CERGE - EI

Center for Economic Research and Graduate Education –

Economics Institute

Essays in High-Dimensional Econometrics and Finance

Vladimir Pyrlik

Dissertation

Prague 2024

Dissertation Committee:

Stanislav Anatolyev, Ph.D. (CERGE-EI, Chair)

doc. PhDr. Jozef Baruník, Ph.D. (Institute of Economic Studies, Charles University)

Veronika Selezneva, Ph.D. (Université Paris Dauphine - PSL)

Referees:

1. Professor Artem Prokhorov, Ph.D. (The University of Sydney)

2. Vladimir Volkov, Ph.D. (University of Tasmania)

"My darling girl, when are you going to realize that being normal is not necessarily a virtue? It rather denotes a lack of courage."

— Aunt Francis, in Alice Hoffman's *Practical Magic*

"We're all mad here."

— The Cheshire Cat, in Lewis Carroll's *Alice's Adventures in Wonderland*

To Natalia G., my therapist, whose steadfast support and guidance have been a source of strength and encouragement throughout my journey, shaping both my personal growth and academic pursuits. Thank you for your unwavering commitment and compassion.

Table of Contents

Prague, Czech Republic Vladimir Pyrlik

March, 2024

Abstract

High dimensionality is a popular contemporary setting in applied statistical analysis of various types of data. Growing dimensionality of the data challenges precise and well-conditioned estimates of statistical models. We address selected high-dimensional methods and their applications to financial market analysis. First, we consider modeling flexible joint distributions using copulas. In Chapter 1, we address estimation of Gaussian and t copulas in ultra-high dimensions, up to thousands of variables that use up to 30 times shorter sample lengths. We employ large covariance matrix shrinkage tools to obtain precise and well-conditioned estimates of the matrix parameters of the copulas. In Chapter 2, we present a new method for estimating the skew-t copula, known for its advantageous properties in characterizing joint distributions, including asymmetry, heavy tails, and asymmetric tail dependence. Our approach involves a two-step procedure based on the simulated method of moments and analytical non-linear shrinkage estimator for large covariance matrices. In both chapters, we also illustrate the benefits of the copula approach in a large stock portfolio allocation. Our analysis shows that copula-based models deliver better portfolios in terms of cumulative returns and maximum downfalls over the portfolio lifetime than the benchmark alternatives. In Chapter 3, we step away from unconditional distribution modeling, and assess the performance of selected machine learning algorithms in forecasting daily realized volatility. We utilize high-dimensional settings and mixed-frequency data set-ups to improve forecasts of selected stocks returns volatility in the Russian stock market in 2018-2020.

Introduction

High-dimensional settings have gained popularity in contemporary statistical analysis, driven by applications characterized by an increasing number of variables or relatively smaller sample sizes. As the dimensionality of the data, denoted by the ratio of variables to the sample size, grows, obtaining precise and well-conditioned estimates of statistical models becomes more challenging. This work addresses selected methods for high-dimensional statistical analysis and their applications to financial market analysis and forecasting.

In the first part, we focus on modeling flexible joint distributions, where copulas serve as a convenient tool. Currently, copula-based high-dimensional settings typically involve a few hundred variables and demand large data samples for accurate estimation. In Chapter 1[1], we tackle the estimation of Gaussian and t copulas in ultra-high dimensions—up to thousands of variables—with sample lengths as much as 30 times shorter. We employ large covariance matrix shrinkage tools to obtain precise and well-conditioned estimates of the matrix parameters for both Gaussian and t copulas.

Moving to Chapter 2[2], we introduce a novel method for estimating the skew-t copula, recognized for its advantageous properties in characterizing joint distributions, including asymmetry, heavy tails, and asymmetric tail dependence. Our approach utilizes a two-step procedure based on the simulated method of moments and an analytical non-linear shrinkage estimator for large covariance matrices. In both chapters 1 and 2, we illustrate the benefits of the copula approach in large stock portfolio allocation. Our analysis reveals that copula-based models yield portfolios with superior cumulative returns and fewer downfalls over the portfolio lifetime compared to benchmark alternatives. The skew-t copula's ability to effectively account for tail dependence within the distributions of asset returns significantly contributes to portfolio performance.

In Chapter 3[3], we shift focus from unconditional distribution modeling to assessing the

[1] the results of the research presented in this chapter were co-authored and published in the Journal of Economic Dynamics and Control (Anatolyev and Pyrlik, 2022)

[2] this chapter presents results of the research that are solo-authored

[3] the research presented in this chapter was co-authored and published (Pyrlik et al., 2021)

performance of selected machine learning algorithms in forecasting the daily realized volatility of returns for top stocks in the Russian stock market. We compare these forecasts with the widely-used benchmark, the heterogeneous autoregressive realized volatility, over the period 2018-2020. Utilizing high-dimensional settings and a mixed-frequency data setup, we enhance the model's predictive power by including various economic indicators that carry information about future volatility. Our findings indicate that Lasso delivers a good combination of easy implementation and forecast precision. Other algorithms require fine-tuning and frequent re-training to outperform the benchmark consistently. Lagged log-RV values emerge as the only significant explanatory variables for the benchmark's in-sample quality.

1 Shrinkage for Gaussian and t Copulas in Ultra-High Dimensions

1.1 Introduction

In this chapter, we present the results of our research of how methods of high-dimensional covariance matrix estimation can be extended to copulas estimation. The results are also published in the Journal of Economic Dynamics and Control (Anatolyev and Pyrlik, 2022).

Copulas are an attractive tool to model joint distributions due to a high degree of flexibility and ability to capture various properties of the real data, both in marginal distributions and dependence structures (Patton, 2009). An important recent challenge in modeling joint distributions is the upward trend in data dimensionality. For example, financial market participants are challenged to deal with thousands of alternative assets to allocate their funds into (Ledoit and Wolf, 2017a; De Nard et al., 2018; Müller and Czado, 2019).

High dimensional datasets are challenging in many applications that involve statistical estimation, computation, and inference. Having hundreds and thousands of variables in the data complicates each step of statistical modeling, with estimation and inference the most problematic. In particular, when the dimensionality of datasets becomes comparable to available sample sizes, a variety of traditional estimators tends to fail to deliver desirable properties that researchers normally seek to obtain (Ledoit and Wolf, 2004a,b, 2022).

Regarding existing high-dimensional copula settings, a common limitation is the actual number of dimensions relative to sample sizes used that are called '*high dimensional*'. What most studies usually explore as high dimensional settings tend to appear rather moderately dimensional. In this paper, we focus the two most commonly used in modeling and practical applications elliptical copulas: Gaussian and t copulas. Their dimensionality of the parameter space is directly connected to the data dimensionality, with the matrix parameter naturally interpretable in the description of the degree of pairwise dependence among the variables. An important property of these copulas is that the matrix parameter is very close to the correlation matrix of pseudo-observations (Demarta and McNeil, 2005; Kojadinovic and Yan,

2010). Hence, in low dimensions, the copulas are effectively estimated via computationally very practical method-of-moments-like techniques based on rank correlations and sample correlation matrices (Demarta and McNeil, 2005). However, in high dimensions the settings and their estimates inherit the same problems as the traditional covariance matrix estimators. This makes it practical to use the shrinkage estimators of Ledoit and Wolf (2004b, 2017b); Ledoit et al. (2020) to estimate the matrix parameters of Gaussian and t copulas in high dimensional datasets.[4]

We consider datasets with up to thousands of variables that use up to 20 times lower sample sizes. Thus, we take the data dimensionality well beyond what is studied in the copula literature; hence the prefix "ultra-" in "high dimensions" in the title.[5] In a simulation study, we compare the quality of performance of different estimators for various ratios of data dimensionality to sample size. We show that the shrinkage estimators significantly outperform the traditional copula matrix parameter estimators based on sample analogs of Kendall's rank correlation and approximate Spearman's rank correlation. The performance of estimators is measured in terms of both the closeness of estimated parameter values to their actual values and the closeness of the entire estimated copula function to its true counterpart. Not only do we show that the shrinkage estimators outperform the traditional estimators of the copula matrix parameters, but also we find that non-linear shrinkage generally tends to dominate the linear one.

As an empirical application, we apply shrinkage-based estimators of copula correlation matrices in high dimensions to a large portfolio of stocks allocation problem and compare emerging portfolios to those from a multivariate normal model and copula models based on traditional estimators. Using daily data on prices of roughly 5000 U.S. stocks, we construct portfolios of up to 3600 assets and simulate buy-and-hold portfolio strategies. The joint

[4] In the case of t copula, one also needs to estimate the scalar degrees-of-freedom parameter that controls the thickness of copula tails. We confirm that once the large matrix parameter is sufficiently precisely estimated, the remaining scalar parameter can be effectively estimated via maximum pseudo-likelihood method.

[5] The maximum of 1000 for data dimensionality in the simulation study is determined by the computational capacities at our disposal. With a thousand variables and largest samples, simulations are computationally very demanding, particularly due to multiple iterations in computing quality criteria. The results suggest, however, that the shrinkage estimators can be effectively used in even higher dimensions; in our empirical example, the t copula is estimated for 3600 variables in the dataset.

distributional models of asset returns are estimated over the period of six months (120 observations), hence the problem is ultra-high dimensional, with the dimensionality ratio of 30. To our knowledge, this is the highest dimensionality of the large portfolio allocation problem considered in the literature. The comparison of the portfolios based on different models to equally weighted portfolios shows that the shrinkage-based estimators applied to t copula based models of return distribution deliver better portfolios in terms of both cumulative return and maximum downfall over the portfolio lifetime than the corresponding portfolios derived from the multivariate normal or copula-based models estimated via traditional estimators.

The rest of this chapter is organized as follows. Section 1.2 presents a more detailed review of related literature and explains the focus of our study. Section 1.3 covers the methodology including a description of chosen copulas and their main properties, existing approaches to copula estimation, drawbacks thereof and the solution we propose. In Section 1.4, we describe the simulation study design and results. An empirical application of the shrinkage estimators to a large portfolio allocation problem is presented in Section 1.5. Section 1.6 concludes. Appendices A, B and C contain some additional technical material, including tables with detailed results of the simulations in Supplementary Appendix.

1.2 Background

Modeling joint distributions has been a major task in a wide variety of applications. One way to deal with dependence in multivariate settings is to directly model the joint distribution of quantities of interest using a family of multivariate distributions. However, in most applications, there are only few such families that can capture the crucial properties of actual data. Although the multivariate normal is popular due to its analytical and computational convenience, it is also widely criticized for symmetry, non-heavy tails, and linearity of conditional means. Asymmetric and heavy-tailed multivariate distributions are much more cumbersome to work with, particularly in higher dimensions.

Copula-based settings are attractive due to a higher degree of flexibility and ability to capture various properties of the real data, both in marginal distributions and dependence structures

(Patton, 2009). In particular, the financial literature has been giving copulas increasing attention since the 2008 financial crisis. One of critical effects of the crisis was that the quantities previously viewed as "almost independent" were unexpectedly co-moving, resulting in a joint crush in several markets (Zimmer, 2012; Patton, 2012; De Leon and Chough, 2013). This effect of so-called tail-dependence appears crucial for modeling joint distributions in financial markets; yet it was absent in the traditional multivariate normal-based settings (Patton, 2013; Oh and Patton, 2017). Various alternative dependence structures have been proposed to account for the critical properties of real data. For example, the t copula of Demarta and McNeil (2005) was exploited in many studies, although it captures only symmetric tail dependence (Sukcharoen et al., 2014; Ning, 2010; Wen et al., 2012). It was then further extended by Kollo and Pettere (2010) and Smith et al. (2012) to account for asymmetric extreme co-movements, and the resulting versions of skewed-t copula have since been a popular choice to model inter- and intra-market dependencies (Kollo and Pettere, 2010; Smith et al., 2012; Patton, 2012, 2013).

Consistently growing data dimensionality constitutes another challenge in modeling joint distributions. While financial market participants can access thousands of options for their funds' allocation, they may be at the same time become limitied in the sample size bound to use the most relevant information and account for recent changes in the market (Ledoit and Wolf, 2017a; De Nard et al., 2018; Müller and Czado, 2019; Engle et al., 2019). In this and many other applications with the number of variables growing potentially well above the limited sample size, multiple statistical estimators that would perform well under low dimensionality fail to result in precise or well-conditioned estimates (Ledoit and Wolf, 2004a,b; Ledoit et al., 2020; Ledoit and Wolf, 2022).

