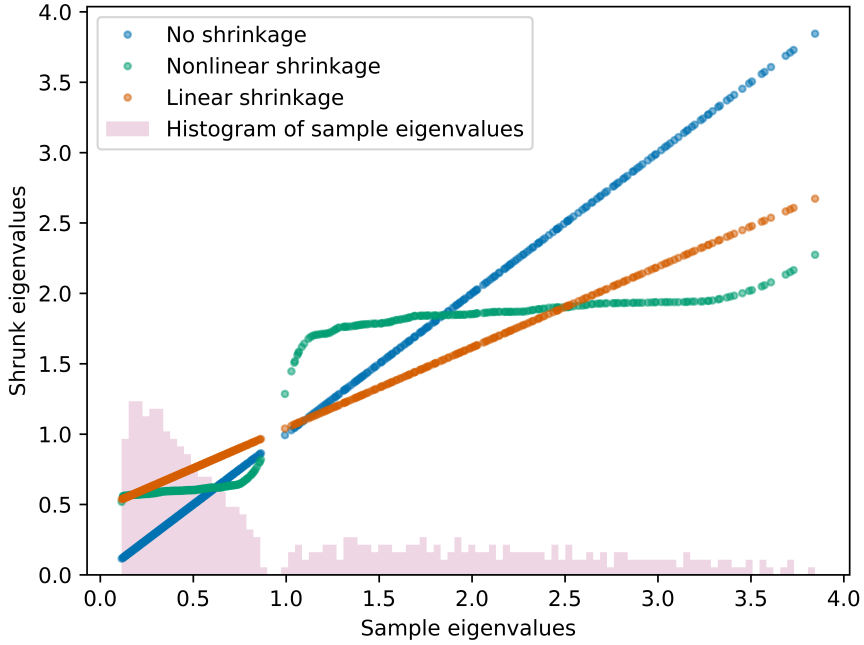
4.2.4 Nonlinear shrinkage

The idea of shrinkage covariance estimators is further developed by Ledoit and Wolf (2012) who argue that given a sample of size $\mathcal{O}(p^2)$, estimating $\mathcal{O}(1)$ parameters, as in linear shrinkage, is precise but too restrictive, while estimating $\mathcal{O}(p^2)$ parameters, as in the sample covariance matrix, is impossible. They argue that the optimal number of parameters to estimate is $\mathcal{O}(p)$. Their proposed nonlinear shrinkage estimator uses exactly p parameters, one for each eigenvalue, to regularize each eigenvalue with a specific shrinkage intensity individually. Linear shrinkage, in contrast, is restricted by a single shrinkage intensity, with which all eigenvalues are shrunk uniformly. Nonlinear shrinkage enables a nonlinear fit of the shrunk eigenvalues, which is appropriate when there are clusters of eigenvalues. In this case, it may be optimal to pull a small eigenvalue (i.e., an eigenvalue that is below the grand mean) further downwards and hence further away from the grand mean. Linear shrinkage, in contrast, always pulls a small eigenvalue upwards. Ledoit and Wolf (2018) find an analytic formula based on the Hilbert transform that nonlinearly shrinks eigenvalues asymptotically optimally with respect to the MV-loss function (as well as the quadratic Frobenius loss). The shrinkage function via the Hilbert transform can be interpreted as a local attractor. Much like the gravitational field extended into space by massive objects, eigenvalue clusters exert an attraction force that increases with the mass (i.e., the number of eigenvalues in the cluster) and decreases with the distance. If an eigenvalue λ_i has many neighboring eigenvalues slightly smaller than itself, the exerted force on λ_i will have large magnitude and downward direction. If λ_i has many neighboring eigenvalues

[1] This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

slightly larger than itself, the exerted force on λ_i will also have large magnitude but upward direction. If the neighboring eigenvalues are much larger or much smaller than λ_i, the magnitude of the force on λ_i will be small. The nonlinear effect this has on the shrunk eigenvalues can be seen in Figure 1. The linearly shrunk eigenvalues, on the other hand, follow a line. Both approaches reduce the dispersion of eigenvalues and hence deserve the name shrinkage.

Figure 1: 1500 variates are drawn from a zero-mean 500-dimensional multivariate-normal distribution with a diagonal covariance matrix, which has 300 eigenvalues equal to 0.5 and 200 eigenvalues equal to 2. The sample eigenvalues are shrunk by the linear and nonlinear shrinkage methods and plotted against the original sample eigenvalues. Depicted at the bottom is a histogram of the sample eigenvalues.



The authors assume that there exists a $n \times p$ matrix $\mathbf{A}$ of i.i.d. random variables with mean zero, variance one, and finite 16th moment, such that the matrix of observations can be written as $\mathbf{X} = \mathbf{A}\boldsymbol{\Sigma}^{1/2}$. Neither $\boldsymbol{\Sigma}^{1/2}$ nor $\mathbf{A}$ are observed on their own. This assumption might not be satisfied, however, if the data generating process is a factor model. A method that does not rely on this assumption is described in Section 4.2.5. Theorem 3.1 of Ledoit and Wolf (2018) states that, under their assumptions and general asymptotics, the MV-loss is minimized by an estimator of the form Equation (32), where the elements of the diagonal matrix are

$$\widehat{\boldsymbol{\delta}}^{(o,nl)} = \left(\widehat{d}_1^{(o,nl)}, \ldots, \widehat{d}_p^{(o,nl)}\right) = \left(d^{(o,nl)}(\lambda_1), \ldots, d^{(o,nl)}(\lambda_p)\right). \tag{42}$$

$d^{(o,nl)}(x)$ denotes the oracle nonlinear shrinkage function and it is defined as

$$d^{(o,nl)}(x) = \frac{x}{[\pi c x f(x)]^2 + [1 - c - \pi c x \mathcal{H}_f(x)]^2} \quad \forall x \in \text{Supp}(F), \tag{43}$$

where $\mathcal{H}_g(x)$ is the Hilbert transform. As per Definition 2 of Ledoit and Wolf (2018), it is defined as

$$\mathcal{H}_g(x) = \frac{1}{\pi} PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x} \quad \forall x \in \mathbb{R}, \tag{44}$$

which uses the Cauchy principal value, denoted as PV, to evaluate the singular integral in the following way:

$$PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x} = \lim_{\varepsilon \to 0^+} \left[\int_{-\infty}^{x-\varepsilon} g(t) \frac{dt}{t - x} + \int_{x+\varepsilon}^{+\infty} g(t) \frac{dt}{t - x} \right]. \tag{45}$$

It is an oracle estimator due to the dependence on the limiting sample spectral density f, its Hilbert transform $\mathcal{H}_f$, and the limiting concentration ratio c, which are all unobservable. Nevertheless, Equation (42) represents progress compared to Equation (33) since it no longer depends on the full population covariance matrix $\mathbf{\Sigma}$, but only on its eigenvalues. This reduces the number of parameters to be estimated from the impossible $\mathcal{O}(p^2)$ to the manageable $\mathcal{O}(p)$. To make the estimator feasible, unobserved quantities have to be replaced by statistics. A consistent estimator for the limiting concentration c is the sample concentration $c_n = p/n$. For the limiting sample spectral density f, the authors propose a kernel estimator. This is necessary, even though $F_n \overset{\text{a.s}}{\to} F$, since F_n is discontinuous at every λ_i, and thus its derivative f_n, which would've been the natural estimator for f, does not exist there.

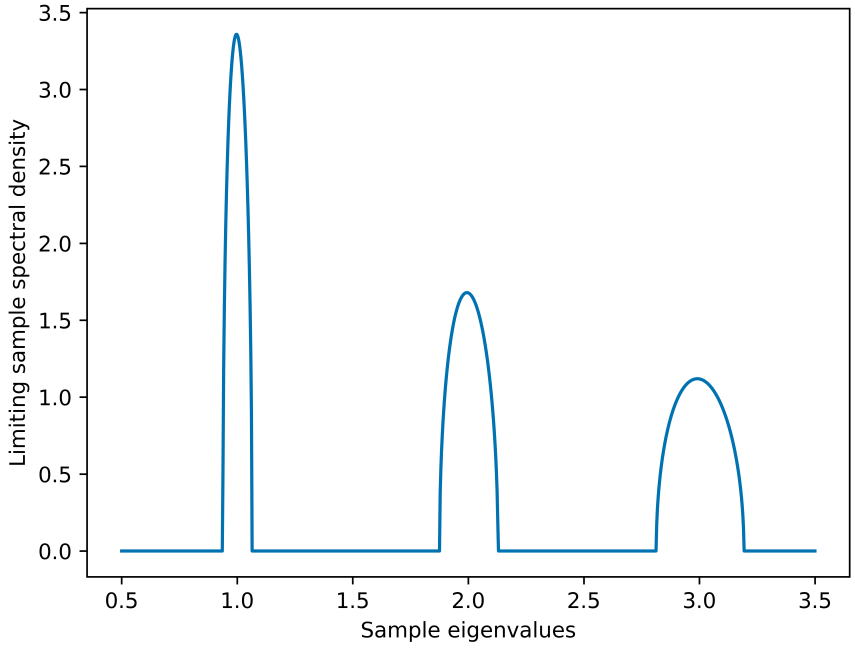
The chosen kernel estimator for the limiting sample spectral density is based on the Epanechnikov kernel of Epanechnikov (1969) with a variable bandwidth proportional to the eigenvalues $h_i = \lambda_i h$ for $i = 1, \ldots, p$, where the global bandwidth is set to $h = n^{-1/3}$. The reasoning behind the variable bandwidth choice can be intuited from Figure 2, which shows that the support of the limiting sample spectral distribution is approximately proportional to the eigenvalue. The Epanechnikov kernel is defined as

$$\kappa^{(E)}(x) = \frac{3}{4\sqrt{5}} \left[1 - \frac{x^2}{5} \right]^+ \quad \forall x \in \mathbb{R}. \tag{46}$$

The kernel estimators of f and $\mathcal{H}$ are thus

$$\widetilde{f}_n(x) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{h_i} \kappa^{(E)} \left(\frac{x - \lambda_i}{h_i} \right) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_i h} \kappa^{(E)} \left(\frac{x - \lambda_i}{\lambda_i h} \right) \quad \forall x \in \mathbb{R} \tag{47}$$

Figure 2: The limiting sample spectral density is an equally weighted mixture of Marchenko-Pastur laws with population eigenvalues 1, 2, 3, and c = 0.001



and

$$\mathcal{H}_{\tilde{f}_n}(x) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{h_i}\mathcal{H}_k\left(\frac{x-\lambda_i}{h_i}\right) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i h}\mathcal{H}_k\left(\frac{x-\lambda_i}{\lambda_i h}\right) = \frac{1}{\pi}PV\int\frac{\widetilde{f}_n(t)}{x-t}dt, \quad (48)$$

respectively. The feasible nonlinear shrinkage estimator is of form Equation (32), where the elements of the diagonal matrix are

$$\widehat{d}_i^{(nl)} = \frac{\lambda_i}{\left[\pi\frac{p}{n}\lambda_i\widetilde{f}_n\left(\lambda_i\right)\right]^2 + \left[1-\frac{p}{n}-\pi\frac{p}{n}\lambda_i\mathcal{H}_{\tilde{f}_n}\left(\lambda_i\right)\right]^2} \quad i=1,\ldots,p. \quad (49)$$

In other words, the feasible nonlinear shrinkage estimator[2] is

$$\widetilde{\mathbf{S}} = \sum_{i=1}^{p}\widehat{d}_i^{(nl)}\mathbf{q}_i\mathbf{q}_i'. \quad (50)$$

4.2.5 Nonparametric nonlinear shrinkage

A very different approach to nonlinear shrinkage is pursued by Abadir et al. (2014). These authors propose a nonparametric estimator based on a sample splitting scheme.