Although there has been significant progress in multivariate methods addressing the high dimensionality challenge, most of the work has been done to restore the properties of estimators up to the second moment. In particular, a variety of estimators robust to growing dimensionality have been recently developed to improve the estimation of large covariance matrices (Ledoit and Wolf, 2017b; De Nard et al., 2018; Anatolyev et al., 2018; Ledoit and Wolf, 2022).

At the same time, significant progress has been observed in the copula theory and applications addressing high dimensional data (Patton, 2009; Müller and Czado, 2019; Smith, 2021). For example, Oh and Patton (2016) suggest a copula version of a high dimensional factor model. Later, Oh and Patton (2017) use mixed frequency data to construct high dimensional distributions. Müller and Czado (2017) develop another type of approach to use the advantages of copulas in high dimensional case that relies on sparse data structures, which allow one to combine copulas with lasso estimation. Another direction in the development of high dimensional copula-based models relies on the pair copula constructions (PCCs, aka vines). Based on hierarchical pair-wise copula construction, the vines presume very flexible settings and an intuitive interpretation of dependence structures that make them an attractive modeling tool (Brechmann and Czado, 2013).

A common limitation of the existing approaches to constructing high dimensional copulas is the actual number of dimensions relative to sample sizes used that are called '*high dimensional*'. What most studies usually explore as high dimensional settings tend to appear rather moderately dimensional. Until recently, the dimensionality of data in empirical applications of PCCs rarely had exceeded a few dozen variables (Brechmann and Czado, 2013), with only several studies applying the PCCs to settings with more than a hundred variables. Currently, the very recent study by Müller and Czado (2019) is the only one with PCCs applied in the framework with more than a thousand variables. Still, the study focuses on sparse structures that are identified heuristically from the data, and uses a considerable number of observations in the sample (viz., $n = 999$ observations and $p = 2131$ variables). Given that the data dimensionality exceeds the number of observations, this setting is indeed high-dimensional. However, in many applications the ratio of the data dimensionality to available sample sizes can be significantly higher, with sparse structures being an excessively strong assumption.

We focus on elliptical copulas in high dimensions, particularly, Gaussian and t copulas that are most commonly used in modeling and practical applications as either main modeling frameworks, important building blocks of more complicated and flexible settings, or benchmark models. Often, Gaussian and t copulas are used to model the joint distribution of characteristics of objects or events located or taking place in different points of geographi-

cal space. This is found particularly useful in environmental and civil engineering studies (Van de Vyver and Van den Bergh, 2018; Li et al., 2018; Valle and Kaplan, 2019) and energy economics (Atalay and Tercan, 2017; Schindler and Jung, 2018). Regression analysis and pattern recognition is another field where these copulas are applied (Fu and Wang, 2016; Kwak, 2017; Li et al., 2017, 2019), including the high-dimensional context, with the data dimensionality exceeding the number of observations (He et al., 2018, 2019). In finance, the Gaussian and t copulas are criticized for inability to capture asymmetric dependence. However, they have proved beneficial for modeling the joint distribution of assets returns as compared to the traditional models that disregard dependencies beyond correlations. Most often, they are applied to model joint distributions of financial assets or indices returns for the task of portfolio allocation (Karmakar, 2017; Han et al., 2017; Lourme and Maurer, 2017), but also in studies of tail dependence (Huang et al., 2009; Zorgati et al., 2019) and asset pricing (Hörmann and Sak, 2010).

Other elliptical copulas, as well as their skewed versions or even selected cases from the more general implicit copulas, may offer very appealing degrees of flexibility and wide range of properties (Smith, 2021). Most of them are easy to extend to high-dimensions, and often the parameters are well interpretable. However, unlike for the Gaussian and t copulas, more complex copula structures are parameterized in the ways that make the parameters not as easy to relate to the observed data and disentangle in estimation (Kollo and Pettere, 2010; Daul et al., 2003; Smith, 2021). It limits the potential choice of estimation techniques and makes many of them computationally demanding beyond low dimensions (Yoshiba, 2018).

In the case of Gaussian and t copulas, the dimensionality of the parameter space is directly connected to the data dimensionality, with the matrix parameter naturally interpretable in the description of the degree of pairwise dependence among the variables. Moreover, the structure of these copulas is such that the matrix parameter is relatively easy to relate to the properties of the observed data. In low dimensions, Gaussian and t copulas are effectively estimated via computationally very practical method-of-moments-like techniques based on rank correlations and sample correlation matrices. However, in high dimensions the settings and their estimates inherit the same problems as the traditional covariance matrix estimators.

Thus, most settings based on the Gaussian and t copulas are low-dimensional, where the number of dimensions varies from two to a few dozen, and the ratio to corresponding sample sizes is considerably less than unity. However, some settings are high-dimensional with the ratio reaching five (He et al., 2018, 2019). More importantly, many applications that are currently low-dimensional can potentially benefit from increased dimensionality. This is particularly relevant for financial applications with more variables in datasets (e.g., more assets in multivariate models used for portfolio management). For applications in which the number of objects is rather low (e.g., in some spatial applications), the high-dimensional case is still relevant due to the necessity of estimating the dependence using small samples.

Recently, a substantial amount of research has focused on developing covariance matrix estimators that are robust to and well-conditioned under the data dimensionality growing along with the sample size. Two main directions towards solving the problem can be distinguished (Fan et al., 2008; Ledoit and Wolf, 2004b). The first approach is based on manipulating the data and relies on dimensionality reduction techniques to impose some structure on the covariances (Wong et al., 2003; Huang et al., 2006; Fan et al., 2008). Alternatively, researchers adjust the traditional sample covariance matrix by directly restricting its structure, eigenvalues or the inverse to achieve better properties under moderate or high data dimensionality (Daniels and Kass, 2001; Ledoit and Wolf, 2004b). Ledoit and Wolf (2012), Ledoit and Wolf (2017b) and Ledoit et al. (2020) developed newer versions of the previously developed estimator by Ledoit and Wolf (2004b). The new estimator relies on the random matrix theory and leads to fast and relatively easy estimation of large covariance matrices of dimensionality higher than had been feasible ever before. It has also proved substantially more efficient than a number of previously developed estimators of the same type (Ledoit and Wolf, 2017b).

These advances in large covariance matrix estimation rather conveniently match with the structure of Gaussian and t copulas. Because their matrix parameter is very close to the correlation matrix of pseudo-observations (Demarta and McNeil, 2005; Kojadinovic and Yan, 2010), shrinkage estimators of Ledoit and Wolf (2004b, 2017b); Ledoit et al. (2020) can be employed to estimate the Gaussian and t copulas, even in high dimensional datasets.

In the next section, we formally introduce the Gaussian and t copulas and their properties

that are crucial for the study. We then give briefly overview of the traditional estimation technique, followed by a rather detailed description of the shrinkage methods for large covariance matrices and suggested algorithm to apply them for the copulas' correlation matrix parameter estimation.

1.3 Methodology

1.3.1 Gaussian and t copulas

The Gaussian copula in p dimensions associated with correlation matrix $P \in \mathbb{R}^{p\times p}$ is defined as

$$C_P^{\mathcal{N}}(u) = F_P\Big(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_p)\Big), \tag{1.1}$$

where $F_P(x)$ is the joint CDF of the p-dimensional random vector drawn from multivariate normal distribution $\mathcal{N}(\mathbb{O}_p, P)$, and $\Phi^{-1}(u)$ is the quantile function of the univariate standard normal distribution. Similarly, the t copula with correlation matrix P and degrees of freedom parameter $\nu > 2$ is defined as

$$C_{P,\nu}^{t}(u) = t_{P,\nu}\Big(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_p)\Big), \tag{1.2}$$

where $t_{P,\nu}(x)$ is the joint CDF of the p-dimensional multivariate Student's t-distribution with ν degrees of freedom and the matrix parameter P, and $t_\nu^{-1}(u_i)$ is the quantile function of the standard univariate t-distribution with ν degrees of freedom.

As any other copula function, the copulas (1.1) and (1.2) are legit CDFs living on the domain $[0,1]^p$, and can be used accordingly. The first important property of these copulas is the relation between Kendall's rank correlation and the regular correlation coefficient[6]. For a pair of random variables $\{U_i, U_j\}$, Kendall's rank correlation, or Kendall's i-τ, is defined as

$$\tau_{ij} \equiv \mathbb{E}\left[\text{sign}\left((U_i - \tilde{U}_i)(U_j - \tilde{U}_j)\right)\right], \tag{1.3}$$

[6]by *regular* correlation we call the correlation coefficient of the underlying multivariate distribution, from which the copula is constructed, i.e. either multivariate normal or multivariate Student's t distribution in our case, that is exactly the coefficients of the matrix parameter P.

where $\{\tilde{U}_i, \tilde{U}_j\}$ is an independent from $\{U_i, U_j\}$ pair of similarly distributed random variables. Then, for $\mathcal{U} = (U_1, ..., U_p)' \sim C(u)$ for either $C(u) = C_P^{\mathcal{N}}(u)$ or $C(u) = C_{P,\nu}^{t}(u)$ it holds that:

$$\tau_{ij} = \frac{2}{\pi} \arcsin\left(P_{ij}\right). \tag{1.4}$$

Another important property is the relation between the matrix parameter P and the correlation of the random variables $\mathcal{U}$ distributed according to the copula function as their CDF[7]. Firstly, in the case of multivariate normal distribution and its copula, i.e. $\mathcal{U} \sim C_P^{\mathcal{N}}(u)$, the relation has the following analytical form:

$$\text{Corr}(\mathcal{U}) = \frac{6}{\pi} \left\{ \text{asin}\left(\frac{P_{ij}}{2}\right) \right\}_{i,j=1,...p}. \tag{1.5}$$

In practical estimation, however, especially beyond the bivariate case, the following approximation of this relation is used (Karmakar, 2017):

$$\text{Corr}(\mathcal{U}) \approx P. \tag{1.6}$$

In the case of t copula, $\mathcal{U} \sim C_{P,\nu}^{t}(u)$, there is no closed form expression for $\text{Corr}(\mathcal{U})$. Nevertheless, the relations (1.5) and (1.6) can be used as reliable approximations, with the corresponding approximation errors diminishing fast as ν grows (Demarta and McNeil, 2005; Karmakar, 2017). In the case of Gaussian copula, the absolute error of this approximation reaches at most 0.018. In the case of t copula, the error is higher, but it approaches the level of that for the Gaussian copula rather fast as the value of degrees of freedom grows. For example, for the t copula with 10 degrees of freedom, the error does not exceed 0.024. See more details in Appendix A.

Thus, (1.6) and (1.4) can be used to estimate the copula matrix parameter. We address the corresponding estimation techniques as traditional/benchmark estimators to compare with the proposed approach. The estimators are presented later in Sections 1.3.2, 1.3.3.

Another construct related to the copula function is the copula density function, the probability

[7] similarly to the *regular* correlation coefficient, in terms of the underlying distributions, from which the copulas are constructed, the correlation of the transformed r.v. $\mathcal{U}$ is called *the Spearman's rank correlation*

density function (PDF) associated with the copula function $C(u)$ as a CDF:

$$c(u) = \frac{\partial^p C(u)}{\partial u_1 ... \partial u_p}. \tag{1.7}$$

In the case of Gaussian and t copulas defined by (1.1) and (1.2) it is easy to show using (1.7) that the corresponding copula log-densities are

$$\log c_P^{\mathcal{N}}(u) = -\frac{1}{2}\log|P| - \frac{1}{2}\phi'(u)\cdot(P^{-1} - I_p)\cdot\phi(u), \tag{1.8}$$

and

$$\begin{aligned}\log c_{P,\nu}^{t}(u) = {} & \log\Gamma\left(\frac{\nu+p}{2}\right) + (p-1)\log\Gamma\left(\frac{\nu}{2}\right) - p\log\Gamma\left(\frac{\nu+1}{2}\right) - \frac{1}{2}\log|P| \\ & - \frac{\nu+p}{2}\log\left(1 + \frac{\psi_\nu'(u)P^{-1}\psi_\nu(u)}{\nu}\right) + \sum_{i=1}^{p}\log\left(1 + \frac{t_\nu^{-1}(u_i)^2}{\nu}\right),\end{aligned} \tag{1.9}$$

where $\phi(u) = \left(\Phi_{0,1}^{-1}(u_1), ..., \Phi_{0,1}^{-1}(u_p)\right)'$ and $\psi_\nu(u) = \left(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_p)\right)'$. The log-densities (1.8) and (1.9) are used in evaluation of estimation precision via the Kullback–Leibler information criterion (KLIC) presented later in Section 1.4.1.