[2]This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

They split the data into two pieces and exploit the independence of observations across the splits to regularize the eigenvalues. Lam (2016) builds on their results and proposes a nonparametric estimator that is asymptotically optimal even if the population covariance matrix $\mathbf{\Sigma}$ has a factor structure. In the case of low-rank strong factor models, the assumption that each observation can be written as $\mathbf{x}_t = \mathbf{\Sigma}^{1/2}\mathbf{a}_t$ for $t = 1, \ldots, n$, where each $\mathbf{a}_t$ is a $p \times 1$ vector of i.i.d random variables a_{it} for $i = 1, \ldots, p$ with zero-mean and unit variance, is violated. Both Ledoit and Wolf (2012) and Ledoit and Wolf (2018) build on this assumption and their estimators are no longer optimal if it is not fulfilled. The proposed nonparametric eigenvalue-regularized covariance matrix estimator (NERCOME) starts by splitting the data into two pieces of size n_1 and $n_2 = n - n_1$. It is assumed that the observations are i.i.d with finite fourth moments, such that the statistics computed in the different splits are likewise independent of each other. The sample covariance matrix of the first partition is defined as $\mathbf{S}_{n_1} = \mathbf{X}_{n_1}\mathbf{X}'_{n_1}/n_1$. Its spectral decomposition is given by $\mathbf{S}_{n_1} = \mathbf{Q}_{n_1}\mathbf{\Lambda}_{n_1}\mathbf{Q}'_{n_1}$, where $\mathbf{\Lambda}_{n_1}$ is a diagonal matrix, whose elements are the eigenvalues $\boldsymbol{\lambda}_{n_1} = (\lambda_{n_1,1}, \ldots, \lambda_{n_1,p})$ and $\mathbf{Q}_{n_1}$ is an orthogonal matrix, whose columns $[\mathbf{q}_{n_1,1} \ldots \mathbf{q}_{n_1,p}]$ are the corresponding eigenvectors. Analogously, the sample covariance matrix of the second partition is defined by $\mathbf{S}_{n_2} = \mathbf{X}_{n_2}\mathbf{X}'_{n_2}/n_2$. Theorem 1 of Lam (2016) shows that

$$\widetilde{\mathbf{d}}_{n_1}^{(\text{NERCOME})} = \text{diag}(\mathbf{Q}'_{n_1}\mathbf{S}_{n_2}\mathbf{Q}_{n_1}) \tag{51}$$

is asymptotically the same as the finite-sample optimal rotation equivariant quantity $\mathbf{d}_{n_1}^*$ per Equation (33), based on the section n_1. The proposed estimator is thus of the form Equation (32), where the elements of the diagonal matrix are chosen according to Equation (51). In other words, the estimator[3] is given by

$$\widetilde{\mathbf{S}}_{n_1}^{(\text{NERCOME})} = \sum_{i=1}^{p} \widetilde{d}_{n_1,i}^{(\text{NERCOME})} \mathbf{q}_{n_1,i}\mathbf{q}'_{n_1,i}. \tag{52}$$

The author shows that this estimator is asymptotically optimal with respect to the Frobenius loss even under factor structure. However, it uses the sample data inefficiently since only one section is utilized for the calculation of each component. The natural extension is to permute the data and bisect it anew. With these sections an additional estimate is computed according to Equation (52). This is done M times

[3]This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

and the covariance matrix estimates are averaged to obtain

$$\widetilde{\mathbf{S}}_{n_1,M}^{(\text{NERCOME})} = \frac{1}{M}\sum_{j=1}^{M}\widetilde{\mathbf{S}}_{n_1,j}^{(\text{NERCOME})}. \tag{53}$$

The estimator depends on two tuning parameters, M and n_1. Higher M give more accurate results but the computational cost grows as well. The author suggests that no more than 50 iterations are generally needed for satisfactory results. n_1 is subject to regularity conditions. The author proposes to search over the contenders

$$n_1 = \left[2n^{1/2}, 0.2n, 0.4n, 0.6n, 0.8n, n - 2.5n^{1/2}, n - 1.5n^{1/2}\right] \tag{54}$$

and select the one that minimizes the following criterion inspired by Bickel and Levina (2008):

$$g(n_1) = \left\| \frac{1}{M}\sum_{j=1}^{M}\left(\widetilde{\mathbf{S}}_{n_1,j}^{(\text{NERCOME})} - \mathbf{S}_{n_2,j}\right)\right\|_F^2. \tag{55}$$

4.3 High-frequency data

Shifting the focus to a high-frequency setting, the considered log-price process is now continuous. The following standard model is assumed. Let $\left(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{0\le t\le T}, \mathbb{P}\right)$ be a filtered probability space on which the p-dimensional log-price process $\{\mathbf{X}_t\}_{0\le t\le T}$ is adapted, where $\mathbf{X}_t = \left(X_t^{(1)}, \ldots, X_t^{(p)}\right)'$, Ω denotes the sample space, $\mathcal{F}$ denotes the σ-algebra, $\{\mathcal{F}_t\}_{0\le t\le T}$ denotes the filtration, i.e., an increasing sequence of sub-σ-algebras of $\mathcal{F}$ over the interval [0,T], and $\mathbb{P}$ denotes the probability measure. It is assumed that $\mathbf{X}_t$ follows a diffusion process satisfying

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t d\mathbf{W}_t \quad t \in [0, T]. \tag{56}$$

The process $\{\mathbf{W}_t\}$ is a p-dimensional standard Brownian motion. The drift $\boldsymbol{\mu}_t \in \mathbb{R}^p$ and the volatility $\boldsymbol{\sigma}_t \in \mathbb{R}^{p\times p}$ are càdlàg. Analogously to the covariance matrix in the low-frequency setting, the integrated covariance matrix over the interval $[a, b] \subset [0, T]$ is defined as

$$\boldsymbol{\Sigma}(a, b) = \int_a^b \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t' dt. \tag{57}$$

The log-price of each asset is observed at discrete times and with market microstructure noise. The sequence of observation times of asset j is denoted as $t^{(j)} = \{t_0^{(j)} < t_1^{(j)} < \ldots < t_{n^{(j)}}^{(j)}\}$. Importantly, observation times are non-synchronous across assets, i.e.,

$t^{(j)} \neq t^{(k)}$ for $j \neq k$ in general. The observed log-price process of asset j is

$$Y_t^{(j)} = X_t^{(j)} + \epsilon_t^{(j)} \quad t \in t^{(j)}, \tag{58}$$

where $\epsilon_t^{(j)}$ is a noise process independent of $\mathbf{X}_t$ with mean 0. The integrated covariance matrix of the multivariate noise process over the interval $[a, b] \subset [0, T]$ is denoted as

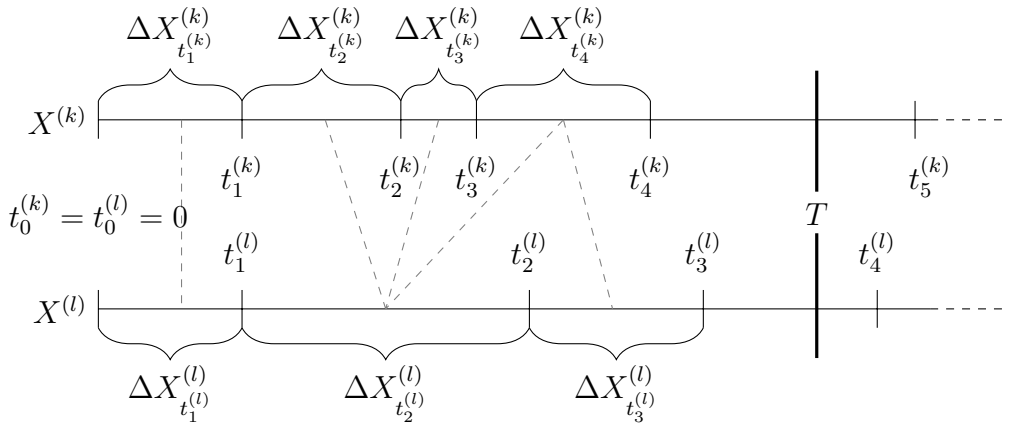
$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(a, b) = \int_a^b \boldsymbol{\sigma}_{\epsilon,t} \boldsymbol{\sigma}'_{\epsilon,t} dt. \tag{59}$$

$\{\boldsymbol{\epsilon}_t\}_{0 \leq t \leq T}$ is assumed to be adapted to $\{\mathcal{F}_t\}_{0 \leq t \leq T}$. Hence, the observed price process $\{\mathbf{Y}_t\}_{0 \leq t \leq T}$ is also adapted. The microstructure noise is the resulting process of an interplay of many effects. Among them are, for example, price discreteness and the bid–ask bounce. It is important to account for the variance due to the noise process. When the observation frequency is high, the variance of the noise process dominates the variance of the underlying process, and estimators that do not cancel the noise are severely biased.

4.3.1 Non-synchronicity

Non-synchronicity of observation times across assets is a problem for estimating the integrated covariance since the naively synchronized sample correlation has a strong bias towards zero as the sampling frequency increases. This effect is described by Epps (1979). A more quantitative statement about the bias is made in Equation (71) in Section 4.3.2. To combat the so-called Epps-effect, Hayashi and Yoshida (2005) propose the following estimator. The Hayashi-Yoshida (HY) estimator[4] for asset l and

Figure 3: A realization of tick-to-tick returns from $[0, T]$.



[4]This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

k is defined as

$$\widehat{\Sigma}_{kl}^{(HY)} = \sum_{i=1}^{n^{(k)}} \sum_{i'=1}^{n^{(l)}} \Delta X^{(k)}_{t_i^{(k)}} \Delta X^{(l)}_{t_{i'}^{(l)}} \mathbf{1}_{\left\{\left(t_{i-1}^{(k)}, t_i^{(k)}\right] \cap \left(t_{i'-1}^{(l)}, t_{i'}^{(l)}\right] \neq \emptyset\right\}}, \tag{60}$$

where $\Delta X^{(j)}_{t_i^{(j)}} = X^{(j)}_{t_i^{(j)}} - X^{(j)}_{t_{i-1}^{(j)}}$ denotes the jth asset tick-to-tick log-return over the interval spanned from $t_{i-1}^{(j)}$ to $t_i^{(j)}$ for $i = 1, \cdots, n^{(j)}$, where $n^{(j)} = |t^{(j)}| - 1$ denotes the number of tick-to-tick returns. In words, the estimator sums all cross products of tick returns that (perhaps partially) overlap in time. On first sight, it might appear that the HY estimator uses all the available data. However, due to the summability of log-returns, some returns effectively cancel in the sum. Figure 3 illustrates the products of returns that are summed in the HY-estimator for the period $t = 0$ to T by the dashed lines. For example, $\left(t_0^{(k)}, t_1^{(k)}\right] \cap \left(t_0^{(l)}, t_1^{(l)}\right] \neq \emptyset$ so that the product $\Delta X^{(k)}_{t_1^{(k)}} \Delta X^{(l)}_{t_1^{(l)}}$ goes into the sum. In contrast, $\left(t_0^{(k)}, t_1^{(k)}\right] \cap \left(t_1^{(l)}, t_2^{(l)}\right] = \emptyset$ so that the product $\Delta X^{(k)}_{t_1^{(k)}} \Delta X^{(l)}_{t_2^{(l)}}$ does not go into the sum. Note that

$$\Delta X^{(k)}_{t_2^{(k)}} \Delta X^{(l)}_{t_2^{(l)}} + \Delta X^{(k)}_{t_3^{(k)}} \Delta X^{(l)}_{t_2^{(l)}} = \Delta X^{(l)}_{t_2^{(l)}} (\Delta X^{(k)}_{t_2^{(k)}} + \Delta X^{(k)}_{t_3^{(k)}}) = \Delta X^{(l)}_{t_2^{(l)}} (X^{(k)}_{t_1^{(k)}} - X^{(k)}_{t_3^{(k)}}).$$

Since $(X^{(k)}_{t_1^{(k)}} - X^{(k)}_{t_3^{(k)}})$ does not overlap with any other returns, the observation at $t_2^{(k)}$ is effectively irrelevant for the HY-estimator. The authors show that this estimator is consistent and, for the special case of Brownian motion with observation times following a Poisson process, the convergence rate is $n^{-1/2}$. Hayashi and Yoshida (2004) show that, in this case, the estimator is asymptotically normal distributed. Unfortunately, this estimator is not applicable in the current setting, since it does not address the microstructure noise problem. In Equation (60) it was implicitly assumed that $\mathbf{X}_t$, and not the contaminated $\mathbf{Y}_t$, was directly observable. There are, however, approaches to correct for the resulting bias, e.g., by Nolte and Voev (2007). A more efficient method is proposed by Christensen et al. (2010). Their estimator is appealing since it uses all data, even the ticks that are irrelevant for the covariance, for the noise reduction. But first, a related but more direct approach is pursued in the next section.

4.3.2 Multi-scale realized covariance estimators

Realized variance estimators based on multiple scales exploit the fact that the proportion of the observed realized variance over a specified interval due to microstructure noise increases with the sampling frequency, while the realized variance of the true underlying process stays constant. The bias can thus be corrected by subtracting a high-frequency estimate, scaled by an optimal weight, from a medium-frequency estimate. The weight is chosen such that the large bias in the high-frequency estimate,

when scaled by the weight, is exactly equal to the medium bias and they cancel each other out as a result. How to determine the optimal time scales and the weights is subject of the following section. Zhang et al. (2005) propose a univariate estimator for the realized volatility that uses two scales to correct the bias due to microstructure noise. Zhang (2006) subsequently improves it by considering multiple scales. Zhang (2011) extends the two-scales estimator to a multivariate setting correcting for both, the Epps effect and noise. Tao et al. (2013), then, proposes a multi-scale estimator of the integrated covariance matrix in a high-dimensional setting.

Zhang et al. (2005) seek an estimator of the realized volatility of the (univariate) process $X^{(j)}$ over time period $[0, T]$ when it is observed irregularly and with microstructure noise as in Equation (58). In the stochastic calculus literature, this object is called quadratic variation and defined as

$$\langle X^{(j)}, X^{(j)} \rangle = \Sigma_{jj} = \int_0^T \sigma_t^2 dt. \tag{61}$$

The two-scales realized volatility (TSRV) estimator[5] is defined as

$$\widehat{\Sigma}_{jj}^{(TSRV)} = \left[Y^{(j)}, Y^{(j)} \right]^{(K)} - \frac{\bar{n}_K}{\bar{n}_J} \left[Y^{(j)}, Y^{(j)} \right]^{(J)}, \tag{62}$$

where

$$\left[Y^{(j)}, Y^{(j)} \right]^{(K)} = \frac{1}{K} \sum_{i=K}^{n} \left(Y^{(j)}_{\tau_i^{(j)}} - Y^{(j)}_{\tau_{i-K}^{(j)}} \right)^2, \tag{63}$$

denotes the realized variance of the Kth scale and the jth asset with K being a positive integer usually chosen much larger than 1 and $\bar{n}_K = (n - K + 1)/K$ and $\bar{n}_J = (n - J + 1)/J$. The authors show that if K is chosen on the order of $K = \mathcal{O}\left(n^{2/3}\right)$, this estimator is asymptotically unbiased, consistent, asymptotically normal and converges at rate $n^{-1/6}$.

By considering M time scales, instead of just two as above, Zhang (2006) improves the rate of convergence to $n^{-1/4}$. This is the best attainable rate of convergence in this setting, as shown by Gloter and Jacod (2001). The proposed multi-scale realized

[5]This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

volatility (MSRV) estimator[6] is defined as

$$\widehat{\Sigma}_{jj}^{(MSRV)} = \sum_{i=1}^{M} \alpha_i [Y^{(j)}, Y^{(j)}]^{(K_i)}, \tag{64}$$

where α_i are weights satisfying

$$\sum \alpha_i = 1 \quad \text{and} \quad \sum_{i=1}^{M} (\alpha_i / K_i) = 0. \tag{65}$$

The optimal weights for the chosen number of scales M, i.e., the weights that minimize the noise variance contribution, are given by

$$a_i = \frac{K_i \left(K_i - \bar{K}\right)}{M \operatorname{Var}(K)}, \tag{66}$$

where

$$\bar{K} = \frac{1}{M} \sum_{i=1}^{M} K_i \quad \text{and} \quad \operatorname{var}(K) = \frac{1}{M} \sum_{i=1}^{M} K_i^2 - \bar{K}^2.$$

If all scales are chosen, i.e., if $K_i = i$ for $i = 1, \ldots, M$, then $\bar{K} = (M+1)/2$ and $\operatorname{var}(K) = (M^2 - 1)/12$, and hence

$$a_i = 12 \frac{i}{M^2} \frac{i/M - 1/2 - 1/(2M)}{1 - 1/M^2}. \tag{67}$$

In this case, and if M is chosen optimally on the order of $M = \mathcal{O}(n^{1/2})$, the estimator is consistent at rate $n^{-1/4}$, as shown by the author in Theorem 4.

Zhang (2011) extends the analysis to a multivariate setting. Multivariate estimators require synchronization of the time series. This can be achieved via a grid. A grid is a subset of $[0, T]$ defined by

$$\mathcal{V} = \{v_0, v_1, \ldots, v_{\tilde{n}}\} \subset [0, T], \tag{68}$$

with $v_0 = 0$ and $v_{\tilde{n}} = T$, where $\tilde{n}$ is the sampling frequency, i.e., the number of grid intervals. Two prominent ways to specify the grid are (i) a regular grid, where $v_m - v_{m-1} = \Delta v$ for $m = 1, \ldots, \tilde{n}$, and (ii) a grid based on 'refresh times' of Barndorff-Nielsen et al. (2011), where the grid spacing is dependent on the observation times. If more than two assets are considered, refresh times can be further classified into 'all-refresh-times' and 'pairwise-refresh times'. Estimators based on pairwise-refresh times

[6] This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

use the data more efficiently but the integrated covariance matrix estimate might not be positive definite. The pairwise-refresh times $\mathcal{V}_p = \{v_0, v_1, \ldots, v_{\tilde{n}}\}$ can be obtained by setting $v_0 = 0$ and

$$v_m = \max\left\{\min\left\{t_i^{(k)} \in t^{(k)} : t_i^{(k)} > v_{m-1}\right\}, \min\left\{t_i^{(l)} \in t^{(l)} : t_i^{(l)} > v_{m-1}\right\}\right\}, \tag{69}$$

where $\tilde{n}$ is the total number of refresh times in the interval $(0, 1]$. This scheme is illustrated in Figure 4. The procedure has to be repeated for every asset pair. In contrast, the all-refresh time scheme uses a single grid for all assets, which is determined based on the trade time of the slowest asset of each grid interval. Hence, the spacing of grid elements can be much wider. This implies that estimators based on the latter scheme may discard a large proportion of the data, especially if there is a very slowly trading asset. An approach between those extremes is proposed by Hautsch et al. (2012). Empirically, Fan et al. (2012) find better accuracy of a suitably adjusted estimate of the integrated covariance matrix using the pairwise-refresh time method. In any case, as stated in Condition C1 of Zhang (2011), there has to be at least one observation time of each asset between any two grid elements. With that condition in mind, the mth 'previous tick time' of asset j is defined as

$$\tau_m^{(j)} = \max\left\{t_i^{(j)} \in t^{(j)} : t_i^{(j)} \leq v_m\right\}. \tag{70}$$

Zhang (2011) begins by quantifying the Epps effect. He shows in Theorem 1 that the finite-sample bias due to the asynchronicity of observed trading times for two assets is given by

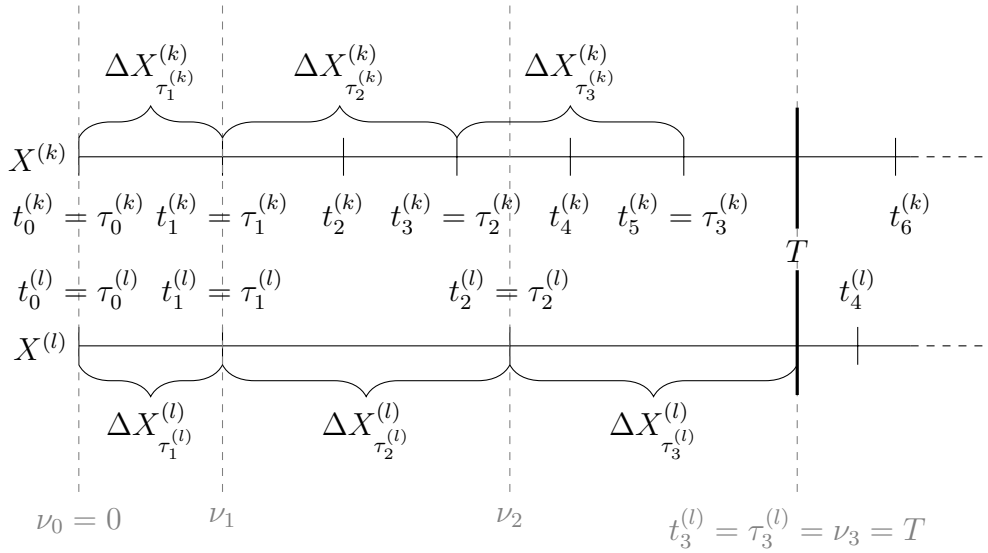
$$\text{Bias} = -\int_0^T \langle X^{(k)}, X^{(l)}\rangle_u' \mathrm{d}F(u) + O_p\left(\frac{1}{n^{(k)} + n^{(l)}}\right), \tag{71}$$

where $F(t) = \sum_{i:\max\left(\tau_i^{(k)}, \tau_i^{(l)}\right) \leq t} \left|\tau_i^{(k)} - \tau_i^{(l)}\right|$. The absolute value of this expression grows with the total absolute time between observations of both previous-tick times over all intervals. This means that the bias is larger in magnitude the less liquid the assets are, unless they are observed synchronously, in which case $F(t) = 0$.

The author proposes the two-scales realized covariance (TSCV) estimator[7] based on previous-tick times of asset k and l, which simultaneously corrects for the bias due to asynchronicity and the bias due to microstructure noise. The TSCV estimator is

[7]This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

Figure 4: A realization of returns using the refresh time scheme from $[0, T]$.



defined as

$$\widehat{\Sigma}_{kl}^{(TSCV)} = c\left(\left[Y^{(k)}, Y^{(l)}\right]^{(K)} - \frac{\bar{n}_K}{\bar{n}_J}\left[Y^{(k)}, Y^{(l)}\right]^{(J)}\right), \tag{72}$$

where

$$\left[Y^{(k)}, Y^{(l)}\right]^{(K)} = \frac{1}{K}\sum_{i=K}^{\tilde{n}}\left(Y^{(k)}_{\tau_i^{(k)}} - Y^{(k)}_{\tau_{i-K}^{(k)}}\right)\left(Y^{(l)}_{\tau_i^{(l)}} - Y^{(l)}_{\tau_{i-K}^{(l)}}\right) \tag{73}$$

denotes the realized covariance of asset k and l at scale K. $c = 1 + o_p\left(\tilde{n}^{-1/6}\right)$ is a small sample correction. K is again a positive integer usually chosen much larger than 1 and $\bar{n}_K = (\tilde{n} - K + 1)/K$ and $\bar{n}_J = (\tilde{n} - J + 1)/J$. The author shows that if $K = \mathcal{O}\left((n^{(k)} + n^{(l)})^{2/3}\right)$, this estimator is asymptotically unbiased, consistent, asymptotically normal, and converges at rate $\tilde{n}^{-1/6}$.

Tao et al. (2013) propose a multivariate multi-scale estimator. This estimator is defined by

$$\widehat{\mathbf{\Sigma}}^{(MSRC)} = \left(\sum_{m=1}^{M} a_m \left[\mathbf{Y}, \mathbf{Y}'\right]^{(K_m)}\right) + \zeta\left(\left[\mathbf{Y}, \mathbf{Y}'\right]^{(K_1)} - \left[\mathbf{Y}, \mathbf{Y}'\right]^{(K_M)}\right), \tag{74}$$

where

$$
\begin{aligned}
[\mathbf{Y}, \mathbf{Y}']^{(m)} &= \frac{1}{m} \sum_{s \in S^(m)} (\mathbf{Y}(s) - \mathbf{Y}(s-m))(\mathbf{Y}(s) - \mathbf{Y}(s-m))' \\
S(m) &= \left\{ s : t_s^{(j)}, t_{s-m}^{(j)} \in t^{(j)} \text{ for all } j \right\} \\
K_m &= N + m \\
a_m &= \frac{12(m+N)(m - M/2 - 1/2)}{M\left(M^2 - 1\right)} \\
\zeta &= \frac{(M+N)(N+1)}{(n+1)(M-1)}.
\end{aligned}
$$

$[\mathbf{Y}, \mathbf{Y}']^{(m)}$ is the realized covariance matrix for all p assets at scale m. N is a constant determining the scale size, introduced by Fan and Wang (2007) who choose it on the order of $n^{1/2}$. Adopting this constant, Tao et al. (2013) are able to prove rate $n^{-1/4}$ convergence if the integrated covariance matrix is sparse. Lam and Qian (unpublished paper), however, show that if sparsity is not given, N must be chosen on the order of $n^{2/3}$ and the estimator loses the optimal convergence rate. It then converges only at rate $n^{-1/6}$.