Finally, the Gaussian and t copulas share the following important property. Consider $C_P(u)$, the Gaussian (or t) copula function (the degrees-of-freedom parameter is unimportant if it is a t copula) of a p-dimensional distribution of random vector $X = (X_1, ..., X_p)'$. Then, for any $\tilde{p}$-dimensional sub-vector $\tilde{X} = (X_{i_1}, ..., X_{i_{\tilde{p}}})$ (with $\tilde{p} < p$, $\{i_s\}_{s=1,...,\tilde{p}} \subset \{1, ..., p\}$, and $\forall s_1 \neq s_2 \in \{1, ..., \tilde{p}\},\ i_{s_1} \neq i_{s_2}$), the copula of the joint distribution of $\tilde{X}$ is also Gaussian (or t) with the matrix parameter $\tilde{P} = \{P_{i_{s_1} i_{s_2}}\}_{s_1, s_2 \in \{1,...,\tilde{p}\}}$.

1.3.2 Traditional estimators

Copulas allow one to separate estimation of the marginal distributions from estimation of the dependence structure embedded in the copula function. Even though for any copula the full maximum likelihood estimation (full MLE, FMLE) problem can be specified, the actual estimation is very demanding, especially in high dimensions. Hence, most of the estimators

of such models are performed in stages.

First, the marginal distributions $\{F_i\}_{i=1,...,p}$ are estimated from the corresponding univariate data on each of the variables $X_i = \{X_{it}\}_{t=1,...,n} = (X_{i1}, ..., X_{in})'$, where n is sample size. The curse of dimensionality does not apply at this stage, and we follow the convention in the copula literature and do not focus on estimating the marginals, assuming one can estimate them efficiently. Second, the estimates of the marginal distributions, $\{\hat{F}_i\}_{i=1,...,p}$, are used to transform the initial data $\{X_i\}_{i=1,...,p}$ into a corresponding set of so-called *pseudo-observations*

$$U_{it} = \hat{F}_i(X_{it}), \tag{1.10}$$

and the copula function is treated as the joint distribution function of the pseudo-observations (2.14), from which the parameters of the copula alone are estimated.

One way to proceed with estimation of copula parameters would be, again, the method of maximum likelihood. The estimation routine in this case is called maximum pseudo-likelihood estimation (MPLE). The method is based on maximization of the traditional conditional likelihood function, so it disregards the fact that the pseudo-observations (2.14) are never i.i.d. (because they are constructed from the estimates of marginal distributions $\hat{F}_i$, each constructed from the whole univariate sample X_i). Still, there is evidence that together with efficient univariate estimation of the marginals, the two-stage procedure as a whole delivers estimates that are very close to and barely worse than the full maximum likelihood (Demarta and McNeil, 2005).

The MPLE is universal among the copula classes, and with its resulting estimates being close to the FMLE, it is often a preferred method of copula estimation. On the other hand, the optimization problem is quite demanding in high dimensions for elliptical and other copulas with high-dimensional parameters. There is another approach to estimating parameters of the dependence structure relevant for the elliptical copulas. It is based on method-of-moments type of estimates for large matrix parameters, and allows one to separate estimation of the large matrix parameters from the rest of the copula function.

In the case of Gaussian and t copulas, the properties (1.6) and (1.4) are used to estimate the

matrix parameter P. Given sample data $\{X_i\}_{i=1,\dots,p}$ and corresponding pseudo-observations $\{U_i\}_{i=1,\dots,p}$, the matrix parameter P of either Gaussian or t copula can be estimated as

$$\hat{P}^{\text{smpl}} = \{\hat{P}_{ij}^{\text{smpl}}\}_{i,j=1,\dots,p} = \left\{\widehat{\text{corr}}(U_i, U_j)\right\}_{i,j=1,\dots,p}, \tag{1.11}$$

and

$$\hat{P}^{\text{i-}\tau} = \{\hat{P}_{ij}^{\text{i-}\tau}\}_{i,j=1,\dots,p} = \left\{\sin\left(\frac{\pi}{2}\hat{\tau}_{ij}\right)\right\}_{i,j=1,\dots,p}, \tag{1.12}$$

where $\widehat{\text{corr}}(U_i, U_j)$ and $\hat{\tau}_{ij}$ are the sample analogs of the correlation coefficients and Kendall's rank correlations for the pseudo-observations $\{U_i\}_{i=1,\dots,p}$.

The most important drawback of Kendall's i-τ estimator (1.12) is that the resulting estimates of correlation matrices are not guaranteed to be positive definite, and this issue naturally escalates under high data dimensionality (Demarta and McNeil, 2005). As for estimators of the type (1.11) based on the sample correlation, they are also sensitive to data dimensionality, as the sample correlation matrix is positive definite if and only if the sample size strictly exceeds data dimensionality.

For the same reason, the estimator based on the exact relation (1.5) is preferred in bivariate case, otherwise there is no guarantee the resulting estimate of the matrix parameter of higher dimensionality will be well-conditioned. The numerical errors of the approximation (1.6) are relatively small, both for Gaussian and t copulas, and the estimator (1.11) turns out precise enough and better-conditioned.

However, these traditional estimators are expected to lose quality under high data dimensionality, which brings forward the main point of this paper. The next subsection briefly covers the basics of shrinkage estimators of large covariance matrices and explains how they can be used to estimate copula matrix parameters.

1.3.3 Shrinkage estimation of copula matrix parameters

Over the years, researchers have come up with a variety of estimators of large covariance matrices to restore the properties of the sample covariance under high dimensionality (Fan et al., 2008; Ledoit and Wolf, 2004b, 2017b). In this paper, to estimate the large matrix

parameters of Gaussian and t copulas, we use the shrinkage estimators of Ledoit and Wolf (2004b, 2017b). These estimators have proved to perform well in general settings of large covariance matrices estimation, and they allow one to take the analysis to the highest data dimensionality achieved so far (Ledoit and Wolf, 2017b; Engle et al., 2019).

The idea behind the shrinkage estimators is the following. Given a p-dimensional random vector X from some distribution F characterized by zero mean (without loss of generality) and some non-random positive-definite covariance matrix $\Sigma = \mathbb{E}[XX'] = \text{cov}(X)$, and an i.i.d. sample of size n from that distribution recorded into $n \times p$ matrix $X_n = \{X_{ti}\}_{t=1,\ldots,n;i=1,\ldots,p}$, the population covariance matrix Σ can be estimated by the sample covariance matrix

$$S_n = \frac{X_n' X_n}{n}. \tag{1.13}$$

The estimator S_n is consistent and well-conditioned under standard asymptotics when p is fixed and $n \to \infty$. However, in high dimensions the sample covariance matrix is not well-conditioned when p is non-negligible compared to n, and even non-invertible for p larger than n. Ledoit and Wolf (2004b) follow the work of Haff (1980) and construct the linear shrinkage estimator as a linear combination of a structural covariance matrix estimator (an equivariate diagonal covariance matrix) and the sample covariance matrix (1.13):

$$\Sigma^* = \rho_1 I_p + \rho_2 S_n. \tag{1.14}$$

However, unlike in the work of Haff (1980), Ledoit and Wolf (2004b) managed to derive the optimal estimator Σ^{**} that minimizes the Frobenius norm of the deviation from the population covariance matrix Σ, $||\Sigma^{**} - \Sigma||^2 = p^{-1}\text{trace}\left[(\Sigma^{**} - \Sigma)(\Sigma^{**} - \Sigma)'\right]$. Next, since the estimator Σ^{**} is not feasible as it depends on the unknown Σ, it itself needs to be estimated. The feasible estimator that can be calculated directly from the data takes the form

$$S^* = \hat{\vartheta}\hat{\mu} I_p + (1 - \hat{\vartheta}) S_n, \tag{1.15}$$

where the coefficients $\hat{\vartheta}$ and $\hat{\mu}$ depend on the data X_n, see the definitions in Lemmas 3.2–3.4 in Ledoit and Wolf (2004b). This estimator is positive definite and consistent for the population

covariance matrix Σ under dimension asymptotics, that is, under $p \to \infty$, $n \to \infty$, and $c \equiv p/n \to \bar{c} \in (0, \infty)$. The value $\hat{\vartheta}$ is called *shrinkage intensity*. The less accurate the sample covariance matrix S_n is, the more it will be shrunk, i.e., more weight in (1.15) is put on the structural estimator (Ledoit and Wolf, 2017b).

An important characterization of the linear shrinkage estimator is in terms of eigenvalues of the covariance matrix. Given that Σ is characterized by its eigenvalues $\lambda_1, ..., \lambda_p$ (let without the loss of generality $\lambda_i \leq \lambda_j \ \forall i < j$), and if $l_1, ..., l_p$ are the eigenvalues of the sample covariance matrix S_n, it is proved that the population and sample eigenvalues share the same grand mean (Ledoit and Wolf, 2004b):

$$\mu = \mathbb{E}\left[\frac{1}{p}\sum_{i=1}^{p} l_i\right] = \frac{1}{p}\sum_{i=1}^{p}\lambda_i. \tag{1.16}$$

Also, Ledoit and Wolf (2004b) show that

$$\frac{1}{p}\,\mathbb{E}\left[\sum_{i=1}^{p}(l_i - \mu)^2\right] = \frac{1}{p}\sum_{i=1}^{p}(\lambda_i - \mu)^2 + \mathbb{E}||S_n - \Sigma||^2. \tag{1.17}$$

Thus, the sample eigenvalues are relatively more dispersed than the population ones, and the excess dispersion exactly equals the expected loss of the sample covariance matrix. Further, as there is particular over-dispersion around the same mean, the higher eigenvalues are biased upward, while the lower ones are biased downward.

Essentially, the shrinkage estimator (1.15) reduces the bias of the sample covariance matrix eigenvalues by shifting them towards their grand mean (1.16) shrinking the distribution of the sample eigenvalues. The shrunk eigenvalues corresponding to the optimal linear shrinkage estimator (1.15) are

$$\lambda_i^* = \vartheta\mu + (1 - \vartheta)l_i, \tag{1.18}$$

where the coefficients ϑ and μ are probability limits, under dimension asymptotics, of $\hat{\vartheta}$ and $\hat{\mu}$ in (1.15), and so the shrunk eigenvalues can then be estimated from the data similarly to how the estimator (1.15) estimates (1.14):

$$l_i^* = \hat{\vartheta}\hat{\mu} + (1 - \hat{\vartheta})l_i, \tag{1.19}$$

and the shrinkage estimator then can be rewritten as a rotation equivariant estimator:

$$S^* = \Gamma_n \text{diag}\{l_i^*\}_{i=1,...,p} \Gamma_n', \tag{1.20}$$

where $\Gamma_n = [\gamma_{n,1}, ..., \gamma_{n,p}]$ is the matrix of sample covariance matrix eigenvectors $\{\gamma_{n,i}\}_{i=1,...,p}$. Later, Ledoit and Wolf (2012) studied the performance of their linear shrinkage estimator and found that it often results in under-shrinkage, i.e. the resulting distribution of sample eigenvalues of the estimator (1.20) is still considerably over-dispersed as compared to the population distribution of eigenvalues of Σ. In their study, Ledoit and Wolf (2012) use the same approach to upgrade to the nonlinear shrinkage by applying different shrinkage intensities to eigenvalues of different magnitude. They build on the work Ledoit and Péché (2011) and show how a feasible estimator can be constructed, in a way similarly to how the optimal linear shrinkage estimator (1.15) estimates the non-feasible estimator (1.14). The non-linear shrinkage estimator preserves the form of the rotation equivariant estimator (1.20), with the linearly shrunk eigenvalues l^*s (1.19) replaced by the non-linearly shrunk versions:

$$l_i^{**} = \frac{l_i}{|1 - \frac{p}{n} - \frac{p}{n} l_i \breve{m}_F^*(l_i)|^2}. \tag{1.21}$$

Here, $\breve{m}_F^*(l)$ is the shrinkage intensity term that depends on sample eigenvalue l. The construction of this term is presented in detail in Section 5 of Ledoit and Wolf (2012).