4.3.3 Kernel realized volatility matrix

Barndorff-Nielsen et al. (2011) propose a multivariate realized kernel estimator that smoothes the autocovariance operator and thereby achieves the optimal convergence rate in the multivariate setting with noise and asynchronous observation times. Incidentally, this estimator is similar in form to the heteroskedasticity- and autocorrelation-consistent estimator of Newey and West (1986), widely used in the statistics and econometrics literature to deal with heteroskedastic and autocorrelated noise. Observations are synchronized with the previously discussed refresh-time scheme. In addition, m observation are averaged at the beginning and at the end of the trading day to estimate the efficient price at these times. The authors call this 'jittering'. In practice, the effect of jittering is negligible but it is needed for proving consistency. The, with parameter m, jittered log-price vectors are denoted as $\mathbf{Y}^{(m)}(s)$ for $s = 1, \ldots, n - 2m + 1$. The kernel estimator[8] is defined by

$$
\widehat{\mathbf{\Sigma}}^{(KRVM)} = \boldsymbol{\gamma}^{(0)}\left(\mathbf{Y}^{(m)}\right) + \sum_{h=1}^{n-2m} \kappa\left(\frac{h-1}{H}\right)\left[\boldsymbol{\gamma}^{(h)}\left(\mathbf{Y}^{(m)}\right) + \boldsymbol{\gamma}^{(-h)}\left(\mathbf{Y}^{(m)}\right)\right], \tag{75}
$$

[8]This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

where

$$\boldsymbol{\gamma}^{(h)}\left(\mathbf{Y}^{(m)}\right) = \sum_{s=h+2}^{n-2m+1} \left(\mathbf{Y}^{(m)}(s) - \mathbf{Y}^{(m)}(s-1)\right)\left(\mathbf{Y}^{(m)}(s-h) - \mathbf{Y}^{(m)}(s-h-1)\right)', \quad h \geq 0 \tag{76}$$

and

$$\boldsymbol{\gamma}^{(h)}\left(\mathbf{Y}^{(m)}\right) = \boldsymbol{\gamma}^{(-h)}\left(\mathbf{Y}^{(m)}\right)', \quad h < 0. \tag{77}$$

$\boldsymbol{\gamma}^{(h)}$ is the hth realized autocovariance. $\kappa(\cdot)$ is the kernel function with bandwidth parameter H. It is assumed that (i) $\kappa(0) = 1$ and $\kappa'(0) = 0$, (ii) $\kappa(\cdot)$ is twice differentiable with continuous derivatives, and (iii) $\int_0^\infty \kappa(x)^2 dx$, $\int_0^\infty \kappa'(x)^2 dx$ and $\int_0^\infty \kappa''(x)^2 dx$ are finite. A slightly adjusted form of this estimator that is positive semidefinite is given by

$$\widehat{\boldsymbol{\Sigma}}^{(KRVM_{psd})} = \boldsymbol{\gamma}^{(0)}\left(\mathbf{Y}^{(m)}\right) + \sum_{h=1}^{n-2m} \kappa\left(\frac{h}{H}\right)\left[\boldsymbol{\gamma}^{(h)}\left(\mathbf{Y}^{(m)}\right) + \boldsymbol{\gamma}^{(-h)}\left(\mathbf{Y}^{(m)}\right)\right]. \tag{78}$$

This form requires the additional assumption $\int_{-\infty}^{\infty} \kappa(x)\exp(ix\lambda)dx \geq 0$ for all $\lambda \in \mathbb{R}$.

Choosing the right kernel function is important. The authors show, for example, that the estimator based on the Bartlett weight function is inconsistent. Instead, the Parzen kernel is suggested as a weight function that yields a consistent estimator and can be efficiently implemented. It is given by

$$\kappa^{(Parzen)}(x) = \begin{cases} 1 - 6x^2 + 6x^3 & 0 \leq x \leq \frac{1}{2} \\ 2(1-x)^3 & \frac{1}{2} \leq x \leq 1 \\ 0 & x > 1 \end{cases} \tag{79}$$

and shall be used to construct the estimator in this thesis. The bandwidth H must be on the order of $n^{3/5}$. The authors choose the scalar H as the average of optimal individual $H^{(j)}$

$$\bar{H} = p^{-1}\sum_{j=1}^{p} H^{(j)}, \tag{80}$$

where

$$H^{(j)} = c^* \xi_j^{4/5} n^{3/5} \tag{81}$$

with

$$c^* = \left\{\kappa''(0)^2/\kappa_\bullet^{0,0}\right\}^{1/5} \tag{82}$$

and

$$\xi_j^2 = \Sigma_{\epsilon,jj}/\Sigma_{jj}. \tag{83}$$

$\mathbf{\Sigma}_\epsilon$ and $\mathbf{\Sigma}$ denote, as previously defined, the integrated covariance matrix of the noise and the efficient return process, respectively. Here these quantities are understood over the interval under consideration. Hence, ξ_j^2 can be interpreted as the ratio of the noise variance and the return variance. c^* is a measure of the relative asymptotic efficiency of the kernel. For the Parzen kernel $c^* = 3.51$, as tabulated by the authors. Σ_{jj} may be estimated via a low-frequency estimator and $\Sigma_{\epsilon,jj}$ via a high-frequency estimator.

4.3.4 Preaveraging

Another approach to canceling microstructure noise is proposed by Podolskij et al. (2009). They use the fact that if the noise is i.i.d with zero mean, then averaging a rolling window of (weighted) returns diminishes the effect of microstructure noise on the variance estimate. The preaveraged log-returns based on the window-length K are given by

$$\bar{\mathbf{Y}}_i = \sum_{j=1}^{K-1} g\left(\frac{j}{K}\right) \Delta_{i-j+1}\mathbf{Y} \quad \text{for } i = K, \ldots, n, \tag{84}$$

where $\mathbf{Y}_i$ have been synchronized beforehand, for example by the refresh-time method described in Section 4.3.2. Note that the direction of the moving window has been reversed compared to the definition in Podolskij et al. (2009) to stay consistent within the thesis. The function g is a weighting function. A popular choice is

$$g(x) = \min(x, 1 - x). \tag{85}$$

Christensen et al. (2010) build on this idea and propose the modulated realised covariance (MRC) estimator[9]. This estimator is analogous to the realized integrated covariance estimator using preaveraged returns. It is thus of the form

[9]This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

$$\widehat{\mathbf{\Sigma}}^{(MRC)} = \frac{n}{n-K+2}\frac{1}{\psi_2 K}\sum_{i=K-1}^{n}\bar{\mathbf{Y}}_i\bar{\mathbf{Y}}_i', \tag{86}$$

where $\frac{n}{n-K+2}$ is a finite-sample correction and

$$\begin{aligned}\psi_1^k &= k\sum_{i=1}^{k}\left(g\left(\frac{i}{k}\right)-g\left(\frac{i-1}{k}\right)\right)^2,\\ \psi_2^k &= \frac{1}{k}\sum_{i=1}^{k-1}g^2\left(\frac{i}{k}\right).\end{aligned} \tag{87}$$

In this form, however, the estimator is biased. To correct for the bias, the authors propose the following adjusted estimator:

$$\widehat{\mathbf{\Sigma}}^{(MRC)} = \frac{n}{n-K+2}\frac{1}{\psi_2 k}\sum_{i=K-1}^{n}\bar{\mathbf{Y}}_i\bar{\mathbf{Y}}_i' - \frac{\psi_1}{\theta^2\psi_2}\hat{\mathbf{\Psi}}, \tag{88}$$

where

$$\hat{\mathbf{\Psi}} = \frac{1}{2n}\sum_{i=1}^{n}\Delta_i\mathbf{Y}\left(\Delta_i\mathbf{Y}\right)'. \tag{89}$$

The rate of convergence of this estimator is determined by the window-length K. The authors show that choosing $K = \mathcal{O}(\sqrt{n})$ delivers the best rate of convergence of $n^{-1/4}$. It is thus suggested to choose $K = \theta\sqrt{n}$, where θ can be calibrated from the data. The authors set $\theta = 1$. Hautsch and Podolskij (2013) suggest values between 0.4 (for liquid stocks) and 0.6 (for less liquid stocks). The bias correction in Equation (88) may result in an estimate that is not positive semi-definite. If positive semi-definiteness is essential, the bias-correction can be omitted. In this case, K should be chosen larger than otherwise optimal with respect to the convergence rate. Of course, the convergence rate is slower then. The optimal rate of convergence without the bias correction, i.e., of the estimator Equation (86) is $n^{-1/5}$, which is attained when $K = \theta n^{1/2+\delta}$ with $\delta = 0.1$. With preaveraging, the HY estimator of Equation (60) can be made robust to microstructure noise as well. It is then of the slightly adjusted form [10]

[10] This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

$$\widehat{\Sigma}_{kl,\theta}^{(HY)} = \frac{1}{(\psi_{HY} K)^2} \sum_{i=K}^{n^{(k)}} \sum_{i'=K}^{n^{(l)}} \bar{Y}_{t_i^{(k)}}^{(k)} \bar{Y}_{t_{i'}^{(l)}}^{(l)} \mathbf{1}_{\left\{\left(t_{i-K}^{(k)}, t_i^{(k)}\right] \cap \left(t_{i'-K}^{(l)}, t_{i'}^{(l)}\right] \neq \emptyset\right)} \tag{90}$$

where

$$\psi_{HY} = \frac{1}{K} \sum_{i=1}^{K-1} g\left(\frac{i}{K}\right).$$

The authors show that the preaveraged HY estimator has optimal convergence rate $n^{-1/4}$. Christensen et al. (2013) subsequently prove a central limit theorem for this estimator and show that it is robust to some dependence structure of the noise process. Since preaveraging is performed before synchronization, the estimator utilizes more data than other methods that cancel noise after synchronization. In particular, the preaveraged HY estimator even uses the observation $t_2^{(j)}$ in Figure 3, which, as previously shown, does not contribute to the covariance due to the log-summability.

4.4 Shrinking an integrated covariance matrix

Applying the regularization methods of Section 4.2 to integrated covariance matrices estimated with high-frequency data poses new challenges. First, observations generally are not i.i.d. This invalidates optimality results of the analytic formulas that build on this assumption. Second, pairwise estimators are even more unstable than their counterparts based on all-refresh times, of which daily close times are a special instance. Higher instability requires a higher shrinkage intensity. The optimal linear shrinkage intensity of Section 4.2.3, for example, estimates the error of the sample covariance matrix as the sample mean of the Frobenius error over all grid points. If a pairwise estimator is used, however, grid points are not synchronous across assets. This prevents the use of this particular formula with pairwise estimators. Similarly, the analytic nonlinear shrinkage formula requires a single sample size parameter n^*. In both cases, optimality of the shrinkage intensities is only given if observations are i.i.d. As previously discussed, the nonlinear shrinkage formula further excludes factor structure of the population covariance matrix. The NERCOME estimator, on the other hand, does not suffer from these limitations and allows for a straightforward extension into the high-frequency setting. Instead of splitting the sample into sets of regular observation time points, the sample is split based on time intervals. The nonparametric eigenvalue-regularized integrated covariance matrix estimator (NERIVE), proposed by Lam and Feng (2018), splits the sample into L partitions. The split points are denoted by

$$0 = \widetilde{\tau}_0 < \widetilde{\tau}_1 < \cdots < \widetilde{\tau}_L = T$$

and the lth partition is given by $(\widetilde{\tau}_{l-1}, \widetilde{\tau}_l]$. The integrated covariance estimator for the lth partition is given by

$$\widehat{\mathbf{\Sigma}}_l^{(NERIVE)} = \mathbf{Q}_{-l} \operatorname{diag}\left(\mathbf{Q}'_{-l}\widetilde{\mathbf{\Sigma}}_l\mathbf{Q}_{-l}\right)\mathbf{Q}'_{-l}, \tag{91}$$

where $\mathbf{Q}_{-l}$ is an orthogonal matrix depending on all observations over the full interval $[0, T]$ except the lth partition. The NERIVE estimator[11] over the full interval $[0, T]$ is then given by

$$\widehat{\mathbf{\Sigma}}^{(NERIVE)} = \sum_{l=1}^{L} \widehat{\mathbf{\Sigma}}_l^{(NERIVE)} = \sum_{l=1}^{L} \mathbf{Q}_{-l} \operatorname{diag}\left(\mathbf{Q}'_{-l}\widetilde{\mathbf{\Sigma}}_l\mathbf{Q}_{-l}\right)\mathbf{Q}'_{-l}. \tag{92}$$

$\widetilde{\mathbf{\Sigma}}$ is an integrated covariance estimator that corrects for asynchronicity and microstructure noise as described in Section 4.3. Importantly, NERIVE does not assume i.i.d. observations but weak dependence between the log-price process and the microstructure noise process within partition l and weak serial dependence of microstructure noise vectors given $\mathcal{F}_{-l}$. Similar to NERCOME, NERIVE allows for the presence of pervasive factors as long as they persist between refresh times.

Lam and Feng (2018) choose the TSRC for $\widetilde{\mathbf{\Sigma}}$. This choice is primarily made for the sake of tractability in the proofs. In this thesis, an ensembled pairwise integrated covariance (EPIC) estimator[12] is proposed. This estimator aims to produce an estimate with better finite-sample properties by combining the MSRC estimator, the MRC estimator, the preaveraged HY estimator, and the KRVM estimator, each pairwise computed. It is defined by

$$\widehat{\mathbf{\Sigma}}^{(EPIC)} = \sum_{i=1}^{4} \left[((\alpha_i - \beta_i)\mathbf{I}_p + \beta_i\mathbf{J}_p) \odot \left(\widehat{\mathbf{\Sigma}}^{(MSRC)}\mathbf{1}_{\{i=1\}} + \widehat{\mathbf{\Sigma}}^{(MRC)}\mathbf{1}_{\{i=2\}} + \widehat{\mathbf{\Sigma}}_\theta^{(HY)}\mathbf{1}_{\{i=3\}} + \widehat{\mathbf{\Sigma}}^{(KRVM)}\mathbf{1}_{\{i=4\}}\right)\right], \tag{93}$$

where $\odot$ denotes the Hadamard product, $\mathbf{I}_p$ is the identity matrix, $\mathbf{J}_p$ is a p-dimensional square matrix of ones, α is a 4-dimensional weight vector with which the diagonal elements of the covariance matrix are weighted and β is a 4-dimensional weight vector with which the off-diagonal elements of the covariance matrix are weighted. The weights of both α and β must each sum to one. Since the preaveraged HY estimator of the off-diagonal elements is more sample efficient than the MSRC, the MRC and the KRVM estimator, it makes sense to overweight it.

[11] This estimator is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

[12] see footnote 11.

5 Monte Carlo evidence

In this section, high-frequency estimators are evaluated in finite-samples. They are then used to verify efficiency gains of the FGLS estimator and the neural network with the generalized MSE loss function.

5.1 Finite sample properties of the integrated covariance estimators

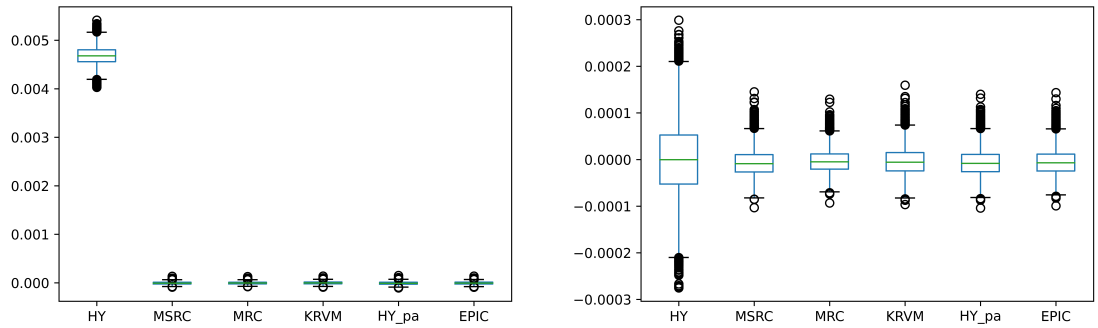
First, a Monte Carlo study is performed to evaluate the finite-sample properties of the estimators described in Section 4.3. For each run, 23,400 two-dimensional Gaussian log-returns are simulated with mean $\mathbf{0}$ and constant integrated covariance matrix $\mathbf{\Sigma}$ with diagonal elements given by $\Sigma_{0,0} = \Sigma_{1,1} = 0.2^2/250$ and off-diagonal elements given by $\Sigma_{0,1} = \Sigma_{1,0} = 0.8\Sigma_{0,0}$. Log-prices are computed by cumulative summation and sampled with probability 0.1. For each asset j, an independent Gaussian microstructure noise process with mean zero and standard deviation $\sigma_{\epsilon_t}^{(j)} = 0.1/100X_t^{(j)}$ is generated. The log-price process is observed according to Equation (58). Hence, on average, one traded price is observed every 6 seconds with noise having a standard deviation of 10 cents for a \$100 stock. The simulation results are summarized in Table 1.

Table 1: RMSE: 10,000 sample-paths of a two-dimensional price process are simulated second-by-second over one trading day. The integrated covariance matrix is estimated element-wise by the standard HY($\theta = 0$), MRC($\theta = 1$), MSRC($M = n^{1/2}$), KRVM($H = n^{3/5}$), HY($\theta = 1/2$), and EPIC, being the element-wise simple average of the MSRC, MRC, KRVM, and preaveraged HY estimators. The table shows the root mean squared error (RMSE) for the diagonal and off-diagonal element estimates.

	HY	MSRC	MRC	KRVM	$\text{HY}_{\theta=1/2}$	EPIC
$\Sigma_{0,0}$	0.0046843	0.0000269	0.0000254	0.0000276	0.0000310	0.0000274
$\Sigma_{0,1}$	0.0000785	0.0000288	0.0000249	0.0000296	0.0000282	0.0000275

Boxplots of the errors of estimates for $\Sigma_{0,0}$ and $\Sigma_{0,1}$ are shown in Figure 5 on the left-hand side and right-hand side, respectively. The standard HY estimator has poor performance. Its integrated variance estimate is significantly biased due to the relatively large noise contamination. The MRC, MSRC, KRVM, and preaveraged HY estimator all have satisfactory results. For comparison, the RMSE for the integrated variance and covariance using the standard integrated covariance estimator on a 5-minute fixed grid is 0.0001671 and 0.0000414, respectively. When computational costs are no constraint, ensembling according to Equation (93) may reduce model risk. In the experiment described above, the RMSE of the integrated covariance estimate

Figure 5: Errors: 10,000 sample-paths of a two-dimensional price process are simulated second-by-second over one trading day. The integrated covariance matrix is estimated element-wise by the standard HY($\theta = 0$), MRC($\theta = 1$), MSRC($M = n^{1/2}$), KRVM($H = n^{3/5}$), HY($\theta = 1/2$), and EPIC, being the element-wise simple average of the MSRC, MRC, KRVM, and preaveraged HY estimators. The figure shows the errors for the estimates of the diagonal (left) and off-diagonal elements (right) of the integrated covariance matrix.



(a) The integrated variance estimates.

(b) The integrated covariance estimates.

could not be reduced below the minimum of any single estimator alone. However, as documented in Table 2, this procedure does improve results in higher dimensions.

5.2 Finite sample properties of feasible GLS estimators

5.2.1 A linear conditional expectation specification

This section compares the GLS, the FGLS, and the OLS estimator efficiency in finite-samples. Sample-paths are drawn from a Universe object [13]. Stocks follow a single-factor model, they belong to industries and have an idiosyncratic component. The factor returns, industry returns and idiosyncratic returns are drawn from the generalized autoregressive conditional heteroskedasticity of order (1,1) model (GARCH(1, 1)) of Bollerslev (1986) with specification[14] $[\sigma_0^2, \mu, \alpha, \beta, \omega] = [0.1^2/250/6.5/60,\ 0,\ 0.0199,\ 0.98,\ \sigma_0^2/(1-\alpha-\beta)]$ each. There are 5 industries and each stock belongs to exactly one industry. The exposure to the market factor is determined by $\beta_{M,j,t}$, which ranges from 1 to 3 with constant increments for $j = 1, \ldots, p$ and is constant through time. Stocks are linearly predictable by a single feature, which is simulated at the daily frequency. The feature is drawn from a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $0.01\mathbf{I}_p$. Ticks are simulated every minute around the clock but sampled only within market hours and only a proportion specified by the liquidity parameter $\tilde{\ell}^{(j)}$. This results in a correlated sample of asynchronously and noisily

[13] The class defining this object is implemented in the hfhd Python library available at https://github.com/jpwoeltjen/hfhd. It is documented at https://hfhd.readthedocs.io/.

[14] This model is implemented in the hfhd Python library. See footnote 13.

observed stock returns with daily conditional expectation determined by the feature in proportion to the true coefficient β. For this simulation, β is chosen to be 0.1. A typical integrated covariance matrix of simulated returns is depicted in Figure 7 in Appendix B.1. The GLS, FGLS, and OLS estimator predict the next daily open-to-close return with the latest observable feature. The GLS estimator uses the ground truth conditional and unconditional integrated covariance matrices of the true residual to transform the model. The FGLS estimator uses observed (demeaned) intraday log-prices and the estimators of Section 4 to estimate the conditional integrated covariance matrix of the residual for each day.

Table 2 displays the results of the main simulation study. The GLS procedure can substantially reduce the RMSE of the parameter estimate compared with OLS. The estimator based on the conditional integrated covariance matrix works better than that based on the unconditional one. This is, of course, expected and these statistics mainly serve as benchmarks in the evaluation of the feasible estimators. When prices are observed regularly and without microstructure noise, the standard realized covariance (RC) estimator outperforms all other estimators since, in this case, it is unbiased and has a faster rate of convergence. If prices are observed non-synchronously and with microstructure noise, the MSRC, MRC, preaveraged HY and EPIC estimators are generally better.

The analytic nonlinear shrinkage estimator needs adjustment of the parameter n^*. If n^* is chosen as the effective sample size as proposed by its authors, the estimates are often ill-conditioned since noise and non-synchronicity invalidate their optimality results. Here, n^* is chosen ad-hoc as $\sqrt{\bar{n}}$, where $\bar{n}$ is the average number of prices observed over all assets. A more rigorous approach to choosing n^* in this setting is of great interest but beyond the scope of the thesis. The same applies to the linear shrinkage intensity, which is set to 0.8.

The nonparametrically eigenvalue-regularized (NER) estimates may be ill-conditioned if the sample size per partition falls below the by Lam and Feng (2018) recommended minimum of $\max(100, p)$. To reduce this risk, the daily sample is split into a modest number of subsamples. This number is chosen to be 4, which seems to work well. Though the probability of an ill-conditioned estimate is then rather small, the effect an outlier has on the variance of the estimate is large. Ensembling and averaging over multiple days can reduce this probability further but some risk remains. To overcome this problem, estimates with an abnormally high condition number are further shrunk iteratively by linear shrinkage until the condition number falls below some threshold. This procedure shall be called condition number targeting. The threshold is chosen to be 800. For reference, the condition number of the true unconditional integrated covariance matrix is 454.95. As can be seen in Table 2d, this procedure successfully

increases stability.

The FGLS model based on the nonparametric eigenvalue-regularized EPIC estimates with condition number targeting, averaged over two days, is the best performing feasible model in this simulation, achieving even better performance than the GLS model based on the true unconditional integrated covariance matrix of the true residual. Compared to OLS, the RMSE of the parameter estimate is reduced by more than 50%. Realizations of estimates are depicted in Figure 8 in Appendix B.2. These boxplots demonstrate the unbiasedness of the FGLS estimators and the successful regularization of the integrated covariance matrix estimates via NER when combined with condition number targeting even when the sample size per partition is small, leading to more accurate estimates.

Table 2: The root mean squared error (RMSE) of $\hat{\beta}$, computed with various least squares estimators over 1000 sample paths. 2 days of minutely price data are simulated. $\beta = 0.1$. GLS is performed with the true conditional covariance matrix $\mathbf{\Sigma}_{\mathbf{u}_t}$, the over 2 days averaged true conditional covariance matrix $\overline{\mathbf{\Sigma}}_{\mathbf{u}_t}$, and the true unconditional covariance matrix $\mathbf{\Sigma}_{\mathbf{u}}$. Feasible GLS is performed with the RC, MSRC, MRC, KRVM, preaveraged HY and EPIC estimators. Parameters of the covariance estimators are half of the exact order recommended by the respective authors. The feasible covariance estimates are computed (i) independently for each of the 2 days and (ii) as the simple average of both day's estimates. The smallest RMSE of the feasible estimators is highlighted in **bold.**

(a) Low-dimension: 10 stocks non-synchronously observed ($\tilde{\ell}^{(j)} = 0.8$) with microstructure noise ($\sigma_{\epsilon_t}^{(j)} = 0.1/100 X_t^{(j)}$).