The intuition behind the estimator is the following. The linear shrinkage performs well when the sample eigenvalues are not too dispersed so that the constant shrinkage intensity is sufficient to shift the distribution of the sample eigenvalues closer to the population analog. However, with a higher dimensionality p/n and sample eigenvalues far from the grand mean appearing more frequently, treating the sample eigenvalues differently is likely to pay off. The estimator of nonlinear shrinkage intensity $\breve{m}_F^*(l_i)$ aims to make the estimator of the asymptotic distribution of eigenvalues as close to the actual limiting distribution of the sample eigenvalues as possible (Ledoit and Wolf, 2012). The resulting estimator is proved to be asymptotically equivalent to the optimal one in terms of Frobenius loss in the class of rotation equivalent estimators of Ledoit and Péché (2011), and thus can outperform the linear

shrinkage estimator (Ledoit and Wolf, 2012). However, implementation of the estimator requires numerical inversion of a particular multivariate nonrandom function, which was later efficiently implemented by Ledoit and Wolf (2017b).

We employ these shrinkage estimators in estimation of the high dimensional correlation matrices of Gaussian and t copulas. The shrinkage estimators are to substitute the sample correlation-based estimator (1.11). Since the shrinkage estimators estimate the population covariance matrix, we transform them to estimates of the correlation matrices to comply with the structure of the copulas' parameters P. It is done by a simple covariance-to-correlation transformation, which cannot change the estimator being well-conditioned or not.

Another possible concern when estimating copulas is that the shrinkage estimators and their properties rely on i.i.d. data samples, while in copula estimation the pseudo-observations (2.14) are not independent. Still, the same issue arises when implementing the MPLE, yet the resulting estimates are shown to be relevant and insignificantly different from the FMLE. Hence, we expect that disregarding the actual "non-iid-ness" of pseudo-observations and applying the shrinkage estimators will perform sufficiently better than the traditional estimators (1.11) and (1.12).

1.4 Simulation study

In this section, we present the results of our simulation study. We consider a variety of Gaussian and t copulas with different values of matrix parameters. We vary both the number of variables in the data and its ratio to the sample size in order to track the performance of the estimators under low and high dimensionality. The estimation quality is evaluated both in terms of closeness of matrix parameter estimates to the true matrix parameter values and closeness of estimates of copula functions to their true counterpart.

When working with the t copula, the degrees of freedom parameter ν needs to be estimated as well. We avoid describing technical details of this estimation; it is a basic uni-dimensional estimation performed via MPLE treating the matrix parameter fixed at its estimated (via one of the moments-like estimators) level. Neither do we report the estimation results of these parameters; the estimates $\hat{\nu}$ are generally very close to the true values and do not

cause any problems. Similarly, we do not focus on details or results of estimating the marginal distributions. We use univariate empirical distributions (EDF) to construct the pseudo-observations (2.14) from the original data.

Next, we present the choice of copula parameters and estimation quality criteria. Then we present simulation design and report the results.

1.4.1 Simulation design

True copula specifications The following specifications of the copulas are used in the simulations:

- The true copulas are either Gaussian or t.
- The data dimensionality p takes one of three values

$$p \in \{10, 100, 1000\}. \tag{1.22}$$

- The sample size is set via fixing particular values of the p-*to*-n ratios to compare the cases of different dimensionality. Generally, we consider the range of the dimensionality ratio from $1/20$ to 20 except the cases with a small number of variables ($p = 10$) and dimensionality higher than 2 (as they imply the sample size of $n < 5$), and the cases with a large number of variables ($p = 1000$) and dimensionality lower than $1/2$ (as they imply sample sizes higher than 2000 which is too computationally demanding). To summarize, the dimensionality varies in the following way:

$$\frac{p}{n} \in \begin{cases} \{1/20,\ 1/10,\ 1/2,\ 1,\ 2\}\,, & p = 10, \\ \{1/10,\ 1/2,\ 1,\ 2,\ 5,\ 10\}\,, & p = 100, \\ \{1/2,\ 1,\ 2,\ 5,\ 10,\ 20\}\,, & p = 1000. \end{cases} \tag{1.23}$$

- For each copula and all pairs of dimensionality and sample size we consider two versions of the true matrix parameter P. First, we use the identity structure $P = I_p$ as an

important benchmark case. Second, for each p we construct an arbitrary and randomly generated matrix parameter P, which is a legit correlation matrix as it is positive definite, far from being degenerate, and has a full range of values for correlation coefficients. The three non-identity matrices are visualized in Figure 3.

- For the t copulas, the degrees of freedom parameter value is always fixed at $\nu = 8$ so that the copulas are sufficiently far from being Gaussian, but also are sufficiently distant from the value of 2 when variance does not exist.

- The marginal distributions are set to univariate standard skewed-t distribution with randomly and independently assigned degrees-of-freedom and skewness parameters. The degrees-of-freedom parameter is drawn from a discrete uniform on $\{6, 7, 8, 9, 10\}$, and the skewness parameter is drawn from $U[-1, 1]$.

Measures of estimation accuracy

Given some true model $C_P(u)$ with the $p \times p$ matrix parameter P and its estimate $\hat{P}$ we evaluate estimation quality using the following three measures:

- *Positive-definiteness.* As all true matrix parameters P are legit correlation matrices, it is a desirable property of the estimates $\hat{P}$ to be such, too. By construction, all estimators we consider deliver $\hat{P}$ that are symmetric with unit diagonal elements and correlation coefficients off the diagonal. Positive-definiteness, however, is not guaranteed for some of the estimators; hence, for every $\hat{P}$ we check whether they satisfy this property. The shrinkage-based estimators deliver positive-definite matrices by construction; still, we assess their positive-definiteness as a sanity check for numerical routines.

- *Closeness of matrix estimate to true values.* Given that the matrix parameters are symmetric, there is a wide choice of measures of closeness of estimates to true values. However, since the matrices at hand are correlation matrices, it is sufficient to measure the closeness of elements off the main diagonal. We use the Euclidean norm of the difference between the half-vectorized true and estimated matrices:

$$L_E(P, \hat{P}) = ||\text{vech}(P - \hat{P})||. \tag{1.24}$$

Note that the use of Frobenius matrix norm would deliver the same rankings because the diagonal elements in both matrices are fixed.

- *Closeness of estimated copula function to true one.* Finally, as the main object of modeling is the copula function C_P itself, we measure the closeness of the estimated one to the true one via the Kullback-Leibler information criterion (KLIC):

$$KLIC_{P|\hat{P}} = \mathbb{E}_{C_P}\left[\log\left(\frac{c_P(u)}{c_{\hat{P}}(u)}\right)\right] = \int\cdots\int_{\mathbb{O}_p}^{\mathbb{1}_p} c_P(u)\log\frac{c_P(u)}{c_{\hat{P}}(u)}\mathrm{d}^p u. \tag{1.25}$$

While the first two criteria are computationally practical even when p is large, calculating KLIC for large p is computationally demanding. To make it operational, we do two simplifications. First, we use the property that Gaussian and t copulas of larger vectors remain the same for their sub-vectors (see Section 1.3.1), so for any data dimensionality p we only consider KLIC for 3-dimensional subsets of the data. For $p = 10$, we compute the KLIC for only one triplet; for $p = 100$, we average KLIC over randomly chosen 30 triplets, and for $p = 1000$ the number of triplets we average over is 100. Second, we estimate the expectation in (1.25) via simulations. For each true copula function $C_{\tilde{P}}(u)$ (where $\tilde{P}$ is a 3×3 matrix parameter corresponding to a chosen triplet and the initial true matrix P), we generate a collection of $M = 10^6$ 3-dimensional vectors $\{\tilde{u}_m\}_{m=1,\dots,M}$ from the true copula function $C_{\tilde{P}}$, and estimate the expectation in (1.25) using the expressions for log-densities of Gaussian and t copulas (1.8) and (1.9):

$$KLIC_{\tilde{P}|\tilde{\hat{P}}} = M^{-1}\sum_{m=1}^{M}\left(\log c_{\tilde{P}}(\tilde{u}_m) - \log c_{\tilde{\hat{P}}}(\tilde{u}_m)\right). \tag{1.26}$$

Simulation design

For a particular combination of number of variables p, true matrix parameter P, marginal distributions $\{F_i\}_{i=1}^p$, true copula function $C_P(u)$, and sample size n, a single simulation is run as follows.

1. We generate the data $X \in \mathbb{R}^{n\times p}$ from $C_P(F_1(u_1),...,F_p(u_p))$, estimate the marginals via EDFs, and transform them to pseudo-observations, $U = \{\hat{F}_i(x_i)\}_{i=1}^p \in [0,1]^p$.

2. We estimate $\mathrm{corr}(U)$ via each of the four estimators and obtain estimates $\hat{P}^{\text{smpl}}$, $\hat{P}^{\text{i-}\tau}$,

$\hat{P}^{\text{LSh}}$ and $\hat{P}^{\text{NLSh}}$.

3. For each estimate, we calculate the following accuracy measures:

   - a binary indicator of positive-definiteness of $\hat{P}$;
   - the Euclidean loss, $L_E(P, \hat{P})$, via (1.24),
   - KLIC, via (1.26) and averaged over randomized triplets of variables;
   - * for t copulas, KLIC are estimated twice: once treating the degrees-of-freedom parameter as known, and then with that estimated by MPLE.

We repeat each simulation 2^{10} times.[8]

1.4.2 Simulation results

The simulation results are presented Tables SA1 – SA12 in the Supplementary Appendix. For each evaluation criterion, we report the median, mean and standard deviation across the simulations. The median values of the criteria are visualized in Figures SA1, SA2. When calculating KLIC for non-positive-definite $\hat{P}$, there is a great chance that the estimate of the expectation does not converge, resulting in an "infinite" value of KLIC. In most of these cases, the median can still be computed (unless KLIC is infinite in all the simulations), but the mean and standard deviation make no sense due to a high share of infinite values. Next, in some cases either the median or the mean and standard deviation are numerically indistinguishable from zero, i.e. they are $< 10^{-23}$. In measuring the performance in terms of any criterion, we say that one estimator outperforms another if the median value of the former estimator's performance criterion is smaller than that of the latter estimator.

The results of the positive-definiteness check are perfectly predictable and appear as expected. The shrinkage estimators always deliver positive-definite estimates of the matrix parameter. The traditional estimators deliver positive-definite estimates only under low dimensionality

[8] The format of a power of two is chosen due to technical reasons of multi-core calculation organization. A higher number of simulations appears very time consuming under large p and n, and the number 2^{10} resulted in sufficiently precise calculations to make the conclusions.

($p/n < 1$), with $\widehat{P}^{\text{i-}\tau}$ not necessarily positive-definite even then (though the fraction of such cases is small).

The case of $p = 10$ is included to show the basic properties of the four estimators and to point out that the ratio of the number of dimensions in the data to the sample size does matter (see Tables SA1–SA4). More importantly, the difference in performance is well observed for higher dimensions and smaller samples (Tables SA9–SA12).

Regarding the two distance criteria, overall the shrinkage estimators confidently outperform the traditional ones. First, under low dimensionality, there is no clear pattern in which type of estimator is the best in terms of the closeness of the estimated matrix to its true counterpart. However, there are very few cases when one of traditional estimators outperforms one of the shrinkage estimators in terms of Euclidean distance. Further, even when the traditional estimators do outperform the shrinkage ones in terms of Euclidean distance, the KLIC are likely to be smaller for the shrinkage estimators.

Second and most interesting, under high dimensionality, the better performance of shrinkage estimators is more obvious. Not only are the estimates always positive definite, but they are also precise enough in terms of both Euclidean distance and KLIC, and the difference in the performance of the shrinkage estimators and traditional ones is substantial. Figure 4 gives a representative picture for a slice of simulations.

Regarding the relative performance of the shrinkage estimators to each other, we additionally report several selected slices of the joint distributions of their performance to check how often each of the estimators outperforms the others, and how that changes with higher dimensionality. This is reflected in Figure 5.

Overall, under high dimensionality ($p/n > 1$), there is a tendency for nonlinear shrinkage based estimators of copulas, both Gaussian and t, to outperform linear shrinkage based either in terms of Euclidean distance between the true and estimated matrix parameter, or the average Kullback-Leibler distance between the true and estimated copula function. Further, the higher the dimensionality, the more likely the nonlinear shrinkage will perform better than the linear one (see, for example, Figure 5). However, there are a few exceptions. First, for either copula with rather dispersed true eigenvalues (e.g., the 100×100 arbitrary true

matrix P in our simulations), the linear shrinkage outperforms the nonlinear one under high dimensionality (see Figures 4 and 5c). We conjecture that the relatively better performance of nonlinear shrinkage for the models with less-dispersed true eigenvalues (e.g., the identity P in our simulations) is explained by the ability of nonlinear shrinkage to shift the right tail (outlier) sample eigenvalues towards the grand mean. Second, there may be a situation (see, e.g., Figure 5d) in which the linear shrinkage based estimator dominates all others, with the nonlinear shrinkage, in this case, only slightly underperforming (see Table SA12c), and the differences between the two can be neglected.