	1 day				2 days			
	Raw	LS	NLS	NER	Raw	LS	NLS	NER
OLS	0.03769	-	-	-	-	-	-	-
$\mathbf{\Sigma}_t$	0.01585	-	-	-	-	-	-	-
$\overline{\mathbf{\Sigma}}_t$	0.01686	-	-	-	-	-	-	-
$\mathbf{\Sigma}$	0.01983	-	-	-	-	-	-	-
RC	0.02175	0.02906	0.02429	0.02277	0.02183	0.02963	0.02417	0.02287
MSRC	3.54427	0.02447	0.29719	0.11615	0.25006	0.02474	0.17748	0.09546
MRC	1.11332	0.02450	0.57164	0.06297	0.10634	0.02484	0.40039	0.02521
KRVM	1.50688	0.02469	1.27517	0.04344	0.19250	0.02487	0.22956	0.02231
HY	0.63066	0.02479	1.18028	0.06884	0.12119	0.02485	2.18458	0.02009
EPIC	0.63753	0.02455	0.37218	0.02259	0.25917	0.02481	0.50474	**0.01954**

(b) High-dimension: 100 stocks synchronously observed ($\tilde{\ell}^{(j)} = 1$) without microstructure noise ($\sigma_{\epsilon_t}^{(j)} = 0$).

OLS	0.01104	-	-	-	-	-	-	-
$\mathbf{\Sigma}_t$	0.00317	-	-	-	-	-	-	-
$\overline{\mathbf{\Sigma}}_t$	0.00354	-	-	-	-	-	-	-
$\mathbf{\Sigma}$	0.00456	-	-	-	-	-	-	-
RC	0.00370	0.00493	0.00427	**0.00356**	0.00381	0.00489	0.00420	0.00377
MSRC	0.31012	0.00513	0.00455	0.00432	2.46654	0.00498	0.00442	0.00427
MRC	3.08513	0.00507	0.00447	0.00420	0.00625	0.00495	0.00438	0.00421
KRVM	0.21107	0.00520	0.00467	0.00431	0.45644	0.00503	0.00450	0.00434
HY	5.39490	0.00534	0.00484	0.00456	0.82055	0.00508	0.00459	0.00442
EPIC	1.01439	0.00515	0.00453	0.00418	0.12111	0.00500	0.00442	0.00422

(c) High-dimension: 100 stocks non-synchronously observed ($\tilde{\ell}^{(j)} = 0.8$) with microstructure noise ($\sigma_{\epsilon_t}^{(j)} = 0.1/100 X_t^{(j)}$).

OLS	0.01068	-	-	-	-	-	-	-
$\mathbf{\Sigma}_t$	0.00353	-	-	-	-	-	-	-
$\overline{\mathbf{\Sigma}}_t$	0.00375	-	-	-	-	-	-	-
$\mathbf{\Sigma}$	0.00489	-	-	-	-	-	-	-
RC	0.17246	0.00579	0.00581	0.00555	0.00674	0.00574	0.00568	0.00548
MSRC	0.73966	0.00540	0.00496	0.61913	2.74424	0.00521	0.00481	0.01637
MRC	0.54219	0.00534	0.00495	0.04803	1.18245	0.00521	0.00483	0.02166
KRVM	1.52750	0.00548	0.00497	0.03263	0.51311	0.00524	0.00479	0.00578
HY	0.77265	0.00567	0.00507	0.02297	0.82621	0.00528	0.00481	0.00517
EPIC	2.84506	0.00539	0.00487	0.02305	0.77550	0.00521	**0.00474**	0.00485

Table 2: *continued.*

(d) High-dimension: 100 stocks non-synchronously observed ($\tilde{\ell}^{(j)} = 0.8$) with microstructure noise ($\sigma_{\epsilon_t}^{(j)} = 0.1/100 X_t^{(j)}$). With condition number targeting ($threshold = 800$).

	1 day				2 days			
	Raw	LS	NLS	NER	Raw	LS	NLS	NER
OLS	0.01183	-	-	-	-	-	-	-
$\mathbf{\Sigma}_t$	0.00329	-	-	-	-	-	-	-
$\overline{\mathbf{\Sigma}}_t$	0.00356	-	-	-	-	-	-	-
$\mathbf{\Sigma}$	0.00463	-	-	-	-	-	-	-
RC	0.55374	0.00698	0.00569	0.00541	0.00695	0.00601	0.00551	0.00528
MSRC	0.89018	0.00674	0.00500	0.00536	0.98717	0.00522	0.00473	0.00475
MRC	6.58552	0.00671	0.00492	0.00514	22.42251	0.00491	0.00473	0.00467
KRVM	10.43589	0.00730	0.00507	0.00524	2.27015	0.00915	0.00472	0.00477
HY	1.96412	0.01038	0.00513	0.00534	1.06475	0.03495	0.00474	0.00481
EPIC	2.16646	0.00516	0.00489	0.00490	1.95265	0.00471	0.00464	**0.00460**

5.2.2 A nonlinear conditional expectation specification

Next, a nonlinear conditional expectation specification is simulated. Stocks follow the same process as described in the linear case but now their conditional expectation is given by

$$\begin{aligned} \mathrm{E}[\Delta y_{j,t} | z_{j,t,1}, z_{j,t,2}, z_{j,t,3}] &= \beta_1 z_{j,t,1} + \beta_2 z_{j,t,2} + \beta_3 z_{j,t,3} \\ &\quad + \beta_4 z_{j,t,1} z_{j,t,2} + \beta_5 z_{j,t,1} z_{j,t,3} + \beta_6 z_{j,t,2} z_{j,t,3} \\ &\quad + \beta_7 z_{j,t,1} z_{j,t,2} z_{j,t,3}, \end{aligned} \tag{94}$$

with $\beta_1 = \beta_3 = \beta_3 = \nu 0.005$ and $\beta_4 = \beta_5 = \beta_6 = \beta_7 = \nu 0.01$, where ν is a constant determining the strength of predictability. It is chosen as 1 and 2 for the cases of weak and strong predictability in the simulation, respectively. $z_{j,t,1}$, $z_{j,t,2}$, and $z_{j,t,3} \sim$ Normal $(0, 0.1)$ are predictive features with daily observation frequency. A neural network with 2 hidden layers each with 100 neurons, batch normalization of Ioffe and Szegedy (2015), and dropout of Srivastava et al. (2014) ($p^{(do)} = 0.2$) is fitted for 1000 epochs. The Adaptive Moment Estimation (Adam) optimizer of Kingma and Ba (2014) is used with hyperparameters $\alpha^{(Adam)} = 0.1$, $\beta_1^{(Adam)} = 0.9$, $\beta_2^{(Adam)} = 0.999$, $\epsilon = 10^{-8}$, and $weight\ decay = 0$ (e.g. Krogh and Hertz (1992)). Weights are initialized according to the Kaiming uniform initialization of He et al. (2015). The activation function of each hidden layer is ReLU. The training set and test set are each 5 days. The performance metric is the information ratio (IR) of the mean-variance efficient portfolio constructed with the predictions as conditional mean estimate and the true conditional integrated covariance of each day as the covariance matrix. The (annualized) information ratio

is defined as

$$IR = \sqrt{252}\frac{\overline{r}}{std(r)}, \tag{95}$$

where $\overline{r}$ and $std(r)$ are the sample mean and standard deviation, respectively, of daily portfolio returns $\{r_t\}_{t\in\{1,\dots,n\}}$. r_t is calculated with daily rebalancing according to the mean-variance efficient weight vector, originating from Markowitz (1952), $r_t = \boldsymbol{\Delta}\mathbf{y}_t\boldsymbol{w}_t'$. Here, the weights are standardized by their L1 norm and thus given by

$$\boldsymbol{w}_t = \frac{\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\widehat{\boldsymbol{\Delta}\mathbf{y}}_t}{\mathbf{1}_p' \operatorname{abs}(\boldsymbol{\Sigma}_{\mathbf{u}_t}^{-1}\widehat{\boldsymbol{\Delta}\mathbf{y}}_t)}. \tag{96}$$

The information ratio has intuitive appeal as a metric since it is closely related to the Sharpe ratio of Sharpe (1966), widely used to assess trading strategy performance, while still isolating the conditional mean model performance by using the same covariance matrix in the construction of the weights for all estimators. The focus lies less on the level of the information ratios as it depends on the amount of training data, model architecture, and hyperparameters but on the difference between the IR achieved with and without the transformation applied to the error. As can be seen from Table 3, without the transformation, the model was not able to to discover any predictive signal, resulting in an insignificant IR, whereas the transformed model achieves a statistically significant mean IR of 1.054954 (p-value = 0.0137). The IRs based on the transformed model have greater mean than that of the standard model at the 2.5% significance level, according to a one-sided t-test. In the simulations with strongly predictive features, both the standard and transformed models achieve statistically highly significant IR means. Still, the treatment improves the IR by an economically meaningful and statistically highly significant amount (p-value = 0.00081). The feasible model based on the NER_EPIC estimator captures most of the theoretical improvement, evidenced by the statistically insignificant difference in means of its IRs and the IRs of the infeasible model (p-value = 0.342995).

6 Empirical findings

In this section, the predictive power of the proposed generalized models are compared with their standard counterparts using real-world historic stock market data. Based on the theoretical results of Section 2 and Section 3, the MC evidence of Section 5, and the intuition that the conditional expectation is likely a nonlinear function of the features, the following predictions are made: (i) The FGLS estimator should perform better than the OLS estimator, (ii) the neural network minimizing the GMSE loss function

Table 3: This table displays the information ratio (IR) of the mean-variance efficient portfolio constructed from the predictions and the true conditional covariance matrix of the true residual with daily rebalancing. A feedforward neural network is fitted to minimize the MSE, the GMSE based on $\widehat{\mathbf{\Sigma}}_t^{-1/2}$ computed via the nonparametrically eigenvalue-regularized (NER) EPIC estimator for each day, and the GMSE based on the true conditional integrated covariance matrix of the true residual for each day. The models are trained on 5 days worth of training data and evaluated on 5 days worth of test data for 100 stocks. For reference, the performance of the ground truth model with $\widehat{\Delta y_{j,t}} = \mathrm{E}\left[\Delta y_{j,t} \mid z_{j,t,1}, z_{j,t,2}, z_{j,t,3}\right]$ is tabulated as well. The simulation is repeated for $M = 1000$ iterations.

		MSE	$GMSE_{NER\_EPIC}$	$GMSE_{\mathbf{\Sigma}_{\mathbf{u}_t}}$	Ground Truth
$\nu = 1$	$\overline{IR}$	0.036527	1.054954	1.014555	10.597263
	$\mathrm{std}(IR)/\sqrt{M}$	0.310837	0.340965	0.347377	0.365767
$\nu = 2$	$\overline{IR}$	1.342032	2.981034	3.190057	23.023675
	$\mathrm{std}(IR)/\sqrt{M}$	0.351061	0.382461	0.347750	0.521955

should perform better than the neural network minimizing the MSE loss function, (iii) the neural network minimizing the GMSE loss function should perform better than the FGLS estimator, and (iv) the performance difference between the standard and transformed neural networks should be greater than the difference between the linear models since neural networks are more likely to overfit on the latent noise process.

6.1 Description of data

Fundamental data at the daily frequency are obtained from Sharadar Core US Fundamentals Data via Quandl. Intraday price data at the 1-minute frequency are downloaded via IQFeed from DTN[15]. To avoid survivorship bias, the 100 largest US-based stocks by market capitalization are selected at the start of the time series. If stocks are delisted, they are kept in the dataset. One exception is the unavailability of intraday data for equities where the ticker was recycled. For example, General Motors Company, with ticker GM, filed for bankruptcy in 2009. The ticker was later reused for the new equity and the data of the old entity is not available anymore. If intraday data is unavailable for a particular stock, the corresponding diagonal element of the covariance matrix is filled with the average value of all available variances for the period and the off-diagonal elements are set to 0. This way there is no survivorship bias.

The daily log-returns are computed as the open-to-close log-price difference. This definition differs from many other empirical studies to reflect the fact that the integrated covariance estimate is for the open-to-close period and not close-to-close, as the

[15]I would like to thank John Fullerton and Brian Grosso of JBF CAPITAL, INC. for funding the data feed subscriptions.

estimate is based on all available intraday returns but not the overnight return.

The top 100 largest stocks by market capitalization based on prices as of January 4, 2010, and the most recent number of shares outstanding publicly available on that date are selected. This selection defines the universe. Features are the earnings-before-interest-and-taxes-to-enterprise-value ratio ($EBIT/EV$), the return on invested capital ($ROIC$), the previous 250-day cumulative return (MOM), the previous 20-day cumulative return MR, the Trailing 12 Months (TTM) log-change of revenue (GR), the TTM net profit margin (NM), the TTM ratio of dividends paid out relative to net income (PR), the TTM debt-to-equity ratio (D/E), and the TTM current-assets-to-current-liabilities ratio (CR). Fundamental data are point-in-time, excluding restatements, and time-stamped to the date one day after the report was submitted to the Security and Exchange Commission (SEC). If any of these variables are not available due to insufficient publicly available records as of January 4, 2010, the observation is dropped from the dataset. If the covariance matrix estimate is still not positive definite after missing data has been filled in as described above and condition number targeting with a target of 1000 was performed, the day is dropped from the dataset.