1.5 Empirical illustration: large portfolio allocation

We apply shrinkage based estimators of copula correlation matrices in high dimensions to allocate large portfolios of stocks and compare their performance with portfolio choices derived from the plain multivariate normal (MVN) model.

Asset allocation is one of the classical applications of multivariate models of assets returns. A number of theoretical settings describing investor's behavior offer analytical solutions for a portfolio structure. However, the more complicated the investor's problem is or the more sophisticated the model for asset returns is, the more likely numerical methods need to be employed for an optimal portfolio choice (DeMiguel et al., 2007; Michaud and Michaud, 2008; Guidolin and Timmermann, 2008; Kolm et al., 2014; Ledoit and Wolf, 2017a). Even in the static case, when the portfolio structure is determined only once per portfolio lifetime, it often appears necessary to simulate the dynamics of asset returns over a portfolio lifetime period to evaluate the performance of different portfolios and pick the optimal structure corresponding to investor's utility function (van Binsbergen and Brandt, 2007; Guidolin and Timmermann, 2008; Harvey et al., 2010).

We perform a static portfolio allocation exercise, i.e. the structure of the portfolio is going to be set once per portfolio lifetime. However, the joint distribution model of asset prices during the portfolio lifetime is based on empirical marginal distributions of asset returns and copula across assets' dependence structure. Hence, simulations of asset price dynamics are required to evaluate the value of portfolios during and at the end its lifetime.

We use historical data from the database *FIZ©2019*.[9] From the CRSP dataset we extract daily close prices of the securities listed in the Wilshire 5000 index for the last 9 months of 2017. There are 4982 assets at our disposal. We randomly choose subsets of size 3600 assets to model the predictive joint distribution of their prices. Based on this model, we simulate future prices and select portfolios with the best Sharpe ratio. To evaluate these portfolios, we compare their actual performance over the period of simulation with the performance of the equally weighted portfolio, or the portfolios based on other models, in terms of cumulative return in the end of portfolio lifetime.

Prior to estimating predictive multivariate distribution, we filter out univariate conditional means and conditional variances of each log-return via ARMA-EGARCH modeling, and extract serially uncorrelated standardized residuals. Then, one of the following multivariate distribution models is applied to these residual terms across the assets:

- MVN,
- t copula, with the marginals estimated as EDFs.

We use either linear or nonlinear shrinkage estimators to estimate the matrix parameter of both the MVN and the t copula models. The d.f. parameter of the t copula is estimated via MPLE. In this exercise we drop the sample correlation estimator of the matrix parameter of either MVN or t copula due to the high dimensional context ($p/n = 30$), and the i-τ estimator for the copula is dropped due to its poor performance shown in simulation results earlier. We use only the t copula as it includes the Gaussian copula as a special case.

Thus, for each set of 3600 assets we obtain 4 different model-based portfolios, each of which is the optimal portfolio in terms of Sharpe ratio corresponding to one of the 4 estimates. To account for differences among randomly chosen subsets of assets, we measure the performance of these portfolios relative to each other or to the return of the equally-weighted portfolio.

The detailed description of the modeling technique and simulation design are relegated to Appendix B. We use historical data over the period of the last 9 months of 2017, with the first 6 months used to fit the models, and the last 3 months used as an out-of-sample period,

[9] Center for Research in Security Prices (CRSP), University of Chicago Booth School of Business.

over which the simulations are run and the performance of the portfolios is evaluated. The distributions of relative performance of portfolios suggested by different models and estimates across the randomly chosen sets of assets are shown in Figure 7. Figure 6 gives examples of dynamics of different model-based portfolio cumulative return in comparison with the one for equally-weighted portfolios.

The intuition behind this approach is the following. The performance of model-based portfolio choices crucially depends on whether the model is capable of capturing the properties of returns properly. In the case of MVN, not only does the model disregard heavy tails and asymmetry in return marginal distributions, but also it ignores possible tail dependence. The resulting portfolios are likely to be vulnerable to the shocks that are rare, but occur simultaneously in the returns of many assets included in the portfolio. Although the t copula based model is also rather limited in capturing the desired properties (only symmetric tail dependence can be captured), it still is able to improve the quality of the portfolios exactly because the assets that are likely to be tail dependent will not be included in the same portfolio with high weights. Further, given the results presented earlier, we expect that under high dimensionality ($p/n = {}^{3600}/_{120} = 30$ in this case), the shrinkage-based estimates of the t copula based models are to deliver more relevant portfolio choices.

The results do confirm this. Overall, from our 135 randomly chosen sets of assets we find that in over 74% of cases the best portfolio is suggested by either of the t copula based models, in about 13% the best portfolio is the model-free equally weighted one, and the rest are the MVN-based choices. Further, when a portfolio is suggested by either MVN or t copula model, it is more often the one based on the nonlinear shrinkage estimator of the matrix parameter. However, in case of the t copula estimates, in over 63% of cases the performance of the two portfolios is indistinguishable in terms of the cumulative return in the end of portfolio lifetime. In terms of relative performance of the models, for t copula based portfolios there is a considerable chance that the resulting return at the end of portfolio's lifetime is going to be higher than the corresponding return of any other portfolio (see Figure 7).

We have intentionally designed this example so that it over-simplifies the dynamic component of the returns modeling, but instead reveals and stresses the potential benefits in the high-

dimensional context. First, we took the number of assets to what, to our knowledge, is the highest dimensionality of portfolios analyzed via copulas. Second, the model is estimated on a (relatively) extremely small sample, which justifies using a very simple dynamic model for asset returns. We believe that this approach can be further developed for the task of dynamic re-balancing of large portfolios.

1.6 Discussion and concluding remarks

We employ large covariance matrix shrinkage estimators in the task of Gaussian and t copulas estimation in high dimensions. This technique allows us to precisely estimate the copulas in (ultra-)high dimensions with up to 1000 variables in a dataset and sample sizes up to 20 times smaller. While it is accepted that the copulas we study cannot capture all of data properties in all empirical applications (e.g., asymmetric dependence, including that in the tail), they remain favored in numerous applications either as a main dependence model or at least as important benchmark models and building blocks for more flexible settings. Many applications that employ the Gaussian and t copulas can benefit from higher dimensionality either by including more variables into the datasets, or by making use of smaller samples.

Our main results show that large covariance shrinkage estimators can effectively be used for copula matrix parameter estimation in (ultra-)high dimensions. Not only are the resulting estimates of the correlation matrices of the pseudo-observations well-conditioned and close to their true values, but also the whole copula function estimates are close to their actual counterparts, including t copulas, for which the scalar degrees-of-freedom parameter controlling for tail dependence is additionally estimated by MPLE. In addition, we show that the non-linear shrinkage estimator generally outperforms the linear one, except when the true matrix parameter is rather sparse, in which case the performance of the two shrinkage estimators is indistinguishable.

Obviously, it is potentially very beneficial in future research to extend the approach we have proposed to other copula-based settings, such as skewed versions of Gaussian and t copulas that are known to be able to capture asymmetric dependence. In this paper, we heavily exploit the symmetry to be able to connect the correlation matrix of the pseudo-

observations with the actual parameters of the copula function. This makes estimation of the actual copula parameters practical. However, we conjecture that there is no obstacle in extending the approach to the estimation of correlation matrices of pseudo-observations for other copulas, including skewed ones. However, it is not operational since for copulas other than Gaussian or t the parameters of copula functions cannot be easily connected with moments of pseudo-observations. One possible way to overcome this is to use the idea of simulated method of moments for copula estimation of Oh and Patton (2013) combined with shrinkage estimation of the covariance matrix of pseudo-observations. Again, currently the approach is rather computationally impractical in high dimensions. Another way to approach it would be to introduce a two-step-like estimation, when on the first step one estimates the lower-dimensional parameters of the copula so that to transform the pseudo-observations according to the quantile functions of the underlying distribution of the copula and use its properties to estimate, on the second stage, the matrix parameter via shrinkage estimators. Potentially, shrinkage estimators application can be extended to skewed versions of elliptical copulas or other implicit copulas. We see this idea potentially very beneficial, yet it requires substantial further investigation.

What may be a beneficial and computationally practical extension of the current approach is to use the most recent advances in non-linear shrinkage estimation of large covariance matrices. In particular, the recently suggested analytical non-linear shrinkage of Ledoit et al. (2020) makes the non-linear shrinkage estimator easier and faster to implement. Similarly, the quadratic shrinkage of Ledoit and Wolf (2019) is potentially beneficial for practical application. According to the authors, it is unlikely that either of these estimators will improve the quality of estimation as compared to earlier numerical implementation of the non-linear shrinkage. We ran a separate short simulation study of this issue confirming that the gain of the analytical non-linear shrinkage is only in terms of computational time.

Another result of our research is an empirical application of the proposed copula estimators to a large portfolio allocation problem. We use the high-dimensional t copula to model the joint distribution of returns of (ultra-)many assets over a short period and construct large portfolios. With the number of assets in the portfolio of 3600 and the sample length for

model estimation of 120 observations, the problem is ultra-high dimensional and, to our knowledge, the highest dimensional portfolio allocation problem in the literature. Hence, precise estimation of the model requires shrinkage estimation of matrix parameters. The results show that although the t copula is symmetric, the suggested portfolios significantly outperform those coming from the multivariate normality or the copula model estimated by traditional estimators. Not only do the portfolios deliver higher returns by the end of the lifetime, but also they persistently avoid substantial downfalls during the lifetime due to accounting for and proper estimation of tail dependence.

The results of the empirical exercise also suggest that the proposed approach can be beneficial for constructing more sophisticated multivariate dynamic models for financial asset returns, particularly if one succeeds in practically applying it to the case of skewed copulas. Alternatively, these results can be used to update some of the existing approaches to modeling the joint dynamics of many assets' returns that yet disregard the the dependence between the variables beyond correlations. For example, Engle et al. (2019) use non-linear shrinkage to bring the dynamic conditional correlation model of assets' returns into high dimensions and use it to construct large portfolios, and De Nard et al. (2020) bring the analysis to even higher dimensions and intra-day data frequency. Yet the standardized innovations follow simple multivariate normal distribution. Our empirical example suggests that a copula-based setting in the part of standardized innovations distribution modeling can be beneficial for the emerging portfolios, and shrinkage estimation is a practical way to keep the whole setting high-dimensional.

2 Estimation of High-Dimensional Skew-t Copula and Application to Portfolio Allocation

2.1 Introduction

Selecting or constructing an appropriate multivariate distribution plays a fundamental role in a wide array of practical statistical applications. While many families of multivariate distributions are commonly used, they often exhibit limitations in their ability to capture all essential data properties.

Copulas have emerged as a potent tool for constructing multivariate distributions. They provide the means to model individual marginal distributions and the interdependence structure, referred to as the copula, separately within a multivariate distribution.

One of the current challenges in copula applications is that, in the contemporary landscape, there is a growing need for modeling frameworks that are effective in high-dimensional settings. Despite the multitude of available copulas, many are unsuitable for high-dimensional applications due to practical constraints.

Specifically, in financial contexts, the estimation of multivariate distributions for asset returns has remained a pivotal endeavor. Over the years, researchers have continually refined these models to better capture essential data properties, enabling more informed decision-making in areas such as pricing, asset risk management, and portfolio allocation.

While univariate processes have seen significant advancements, the modeling of multivariate asset return distributions has lagged behind. Much of the literature in this domain focuses primarily on modeling joint second moments. Aspects of financial data, such as asymmetric dependence and tail dependence, hold significant importance in practical applications. Nevertheless, these properties have received relatively limited attention in the existing literature.

In this paper, we investigate the skew-t copula, based on the multivariate skew-t distribution introduced by Azzalini and Capitanio (2003). This copula leverages the foundation of elliptical

distributions, enabling easy extension to high dimensions. Moreover, owing to its inheritance of the skewness and heavy tails of the skew-t distribution, it possesses the capability to model asymmetry and tail dependence. However, its analytical characteristics are limited, rendering estimation of it challenging, even in scenarios with moderate dimensions.