The data set is split into training, validation, and test set containing 1000, 750, and 746 days of data, respectively. The features of the whole dataset are normalized with their mean and standard deviation of the training set. The test set is then separated to avoid any information leakage.

6.2 Description of models

A feedforward neural network with 5 hidden layers each with 100 neurons, ReLU activation function, and batch normalization is fitted on the training set via the Adam optimizer with $\alpha^{(Adam)} = 10^{-5}$, $\beta_1^{(Adam)} = 0.9$, $\beta_2^{(Adam)} = 0.999$, and $\epsilon = 10^{-8}$ for 200 epochs. Each batch contains one day's worth of features for all stocks and is shuffled after each epoch. All combinations of the following regularization hyperparameters are tried: $dropout = [0, 0.2, 0.5, 0.7]$, $weight\ decay = [0, 0.01, 0.1]$. The MSE-net denotes the model described above minimizing the MSE loss function and the FGMSE-net denotes the model described above minimizing the feasible GMSE (FGMSE) loss. The linear models considered are the OLS estimator and the FGLS estimator. The covariance matrix for the feasible generalized models is estimated via the nonparametrically eigenvalue-regularized ensembled pairwise integrated covariance (NER_EPIC) estimator using 4 sample splits applied to all available centered minutely log-prices.

6.3 Model selection and out-of-sample results

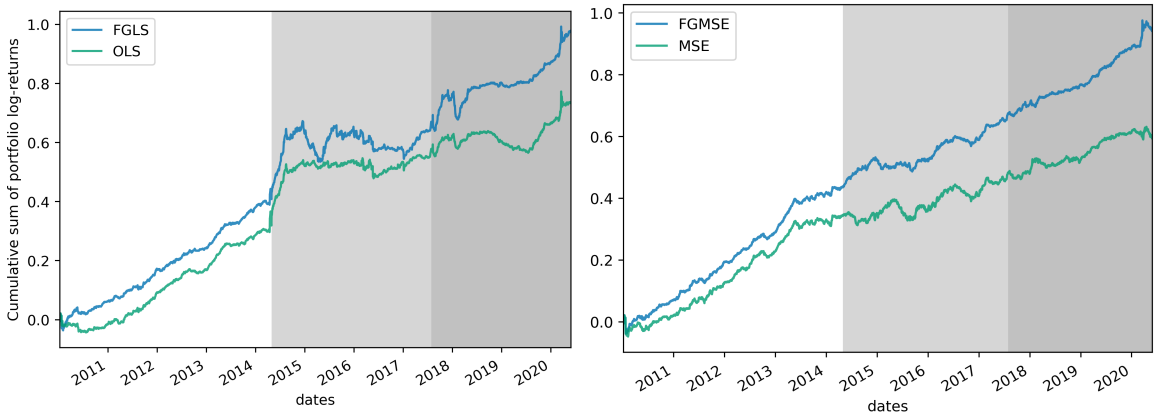
The FGMSE-net model with the largest IR of the portfolio returns on the validation set is selected. This model (and only this single model, to avoid selection bias) is then tested on the test set. Similarly as in Equation (96), the portfolio weight vector is computed for each day via

$$\boldsymbol{w}_t = \frac{\widehat{\boldsymbol{\Sigma}_t^{-1}}\widehat{\boldsymbol{\Delta}\mathbf{y}_t}}{\mathbf{1}_p' \operatorname{abs}(\widehat{\boldsymbol{\Sigma}_t^{-1}}\widehat{\boldsymbol{\Delta}\mathbf{y}_t})}. \tag{97}$$

Note that it is defined with an (at day t) unobservable covariance matrix estimate to crystallize the performance of the conditional expectation model rather than the covariance forecasting model. In contrast to the simulation, the true covariance matrix is, of course, unknown and is estimated via the NER_EPIC estimator. Hence, the IR is not a measure of realistic trading performance as the integrated covariance matrix estimate for the day will not be observable at the beginning of the day and transaction costs are ignored. To reiterate, the reasoning behind this metric is the focus on efficiency enhancement of the conditional expectation model (i.e., the difference of the IRs) and not on the covariance matrix forecast (which is the same for all considered models).

The FGMSE-net model with *dropout* = 0.2 and *weight decay* = 0.1 performs the best on the validation set. The test set predictions are obtained by feeding the separated and untreated features through the network. Hence, there is no risk of contamination with future information. The IR achieved on the test set is 2.58809. The mean of portfolio returns is statistically significantly greater than zero (p-value = 2.1919×10^{-5}). The FGLS estimator also generates a statistically significantly positive out-of-sample portfolio mean return (p-value = 0.001444). Its test-set IR is, however, lower at 1.88121. The OLS estimator, fitted on the training set, generates an IR of only 1.1890. Similarly, the MSE-net, fitted with the same hyperparameters as the FGMSE-net, achieves an IR of only 1.084839 on the test set. These results indicate that the proposed linear transformation can increase parameter estimation accuracy, leading to improved trading strategy performance in real-world data. The substantially smoother cumulative-profit-curve of the FGMSE-net relative to the linear FGLS estimator indicates important nonlinear effects. Figure 6 plots the cumulative portfolio log-returns based on the four models' predictions for the training set, validation set and test set. These results confirm the predictions made at the beginning of this section.

Figure 6: The cumulative sum of portfolio open-to-close log-returns, based on the conditional mean predictions and the daily integrated covariance matrix estimate, is plotted for the training set (white background), validation set (light gray background) and test set (dark gray background).



(a) Linear regression

(b) Neural network

7 Conclusion

Returns of financial assets exhibit significant cross-sectional correlation and volatility clusters. Conditional expectation models typically capture only small amounts of the variance, which causes these properties to be propagated into the residual. This attribute renders estimators that neglect it, such as OLS or neural networks with standard loss functions, inefficient relative to their generalized counterparts that account for it. Increasing amounts of intraday data and computational resources, as well as recent advances in covariance matrix regularization, make it possible to compute accurate estimates of the conditional covariance matrix. These estimates can be used to increase the efficiency of predictive models. Simulations of reasonable specifications show that the standard error of a univariate OLS model parameter estimate can be reduced by more than 50%. Monte Carlo evidence also shows that the proposed generalized MSE loss function helps deep neural networks to discover weakly predictive nonlinear relationships, which can then be used to create trading strategies with economically and statistically significant information ratios, whereas the neural network with the standard MSE objective function was unable to do so. For higher signal-to-noise ratio simulations the treatment increases parameter estimation accuracy resulting in greater information ratios with high statistical significance. In an empirical study of the 100 largest US-based publicly traded corporations, a deep neural network, optimizing the proposed generalized loss function, was able to discover predictive signals evidenced by its corresponding trading strategy's out-of-sample information ratio of 2.58809.

A Code

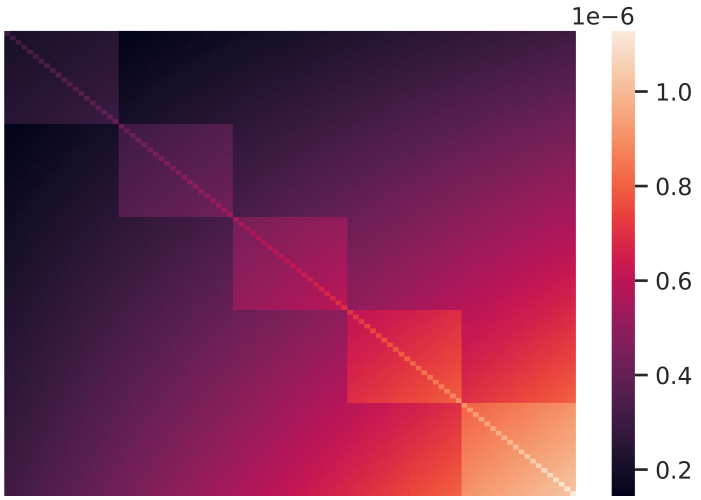
A.1 The Generalized MSE loss function

```
import torch

def gmse(y, y_hat, delta):
    """
    The Generalized MSE loss function. Increasing efficiency of
    panel regressions by accounting for dynamic cross-sectional
    error correlation of p assets for n time periods.

    Parameters
    ----------
    y : tensor, shape=(n, p)
        The targets.
    y_hat : tensor, shape=(n, p)
        The predictions.
    delta : tensor, shape=(n, p, p)
        The square root matrix of the inverse covariance matrix of p
        assets for n time periods.

    Returns
    -------
    out : tensor, shape=(1, 1)
        The GMSE loss.
    """
    error = y - y_hat
    error_transformed = torch.einsum('np, npj -> nj', error, delta)

    return torch.mean(error_transformed**2)
```

B Figures

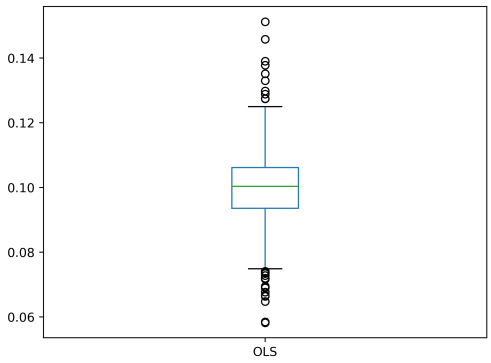
B.1 The Universe simulation

Figure 7: A heatmap of the unconditional integrated covariance matrix for 100 stocks, using a universe specification with 1 market factor, 5 industries, and an idiosyncratic component for each stock. The factor returns, industry returns and idiosyncratic returns are drawn from a GARCH(1, 1) model with specification $[\sigma_0^2, \mu, \alpha, \beta, \omega] = [0.1^2/250/6.5/60, 0, 0.0199, 0.98, \sigma_0^2/(1-\alpha-\beta)]$ each. Each stock belongs to exactly one industry. The exposure to the market factor is determined by $\beta_{M,j,t}$, which ranges from 1 to 3 with constant increments.

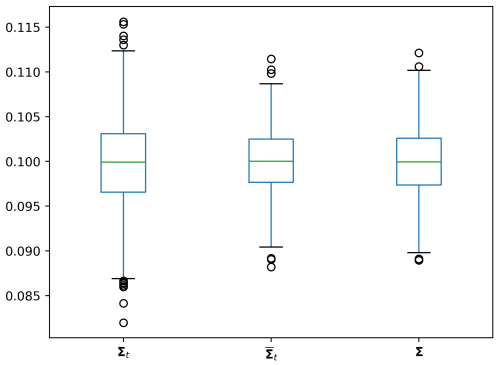


B.2 The simulation parameter estimates

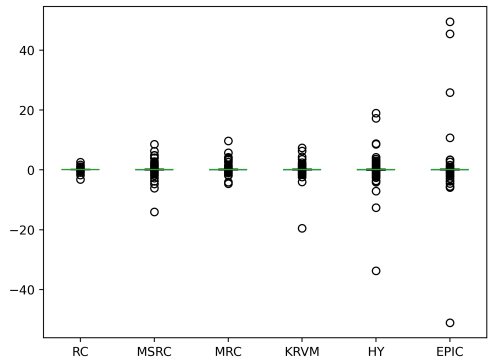
Figure 8: Estimates $\widehat{\beta}$. 100 stocks non-synchronously observed with microstructure noise. $\beta = 0.1$.



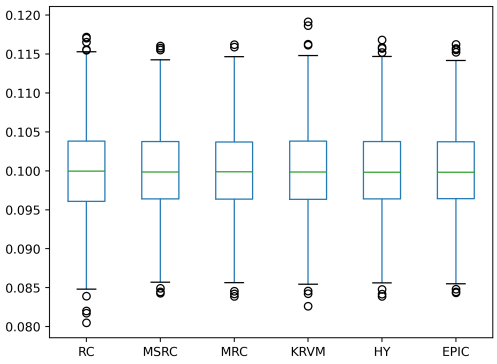
(a) OLS.



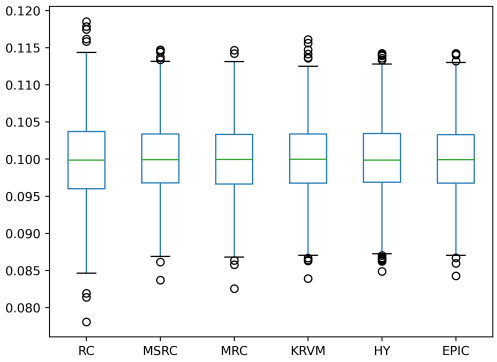
(b) GLS.



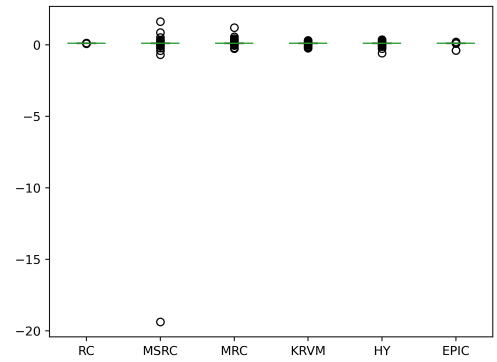
(c) FGLS with raw covariance estimates.



(d) FGLS with linearly shrunk covariance estimates.

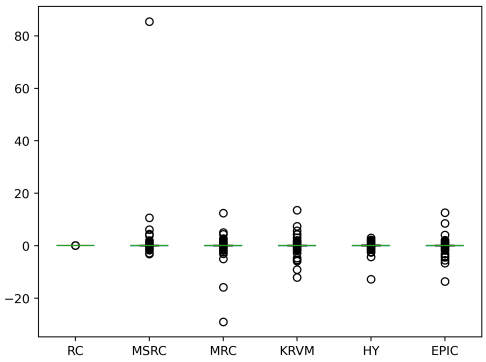


(e) FGLS with nonlinearly shrunk covariance estimates.

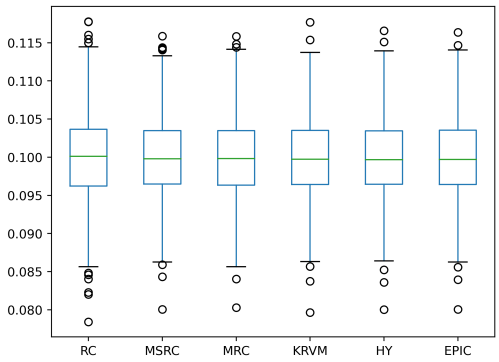


(f) FGLS with nonparametric eigenvalue-regularized covariance estimates.

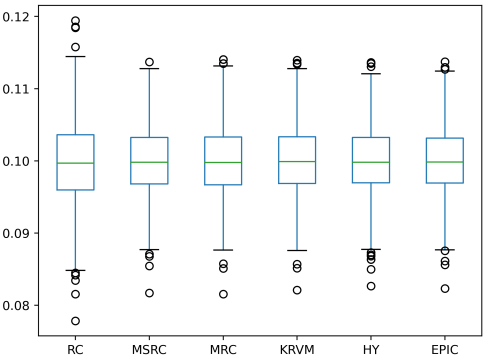
Figure 9: Estimates $\widehat{\beta}$. 100 stocks non-synchronously observed with microstructure noise. $\beta = 0.1$. 2 day average.



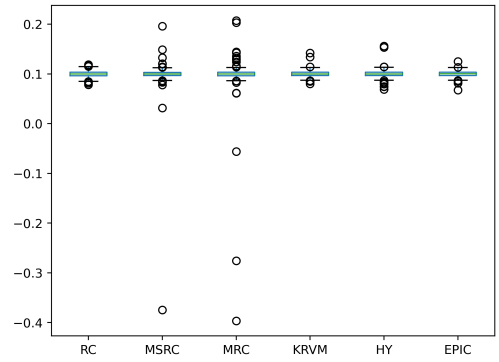
(a) FGLS with raw covariance estimates. 2 day average.



(b) FGLS with linearly shrunk covariance estimates. 2 day average.

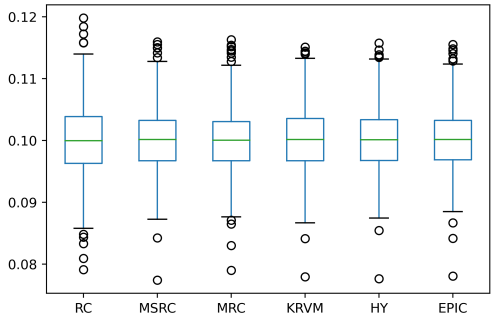


(c) FGLS with nonlinearly shrunk covariance estimates. 2 day average.

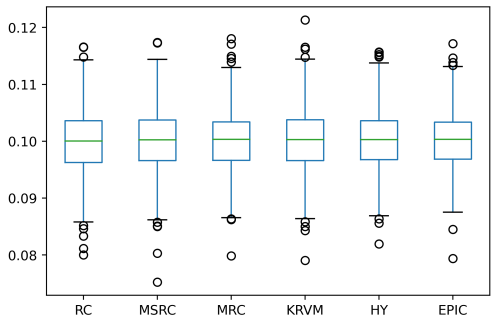


(d) FGLS with nonparametric eigenvalue-regularized covariance estimates. 2 day average.

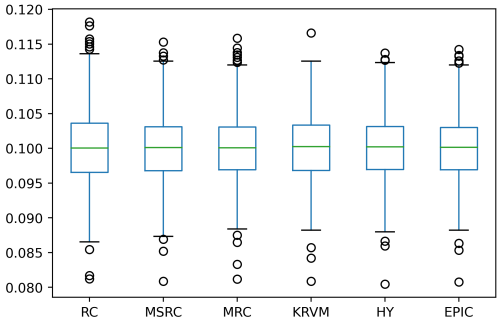
Figure 10: Estimates $\widehat{\beta}$. 100 stocks non-synchronously observed with microstructure noise. $\beta = 0.1$. With condition number targeting to reduce outliers.



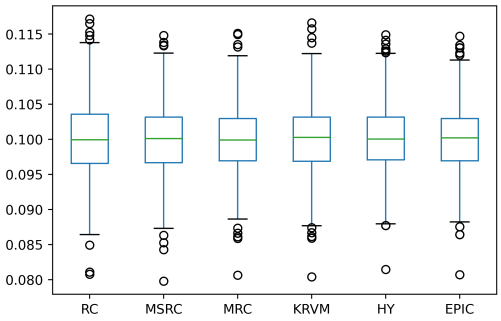
(a) FGLS with nonlinearly shrunk covariance estimates with condition number targeting.



(b) FGLS with nonparametric eigenvalue-regularized covariance estimates with condition number targeting.



(c) FGLS with nonlinearly shrunk covariance estimates with condition number targeting. 2 day average.



(d) FGLS with nonparametric eigenvalue-regularized covariance estimates with condition number targeting. 2 day average.

References

Abadir, K. M., Distaso, W. and Žikeš, F. (2014). Design-free estimation of variance matrices, *Journal of Econometrics* **181**(2): 165–180.

Ait-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data, *Journal of Econometrics* **201**(2): 384–399.

Aitken, A. C. (1936). Iv.—on least squares and linear combination of observations, *Proceedings of the Royal Society of Edinburgh* **55**: 42–48.

Avramov, D. and Chordia, T. (2006). Asset pricing models and financial market anomalies, *The Review of Financial Studies* **19**(3): 1001–1040.

Bai, J. (2003). Inferential theory for factor models of large dimensions, *Econometrica* **71**(1): 135–171.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading, *Journal of Econometrics* **162**(2): 149–169.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices, *The Annals of Statistics* **36**(1): 199–227.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of econometrics* **31**(3): 307–327.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. econometrica 51 1281–1304, *Mathematical Reviews (MathSciNet): MR736050 Digital Object Identifier: doi* **10**: 1912275.

Christensen, K., Kinnebrock, S. and Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data, *Journal of Econometrics* **159**(1): 116–133.

Christensen, K., Podolskij, M. and Vetter, M. (2013). On covariation estimation for multivariate continuous itô semimartingales with noise in non-synchronous observation schemes, *Journal of Multivariate Analysis* **120**: 59–84.

De Nard, G., Ledoit, O. and Wolf, M. (2018). Factor models for portfolio selection in large dimensions: The good, the better and the ugly, *University of Zurich, Department of Economics, Working Paper* (290).

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *Journal of Business & Economic Statistics* **20**(3): 339–350.

Engle, R. F., Ledoit, O. and Wolf, M. (2019). Large dynamic covariance matrices, *Journal of Business & Economic Statistics* **37**(2): 363–375.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density, *Theory of Probability & Its Applications* **14**(1): 153–158.

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

Fan, J., Li, Y. and Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection, *Journal of the American Statistical Association* **107**(497): 412–428.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4): 603–680.

Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data, *Journal of the American Statistical Association* **102**(480): 1349–1362.

Gloter, A. and Jacod, J. (2001). Diffusions with measurement errors. i. local asymptotic normality, *ESAIM: Probability and Statistics* **5**: 225–242.

Gourieroux, C. and Jasiak, J. (2001). Dynamic factor models, *Econometric Reviews* **20**(4): 385–424.

Hautsch, N., Kyj, L. M. and Oomen, R. C. (2012). A blocking and regularization approach to high-dimensional realized covariance estimation, *Journal of Applied Econometrics* **27**(4): 625–645.

Hautsch, N. and Podolskij, M. (2013). Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: theory, implementation, and empirical evidence, *Journal of Business & Economic Statistics* **31**(2): 165–183.

Hayashi, T. and Yoshida, N. (2004). Asymptotic normality of nonsynchronous covariance estimators for diffusion processes, *preprint* .

Hayashi, T. and Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes, *Bernoulli* **11**(2): 359–379.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *CoRR* **abs/1502.01852**. **URL:** *http://arxiv.org/abs/1502.01852*

Ibragimov, I. (1963). A central limit theorem for a class of dependent random variables, *Theory of Probability & Its Applications* **8**(1): 83–89.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* .

Krogh, A. and Hertz, J. A. (1992). A simple weight decay can improve generalization, *Advances in neural information processing systems*, pp. 950–957.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator, *The Annals of Statistics* **44**(3): 928–953.

Lam, C. and Feng, P. (2018). A nonparametric eigenvalue-regularized integrated covariance matrix estimator for asset return data, *Journal of Econometrics* **206**(1): 226–257.

Lam, C. and Qian, C. (unpublished paper). Integrated volatility matrix estimation with nonparametric eigenvalue regularization.

Laurent, S. and Shi, S. (2020). Volatility estimation and jump detection for drift–diffusion processes, *Journal of Econometrics* .

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* **88**(2): 365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices, *The Annals of Statistics* **40**(2): 1024–1060.

Ledoit, O. and Wolf, M. (2018). Analytical nonlinear shrinkage of large-dimensional covariance matrices, *University of Zurich, Department of Economics, Working Paper* (264).

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices, *Matematicheskii Sbornik* **114**(4): 507–536.

Markowitz, H. (1952). Portfolio analysis, *Journal of Finance* **8**: 77–91.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines, *ICML*.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Technical report*, National Bureau of Economic Research.

Nolte, I. and Voev, V. (2007). Estimating high-frequency based (co-) variances: A unified approach, *Available at SSRN 1003201* .

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A. (2017). Automatic differentiation in pytorch.

Podolskij, M., Vetter, M. et al. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps, *Bernoulli* **15**(3): 634–658.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms, *Technical report*, Cornell Aeronautical Lab Inc Buffalo NY.

Sharpe, W. F. (1966). Mutual fund performance, *The Journal of business* **39**(1): 119–138.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices, *Journal of Multivariate Analysis* **55**(2): 331–339.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15**(1): 1929–1958.

Stein, C. (1975). Estimation of a covariance matrix, *39th Annual Meeting IMS, Atlanta, GA, 1975*.

Stein, C. (1986). Lectures on the theory of estimation of many parameters, *Journal of Soviet Mathematics* **34**(1): 1373–1403.

Tao, M., Wang, Y. and Chen, X. (2013). Fast convergence rates in estimating large volatility matrices using high-frequency financial data, *Econometric Theory* **29**(4): 838–856.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75, *Time series analysis: theory and practice* **1**: 203–226.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American statistical Association* **57**(298): 348–368.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach, *Bernoulli* **12**(6): 1019–1043.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise, *Journal of Econometrics* **160**.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data, *Journal of the American Statistical Association* **100**(472): 1394–1411.

Affirmation

I hereby declare that I have composed my Master's thesis "Efficient estimation of predictive models using high-frequency high-dimensional data" independently using only those resources mentioned, and that I have as such identified all passages which I have taken from publications verbatim or in substance. I agree that the work will be reviewed using plagiarism testing software. Neither this thesis, nor any extract of it, has been previously submitted to an examining authority, in this or a similar form.

I have ensured that the written version of this thesis is identical to the version saved on the enclosed storage medium.

Kiel, August 9, 2020

_____

(Jan Wöltjen)