We introduce a new method for estimating the skew-t copula in both low and high dimensions. Our approach takes advantage of the convenient stochastic representation of the underlying distribution and involves a two-step procedure for estimating the copula's parameters. In the first step, we use the simulated generalized method of moments to identify and estimate the lower-dimensional skewness and degrees of freedom parameters of the copula. These estimates help us transform the data from the copula distribution into a representative sample that aligns with the underlying multivariate skew-t distribution. In the second step, we apply analytical non-linear shrinkage of Ledoit et al. (2020) to the transformed data, resulting in a well-conditioned estimate of the large matrix parameter. We then put the skew-t copula and our proposed estimator to the test in practical portfolio allocation within high-dimensional settings.

To demonstrate the real-world relevance of our approach, we construct a series of portfolios using the top 30 components of the EUROSTOXX50 index in the first half of 2022. Our dynamic portfolio exhibits significant outperformance of the market. A closer look at the changes in the portfolio structure highlights the skew-t copula's crucial role in accounting for tail dependence among assets.

The rest of the chapter is organized as follows. In section 2.2 we give an overview of the multivariate skew-t and its copula. Section 2.2.3 presents the suggested two-step algorithm of skew-t copula estimation. Section 2.4 contains a summary and discussion of the empirical exercise. Section 2.5 concludes the chapter.

2.2 Multivariate skew-t distribution and its copula

2.2.1 Sklar's theorem and foundations of copulas

A copula is a multivariate cumulative distribution function (CDF) with $U[0,1]$ marginals. Such CDFs are vital for multivariate distribution construction due to the key result in copula theory – Sklar's theorem (Sklar, 1959). The theorem shows that any multivariate distribution with CDV $F(x)$ and marginal CDFs $\{F_i(x_i)\}_{i=1,...,p}$ has a unique representation

$$F(x) = C\Big(F_1(x_1), ..., F_p(x_p)\Big), \tag{2.1}$$

and the function $C : [0,1]^p \to [0,1]$ is called the copula function of F.

The converse of the theorem also holds and allows one to combine a copula function of one distribution with a set of any marginal CDFs $\{F_i'(x_i)\}_{i=1,...,p}$ to plug them into (2.1) and obtain a legitimate new CDF $F'(x)$ with those marginals. Thus, a copula embeds the structure of the interdependence in a multivariate distribution, that can be modeled and analyzed separately from the marginal distributions.

The copula approach to multivariate distribution modeling has been widely used in the research, and it engendered a vast variety of kinds of copulas. While Archimedean copulas proved very convenient in bivariate settings, they are not flexible when extended to moderate or high dimensions (Hofert et al., 2012). Hence, Archimedean copulas are rarely used outside bivariate settings. The vine copulas allow one to attain maximal flexibility in copula construction, but at the cost of a growing number of alternative specifications and parameters with higher data dimensionality. Often, constructing a vine copula requires one to make rather strict structural assumptions about the data-generating process and to use heuristic algorithms for specification selection (Aas et al., 2009; Brechmann et al., 2012; Brechmann and Czado, 2013; Czado et al., 2013; Dissmann et al., 2013). Another approach to constructing a copula is to directly extract the copula function from a multivariate distribution (2.1). For example, the elliptical copulas are copulas of elliptical distributions. The most famous members of this class are Gaussian and t copulas, derived correspondingly from the multivariate normal and

Student's t distributions (Demarta and McNeil, 2005). Elliptical copulas are often used in the construction of multivariate distributions to capture desired properties of the data (Zimmer, 2012; Patton, 2012; De Leon and Chough, 2013; Patton, 2013; Oh and Patton, 2017).

Thus, finding a relevant copula for a particular task of constructing a multivariate distribution seeks a compromise between the ability to capture all the desired properties of the data and how practical it is in applications to actual data. In this paper, we consider the copula of multivariate skew-t distribution that builds on elliptical copulas. On one hand, it inherits the skewness of the skew distribution that allows it to capture asymmetrical dependence and tail dependence. On the other hand, the lack of analytical properties makes it sufficiently difficult to apply, particularly in high-dimensional settings.

In the next sections, we first introduce the multivariate skew-t distribution and its properties that are essential for skew-t copula analysis. Next, we introduce the corresponding copula and the approach to its estimation that we suggest.

2.2.2 Multivariate skew-t distribution of Azzalini and Capitanio

In the literature, there are several different distributions under the label of "multivariate skew-t distribution" (MSTD). They are generally built by introducing additional skewness to the multivariate t distribution. In our study, we stick to the definition of MSTD by (Azzalini and Capitanio, 2003). This definition introduces a convenient stochastic representation of the distribution that corresponds to our purposes. Moreover, there are several other studies that introduce skew-t copula based on this distribution (Yoshiba, 2018).

The stochastic representation of the standardized distribution is the following[10]. A p-dimensional random vector Y is distributed according to MSTD with parameters $\{P, \delta, \nu\}$, $MSTD(P, \delta, \nu)$, with P $p \times p$ correlation-like[11] matrix, δ $p \times 1$ vector from $[-1, 1]^p$, scalar

[10] The original definition of MSTD by Azzalini and Capitanio (2003) is an equivalent but a different parameterization in terms of the parameters $\{P, \alpha, \zeta, \nu\}$ introduced below. We stick to the parameterization in terms of $\{P, \delta, \nu\}$ that is more convenient for the copula definition, random numbers generation, and estimation.

[11] symmetric, positive-definite, unity diagonal, from $[-1, 1]$ off-diagonal elements matrix

$\nu > 2$, if the extended matrix

$$\Omega = \begin{pmatrix} 1 & \delta' \\ \delta & P \end{pmatrix} \tag{2.2}$$

is positive-definite, and

$$Y = V^{-1/2}\tilde{Z}, \tag{2.3}$$

where the scaling scalar r.v. $V^{-1/2}$ is such that

$$\nu \cdot V \sim \chi^2_\nu, \tag{2.4}$$

and the p-variate r.v.

$$\tilde{Z} = \begin{cases} Z, \text{ if } Z_o \geq 0, \\ -Z, \text{ o.w.}, \end{cases} \tag{2.5}$$

with

$$\begin{pmatrix} Z_o \\ Z \end{pmatrix} \sim \mathcal{N}(\mathbb{O}_{p+1}, \Omega), \tag{2.6}$$

and V is independent of $(Z_o \; Z')'$.

Neither the CDF of this distribution, nor the marginal CDFs, nor the marginal quantile functions have closed-form representations. The PDF can be represented in the following way:

$$f_{MSTD}(Y; P, \delta, \nu) = 2 f_{t_p}(Y; P, \nu) F_{t_1}\left(\alpha' Y \sqrt{\frac{\nu + p}{Y' P^{-1} Y + \nu}}; \nu + p\right), \tag{2.7}$$

where $f_{t_p}(.; P, \nu)$ is the PDF of p-variate Student's t distribution with correlation matrix P and ν degrees of freedom, $F_{t_1}(.; \nu)$ is the CDF of univariate Student's t distribution with ν d.f., and

$$\alpha = \frac{P^{-1}\delta}{\sqrt{1 - \delta' P^{-1} \delta}}. \tag{2.8}$$

The marginal distributions of $MSTD(P, \delta, \nu)$ are the univariate skew t distributions with parameters $\{\delta_i, \nu\}$, $STD(\delta_i, \nu)$. The corresponding PDF is

$$f_{STD}(Y_i; \delta_i, \nu) = 2 f_{t_1}(Y_i; \nu) F_{t_1}\left(\zeta_i Y_i \sqrt{\frac{\nu + 1}{Y_i^2 + \nu}}; \nu + 1\right), \tag{2.9}$$

where $f_{t_1}(.;\nu)$ is the PDF of standard univariate t distribution with ν d.f., and

$$\zeta_i = \frac{\delta_i}{\sqrt{1-\delta_i^2}}. \tag{2.10}$$

The PDF (2.9) can be used to numerically calculate $F_{STD}(.;.)$ and $F_{STD}^{-1}(.;.)$ that are required for the switch to the copula.

Azzalini and Capitanio (2003) derive various properties of this distribution. For our purposes, the closed-form expressions for some moments are useful. Particularly, for $Y \sim MSTD(P, \delta, \nu)$,

$$\mathbb{E}(Y) = \delta(\nu/\pi)^{1/2}\frac{\Gamma\Big((\nu-1)/2\Big)}{\Gamma(\nu/2)},\ \nu > 1, \tag{2.11}$$

$$\mathbb{E}(YY') = P\frac{\nu}{\nu-2},\ \nu > 2, \tag{2.12}$$

and it obviously implies that

$$\mathrm{Cov}(Y) = P\frac{\nu}{\nu-2} - \delta\delta'\frac{\nu}{\pi}\left[\frac{\Gamma\Big((\nu-1)/2\Big)}{\Gamma(\nu/2)}\right]^2,\ \nu > 2. \tag{2.13}$$

2.2.3 Skew-*t* copula

The copula corresponding to MSTD above is the skew-t copula. Yoshiba (2018) gives a good description of the copula and its parameterizations, with a focus on maximum likelihood estimation of its parameters.

The copula function can be expressed via inversion of (2.1), yet it involves the joint and marginal CDFs of MSTD that can only be calculated numerically from (2.7) and (2.9). The same holds for the copula density function. Moreover, this transformation requires the marginal quantile functions of MSTD, $F_{STD}^{-1}(.;.)$ that can only be calculated numerically.

Further, unlike the related symmetric Gaussian or t copulas (the copulas of multivariate normal and multivariate t distributions), the skew-t copula has no closed-form expressions that

would relate the parameters of the copula to the dependency measures (like rank correlations).

The lack of analytical properties and the necessity to numerically integrate and invert the marginal CDFs complicate the application of the copula, even in low dimensions. Nevertheless, the copula remains an attractive tool due to its flexibility.

Figure 8 gives an example of how the skewness of the copula can influence the joint distribution. Both joint densities share the same standard normal marginals, the same values of the degrees of freedom parameter ν, and the off-diagonal element of the matrix parameter P, ρ. The difference in the values of the skewness parameter δ has a significant effect on the shape of the resulting density. In the left panel, the skewness creates a rather minor nonlinear though asymmetric effect, but the random variables remain correlated. However, due to a lot more significant skewness in the right panel, the random variables appear nearly uncorrelated, while there is obvious non-linear asymmetric dependence in means. Moreover, a significant portion of the joint probability mass is relocated in the direction of the left-bottom tail, which lays the basis for asymmetric tail dependence.

2.3 Estimation of the skew-t copula

2.3.1 General notes on copula estimation

There are several approaches to copula estimation. The parameters of the copula can be estimated together with the parameters of the marginal distribution via any relevant method. This approach requires knowledge of the particular properties of the resulting joint distribution and in some sense ignores the advantages of the copula approach. It is sometimes chosen in low-dimensional (mostly bivariate) settings with a relatively low number of parameters and well-established closed-form or computationally practical numerical properties of the distributions.

In more complicated settings, the copula is estimated separately from the marginal distributions. Consider a target joint distribution with CDF G that has the marginals G_j and the copula function C. Given a sample $\{x_{ij}\}_{i=1,\dots n}$ from G, it is transformed to the corresponding sample of so-called pseudo-observations

$$u_{ij} = \hat{G}_j(x_{ij}), \tag{2.14}$$

where $\hat{G}_j$ is an estimate of the univariate marginal CDF G_j. The pseudo-observations (2.14) are then considered as a sample from the distribution with the joint CDF C and are used to estimate its parameters.

One way to approach the estimation is to relate the parameters of the copula to particular properties of the distribution C, and use these relations to estimate the parameters. For example, because the transformation (2.14) is essentially equivalent to the transformation into ranks, for some copulas, rank correlations of the underlying distributions are used to estimate the correlation-like parameters (Gaussian copula, t copula) (Demarta and McNeil, 2005; Anatolyev and Pyrlik, 2022). However, this is not a practical option in the case of the skew-t copula, because no such properties are available in closed form.

For the skew-t copula, Yoshiba (2018) suggests a maximum likelihood estimator for the copula, based on the fast numerical evaluation of the copula density function and a convenient reparameterization of the matrix parameter that guarantees well-conditioned estimates. The approach proves to be practical in low-dimensional settings. However, as the number of dimensions grows, the number of scalar parameters escalates and makes full maximum likelihood optimization unduly heavy.

We suggest a different approach that utilizes the convenient stochastic representation (2.2 - 2.6) of MSTD. Instead of relating the parameters of C to its properties, we use the simulated method of moments to identify the lower-dimensional parameters of the copula δ, ν and use their estimates to perform another inversion transformation to MSTD distribution, for which the analytical properties (2.11 - 2.13) can be used to estimate the remaining high-dimensional matrix parameter. The algorithm is presented in detail in the next few sections.

2.3.2 Suggested two-step algorithm

2.3.2.1 The general scheme

Consider a random variable that generates pseudo-observations from skew-t copula

$$(u_1, ...u_p)' \sim C(P, \delta, \nu). \tag{2.15}$$

This can be considered as a stage of a broader task of estimating a joint distribution of the originally observed random variable x from the distribution G, whose marginals are not of interest or are known. The pseudo-observations (2.15) then should be considered as the result of transformation (2.14) of the original data.

The estimation task is then to obtain estimates $\hat{P}, \hat{\delta}, \hat{\nu}$ based on a sample of the pseudo-observations $\{u_{i1}, ...u_{ip}\}_{i=1,...n}$ (possibly, under $n \geq p$).

The estimation is challenging in two ways. Firstly, the number of scalar parameters grows fast with growing p. Secondly, the estimates of the parameters P, δ must result in a well-conditioned estimate of the corresponding extended matrix Ω (defined by (2.2)).

The rationale behind the suggested two-step estimation technique is the following. Suppose, the true values of the parameters δ, ν are known. Then consider the inverse transformation of the pseudo-observations $\{u_{ij}\}$

$$y_{ij}(\delta_j, \nu) = F_j^{-1}(u_{ij}; \delta_j, \nu), \tag{2.16}$$

where $F_j^{-1}(u_{ij}; \delta_j, \nu)$ is the marginal CDF of MSTD with parameters P, δ, ν. By construction of the skew-t copula, then,

$$Y = (y_1, ...y_p)' \sim MSTD(P, \delta, \nu). \tag{2.17}$$

Using the property (2.13), the remaining matrix parameter P can be estimated using the method of moments:

$$\hat{P}(\delta, \nu) = \frac{\nu - 2}{\nu} \widehat{\mathrm{Cov}}(Y) + \delta\delta' \frac{\nu - 2}{\pi} \left[\frac{\Gamma\Big((\nu - 1)/2\Big)}{\Gamma(\nu/2)} \right]^2. \tag{2.18}$$

Although the parameters δ, ν are not known in actual applications, they can be estimated from lower-dimensional subsets of the data $\{u_{ij}\}$. The stochastic representation of MSTD

(2.2 - 2.6) implies that for any pair $u_{j_1}, u_{j_2}, j_1 \neq j_2$ from (2.15) it holds that

$$(u_{j_1}, u_{j_2})' \sim C\left(\begin{bmatrix} 1 & \rho_{j_1 j_2} \\ \rho_{j_1 j_2} & 1 \end{bmatrix}, \begin{bmatrix} \delta_{j_1} \\ \delta_{j_2} \end{bmatrix}, \nu\right). \tag{2.19}$$

Breaking the sample from (2.15) into such pairs (2.19), even in a high-dimensional setting with relatively large p, we can estimate the parameters $\delta_1, ...\delta_p, \nu$ consistently[12]. Moreover, each parameter appears in different pairs and can be estimated more than once, which makes the final estimates more robust (for example, by taking the median estimator for each of them). The estimates $\hat{\delta}, \hat{\nu}$ then can be used to estimate the quantile functions F_j^{-1} to perform the approximate transformation (2.16) and the estimation (2.18).

Thus, given a sample of pseudo-observations $\{u_{i1}, ...u_{ip}\}_{i=1,...n}$ from a skew-t copula $C(P, \delta, \nu)$, the two-step estimation procedure is the following:

Step 1. Using a sufficient set of pairs $\{j_1, j_2\}, j_1, j_2 \in \{1, ...p\}, j_1 < j_2$, estimate the parameters $\delta_1, ...\delta_p, \nu$ from the subsets of data $\{u_{j_1}, u_{j_2}\}_{i=1,...n}$ using the property (2.19).

Step 2. Use the estimates from *Step 1* $\hat{\delta}, \hat{\nu}$ to estimate the marginal quantile functions of the $MSTD(P, \delta, \nu)$ and perform the transformation of the data (2.16). Use the transformed data to estimate the large matrix parameter P using (2.18).

The two steps are described in detail in the next two sections.

2.3.2.2 SGMM estimation of bivariate skew-t copula

2.3.2.2.1 Population moment conditions

The bivariate skew-t copula (2.19) is a 4-parameter distribution function with very limited analytical properties. However, the stochastic representation (2.2) - (2.6) of MSTD makes simulating it very convenient, including the calculation of mixed moments

[12]Naturally, the parameters elements of the matrix P, $\rho_{j_1 j_2}$ will also be identified and individually consistently estimated. However, the resulting estimates of either the matrix P or the extended matrix Ω are not expected to be either consistent or well-conditioned. Hence, these estimates are disregarded in this step.

$$m_{k_1,k_2}(\rho_{j_1\,j_2},\delta_{j_1},\delta_{j_2},\nu)=\mathbb{E}\left[u_{j_1}^{k_1}u_{j_2}^{k_2}\right]. \tag{2.20}$$

As the domain of the random variable (2.19) is, by definition of the copula, $[0,1]^2$, the moments (2.20) are also bound in $[0,1]$ under $k_1,k_2\geq 0$. The simulation-based calculation of such moments is rather practical.

The same holds for the sample analogs of these moments. Under any $k_1,k_2\geq 0$, the sample moment $n^{-1}\sum_{i=1}^{n}\left[u_{j_1,i}^{k_1}u_{j_2,i}^{k_2}\right]$ is easy to evaluate.

Thus, we define population moment conditions for a simulated generalized method of moments (SGMM) estimation of the bivariate skew-t copula (2.19):

$$\mathbb{E}f_{k_1,k_2}(u_{j_1},u_{j_2};\rho_{j_1\,j_2},\delta_{j_1},\delta_{j_2},\nu)=0,\ \forall(k_1,k_2)\in K \tag{2.21}$$

where

$$f_{k_1,k_2}(u_{j_1},u_{j_2};\rho_{j_1\,j_2},\delta_{j_1},\delta_{j_2},\nu)=u_{j_1}^{k_1}u_{j_2}^{k_2}-m_{k_1,k_2}(\rho_{j_1\,j_2},\delta_{j_1},\delta_{j_2},\nu), \tag{2.22}$$

and K is a set of pairs of indices (k_1,k_2) such that the set of conditions (2.21) is sufficient to identify the four parameters of (2.19).

2.3.2.2.2 Choice of conditions

Set of $|K|=4$ non-redundant and mutually non-repeating conditions results in exact identification and method of moments estimation. $|K|>4$ leads to over-identification and results in a generalized method of moments estimation.

In our work, we present estimation results obtained under

$$K=\{(1/2,1/2),\ (1,1/2),\ (1/2,1),\ (1,1),\ (2,2)\}. \tag{2.23}$$

We made the choice of this set K based on the results of our preliminary simulations. The simulations suggested that, for stronger robustness of the simulation-based calculation of the

moments (2.20) and identification of the parameters from the conditions (2.21), negative k and $k > 3$ are best avoided, as are too close pairs (k_1, k_2).

Inefficient estimator

Given a sufficient set of conditions (2.21), the inefficient SGMM estimator of the copula (2.19) parameters is

$$\left\{\hat{\rho}_{j_1\,j_2}, \hat{\delta}_{j_1}, \hat{\delta}_{j_2}, \hat{\nu}\right\} = \arg \min_{r,d_2,d_2,v} \sum_{(k_1,k_2)\in K} \left(n^{-1} \sum_{i=1}^{n} f_{k_1,k_2}(u_{j_1,i}, u_{j_2,i}; r, d_1, d_2, v) \right)^2. \tag{2.24}$$

The estimator (2.24) is a typical inefficient GMM estimator. Although it is rather computationally practical in the context of a single bivariate copula estimation, the advantages of simulation-based estimation would be lost in any extension to the efficient GMM estimation. Moreover, when it is used to estimate the parameters of bivariate blocks of a higher-dimensional MSTC, the procedure needs to be repeated multiple times, which makes the estimation of parameters of the multivariate copula significantly computationally heavier even via the inefficient estimator. Last but not least, proceeding to a well-conditioned estimation of the parameters of the multivariate copula requires consistent estimates of the lower-dimensional blocks, yet those estimates will be used to transform the sample and will not appear directly in the final estimator. Thus, in *Step 1* of the two-step estimation algorithm of MSTC, the inefficient SGMM estimator of bivariate skew-t copula proves to be sufficient.

2.3.2.2.3 Application to STC estimation

An SGMM estimation of bivariate skew-t copula is used in *Step 1* of our suggested two-step estimation algorithm of MSTC. The step is performed in the following way. We observe a sample $\left\{\left(u_{1,i}, ...u_{p,i}\right)'\right\}_{i=1,...n}$ from a p-dimensional skew-t copula (2.15). The goal of *Step 1* is to estimate the copula's lower-dimensional parameters $\delta = (\delta_1, ...\delta_p)'$ and ν so that the transformation (2.16) is available. These parameters can be identified from pairs $(j_1, j_2) \in J$, the set of which is defined as

$$J = \{(j_1, j_2) \mid j_1 < j_2,\ j_1, j_2 \in \{1, ...p\}\},$$

and

$$\forall j \in \{1, ...p\}\ \exists j' :\ \text{either } (j, j') \in J,\ \text{or } (j', j) \in J.$$

The set J does not require inclusion of all the pairs of indices, as the parameters $\rho_{j_1 j_2}$ are not required to perform the transformation (2.16). However, each $j \in \{1, ...p\}$ should appear in at least one of the pairs in J.

Thus, $\forall (j_1, j_2) \in J$ SGMM above will provide estimates $\hat{\delta}_{j_1}(j_1, j_2)$, $\hat{\delta}_{j_2}(j_1, j_2)$, and $\hat{\nu}(j_1, j_2)$. The resulting estimates of *Step 1* of the two-step algorithm are defined as

$$\hat{\delta}_j^I = \text{median}\left\{\{\hat{\delta}_j(j, j') | (j, j') \in J\} \cup \{\hat{\delta}_j(j', j) | (j', j) \in J\}\right\},\ j = 1, ...p, \tag{2.25}$$

$$\hat{\nu}^I = \text{median}\left\{\hat{\nu}(j, j') \mid (j, j') \in J\right\}. \tag{2.26}$$

The estimate $\hat{\nu}^I$ is also the final estimator of the degrees of freedom parameter that remains unchanged on *Step 2*. The estimates $\hat{\delta}_j^I$, however, are subject to change in *Step 2*, as a part of the large matrix parameter Ω.

2.3.2.3 Shrinkage-based estimation of the large matrix parameter

2.3.2.3.4 First approximate of the extended matrix parameter

The estimates of *Step 1* (2.25) and (2.26) are next used to transform the sample of pseudo-observations $\{u_{ij}\}$ into $y_{ij}(\hat{\delta}_j^I, \hat{\nu}^I)$ via the inverse transformation (2.16). The resulting sample $\{y_{ij}\}$ is then treated as coming from the multivariate skew-t distribution $MSTD(P, \delta, \nu)$, and its parameters are estimated using the analytical properties of the distribution.

We begin by using (2.18) to obtain the first approximation of the large matrix parameter P as $\hat{P}(\hat{\delta}^I, \hat{\nu}^I)$. The quality of the estimate depends on the estimator $\widehat{\text{cov}}Y$ used in (2.18)

and the estimates $\hat{\delta}^I$ and $\hat{\nu}^I$. The main challenge of the estimation is that not only $\hat{P}$ should be a well-conditioned estimate of a correlation matrix, so should be also the corresponding estimate of the extended matrix (2.2). The corresponding to the first approximation of $\hat{P}$ estimate of Ω, thus, is

$$\hat{\Omega}^I(\hat{\delta}^I, \hat{\nu}^I) = \begin{pmatrix} 1 & \hat{\delta}^{I\prime} \\ \hat{\delta}^I & \hat{P}(\hat{\delta}^I, \hat{\nu}^I) \end{pmatrix}. \tag{2.27}$$

Although the estimator (2.27) is by-element consistent to its population analog (2.2), it is not guaranteed to be well-conditioned. Not having a well-conditioned estimate for the matrix Ω crucially limits the application of the estimated skew-t copula. To improve the properties of the estimator (2.27), we suggest using the analytical large covariance matrix shrinkage of Ledoit et al. (2020).

2.3.2.3.5 Analytical shrinkage of large covariance matrix

A detailed description of the analytical shrinkage estimator, its properties, and its relative performance compared to other available covariance matrix shrinkage methods is provided in Ledoit et al. (2020).

Unlike previous numerical versions of non-linear shrinkage methods (Ledoit and Wolf, 2012, 2017a), the analytical approach is significantly faster, and preserves accuracy. It not only resolves the challenging scenario in which the matrix dimension exceeds the sample size, but it also extends the method's applicability to high-dimensional covariance matrices of up to thousands of variables (Ledoit et al., 2020; Ledoit and Wolf, 2022).

Another advantage of the analytical non-linear shrinkage, which is particularly important in the context of skew-t copula estimation, is that the estimator is directly applied to the "non-shrunken" matrix, and it does not require observation of the underlying data. Thus, we directly apply the analytical non-linear shrinkage to the estimator (2.27).

If we rewrite the approach of Ledoit et al. (2020) using the notation of this paper, the analytical non-linear shrinkage estimation of the extended matrix Ω based on the initial, ill-conditioned estimate $\hat{\Omega}$ goes as follows.

1. Perform the spectral decomposition of the matrix $\hat{\Omega}$:

$$\hat{\Omega} = V\Lambda V', \tag{2.28}$$

where

$$\Lambda = \text{diag}\{\lambda_1, ...\lambda_{p+1}\} \tag{2.29}$$

is the diagonal matrix with $\hat{\Omega}$'s eigenvalues. Without loss of generality, assume $\lambda_1 \leq ...\lambda_{p+1}$. The matrix V is the corresponding matrix of $\hat{\Omega}$'s eigenvectors.

2. Set the "global" bandwidth for the estimator

$$h_n = n^{-1/3}, \tag{2.30}$$

and the "locally adaptive" bandwidths

$$h_{n,j} = \lambda_j h_n, \ j = 1, ...p+1. \tag{2.31}$$

3. Estimate the spectral density using Epanechnikov kernel:

$$\tilde{f}_n(\lambda_i) = \frac{1}{p+1}\sum_{j=1}^{p+1}\frac{3}{4\sqrt{5}h_{n,j}}\left(1 - \frac{1}{5}\left(\frac{\lambda_i - \lambda_j}{h_{n,j}}\right)^2\right)^+, \tag{2.32}$$

and its corresponding Hilbert transform:

$$\mathcal{H}_{\tilde{f}_n}(\lambda_i) = \frac{1}{p+1}\sum_{j=1}^{p+1}\left[-\frac{3(\lambda_i - \lambda_j)}{10\pi h_{n,j}^2} + \frac{3}{4\sqrt{5}h_{n,j}}\left(1 - \frac{1}{5}\left(\frac{\lambda_i - \lambda_j}{h_{n,j}}\right)^2\right) \cdot \log\left|\frac{\sqrt{5}h_{n,j} - \lambda_i + \lambda_j}{\sqrt{5}h_{n,j} + \lambda_i - \lambda_j}\right|\right]. \tag{2.33}$$

4. Calculate the estimated "shrunk" eigenvalues:

$$\tilde{\lambda}_j = \frac{\lambda_j}{\left(\pi\frac{p+1}{n}\lambda_i\tilde{f}_n(\lambda_j)\right)^2 + \left(1 - \frac{p+1}{n} - \pi\frac{p+1}{n}\lambda_i\mathcal{H}_{\tilde{f}_n}(\lambda_j)\right)^2}, j = 1, ...p+1. \tag{2.34}$$

5. The resulting estimator of the large matrix parameter is

$$\hat{\Omega}^{ASh} = V \text{diag}\{\tilde{\lambda}_1, ...\tilde{\lambda}_{p+1}\}V'. \tag{2.35}$$

2.3.2.3.6 Remarks on using the analytical shrinkage to estimate skew-t copula

Our proposed algorithm involves application of the method of moments in *Step 1* and subsequent large covariance matrix shrinkage in *Step 2*. Both methods presume the independence of the observations in the sample. However, in the context of copula estimation, this assumption is not met, as both the pseudo-observations (2.14) and the inverted observations (2.16) used in the second step are inherently dependent across observations. It is essential to acknowledge that this issue is common to copula estimation tasks. Thus, we maintain confidence in the finite sample performance of our estimator as in the cases of the other copula estimators.

The utilization of the large covariance matrix shrinkage estimator in *Step 2* of the algorithm should be regarded as a practical means to obtain a well-conditioned estimate of the extended correlation matrix parameter of the copula. This should not be misconstrued as an optimal shrinkage estimator, as it is not applied to a sample correlation matrix corresponding to a traditional data sample. However, it is important to note that the extended matrix subject to shrinkage is, in fact, a correlation matrix. Its high-dimensional nature renders it similarly ill-conditioned to a typical high-dimensional sample correlation matrix. Consequently, we anticipate that the finite sample properties of the estimator will remain intact, though a thorough evaluation of its performance awaits future research.

The efficacy of our proposed two-step estimator is contingent upon the selection of technical details in its application, including the choice of sets K and J. In this paper, we intentionally opt for simpler and smaller sets to prioritize computational efficiency over precision. Nonetheless, our objective is to underscore the practicality of the estimator, recognizing that more comprehensive evaluations can be pursued in subsequent research.

2.4 Dynamic portfolio allocation example

2.4.1 Background and data

As an empirical example, we perform a dynamic portfolio allocation exercise in the European stock market in the first half of 2022. The choice of this particular period is intentional. The baseline market dynamics for the selected period are negative, with the representative market index EUROSTOXX50 losing between 10% and 20% of its value at the beginning of the year by June; see Figure 9. However, we aim to show that using skew-t copula to account for tail-dependency across market assets' returns results in portfolios that hedge successfully against significant shocks and outperform the market.

From Yahoo Finance (2022), we retrieve daily prices of the top 30 components of the EUROSTOXX50 index for the period from November 30, 2021, till June 1, 2022. The period is divided into rolling samples of 2 months (roughly 40 observations). We use the first half of the sample to evaluate the joint distribution of the assets' returns and construct the optimal portfolio for the second half of the period. Then, we evaluate the performance of the portfolio out-of-sample by rolling the sample by 1 month forward. We repeat the evaluation and update the portfolio structure using the data most recent to the update date. Thus, during the whole analysis period, we update the portfolio 4 times. Throughout the period, the benchmarks to compare the performance of the portfolio are the market index EUROSTOXX50 and a naïve portfolio of the index's top 30 components that includes them with equal weights.

Prior to presenting and discussing the results of this exercise, a few remarks are necessary. It is important to note several deliberate choices made in the design of the empirical exercise. Firstly, our use of only 20 historical data points for each portfolio structure update may seem artificially small. However, this intentional decision serves the purpose of highlighting the application of our suggested algorithm of high-dimensional skew-t copula estimation. Should one contemplate extending the time period for portfolio evaluation, it would be crucial to recognize that the number of assets in the portfolio may also increase, maintaining the relationship "$p > n$".

Similarly, our selection of a one-month (20 observations) period for evaluating the portfolio

before the update is somewhat arbitrary. This choice aligns with the length of the evaluation period, which may not be essential in real-life applications. Importantly, in this exercise, the choices of sample size and the number of assets in the portfolio serve better interpretability and computational feasibility.

Thirdly, it is essential to clarify that our primary objective is not to compare the performance of this portfolio construction approach to other allocation techniques. While our portfolio significantly outperforms the market at the period's end, this outcome is influenced, in part, by the overall market conditions during the chosen period. It remains plausible that alternative portfolio allocation techniques could also propose sound improvements beyond a naïve investment strategy. However, our findings highlight the significant contribution of the skew-t copula in capturing tail dependence among asset returns and translating it into a portfolio structure.

Therefore, this empirical exercise should be regarded as 'toy' example, demonstrating the advantages of the high-dimensional copula estimation algorithm and providing intuition for its application in portfolio allocation. We deliberately leave the extension of this technique to a comprehensive portfolio allocation study for future research.

The portfolio evaluation technique we apply is close to the one used by Anatolyev and Pyrlik (2022). The main difference is that, in this study, we use only one model of the joint distribution of the assets' returns after filtration, which is the skew-t copula. A short recap of the portfolio evaluation technique adjusted for this research is presented in Appendix D.

2.4.2 Results

Figure 10 illustrates the comparative dynamics of accumulated returns for the skew-t copula-based investment strategy, a naïve portfolio, and the market over the specified time period. Notably, the model-driven strategy consistently outperforms the market, with a particularly sound result at the end of the period: the model-based portfolio achieved positive accumulated returns, while the market's dynamics, on average, exhibited a significant negative trend.

Upon closer examination of the portfolio's performance over distinct periods, occasional co-movement with the market was observed. However, the model-based portfolio demonstrated

more frequent instances of positive dynamics compared to the negative returns observed in the overall market. Additionally, certain negative shocks affected all portfolios, given their impact on the broader market, making them challenging to hedge through conventional strategies.

Further insights into the portfolio's dynamics are provided by examining specific examples of its structural changes based on copula estimates and their alignment with performance. For example, in Period 1, Sanofi (SAN.PA) and Carrefour SA (CA.PA) stocks constituted up to 30% of the portfolio, then decreased to 2% in Period 2. This shift corresponds to the copula density function's estimate for these assets' returns in Period 1, indicating asymmetric tail dependence (Figure 11). Similarly, the share of Inditex (ITX.MC) and Anheuser-Busch InBev (ABI.BR) decreased from 35% in Period 2 to 9% in Period 3 (Figure 12). The share of Eni (ENI.MI) and Assicurazioni Generali (G.MI) decreased from 24% in Period 3 to 7% in Period 4 (skew-t copula density in Figure 13). These adjustments underscore the skew-t copula's pivotal role in accommodating tail dependence among various asset pairs within the portfolio.

2.5 Discussion and concluding remarks

In this paper, we introduce a novel approach to high-dimensional skew-t copula estimation, surpassing the copula size studied in the existing literature. Our estimation algorithm is robust even when the number of observations is comparable to or smaller than the copula size. Given the copula's lack of analytical properties, we propose a two-step algorithm. In Step 1, we leverage the stochastic representation of the multivariate skew-t distribution to estimate lower-dimensional copula parameters using a simulated generalized method of moments. Although the method of moments is inefficient, it provides consistent estimates of degrees of freedom and asymmetry parameters, which we use to transform the data into inverted observations treated as a sample from the multivariate skew-t distribution. In Step 2, we use these transformed data for high-dimensional copula parameter estimation. To ensure a well-conditioned estimate of the matrix, we employ analytical non-linear shrinkage of large covariance matrix estimation.

Illustrating the application of the skew-t copula and the high-dimensional estimation algorithm, we conduct a dynamic portfolio allocation exercise in the European stock market for the first half of 2022. Our model-based portfolio exhibits positive accumulated returns at the period's end, in stark contrast to the market's overall negative trend. A closer examination of the portfolio's structure emphasizes the skew-t copula's crucial role in accommodating tail dependence among various asset pairs.

Our proposed estimation algorithm involves the application of the method of moments and large covariance matrix shrinkage, assumptions of which are not fully met in copula estimation. The algorithm is not guaranteed to be either efficient or optimal due to the inter-dependence of pseudo-observations and inverted observations. The efficacy of the two-step estimator hinges on technical details, such as the choice of moment conditions. We deliberately opt for simpler choices to prioritize computational efficiency, underscoring the practicality of the estimator. However, a comprehensive evaluation awaits future research, addressing finite sample performance and properties.

Additionally, our suggested application of the large covariance matrix shrinkage estimator to achieve well-conditioned estimates of the extended correlation matrix parameter should not be regarded as the optimal shrinkage estimator. While we anticipate that the finite sample properties will remain intact, a thorough evaluation awaits future research. For potentially improved finite sample performance, we encourage deviations from the original parameters of the analytical shrinkage formula. Asymptotic optimality is not the primary goal, emphasizing the need for a nuanced exploration of parameters such as global and local adaptive bandwidths or the spectral density estimator's kernel in future research.

3 Forecasting Realized Volatility Using Machine Learning and Mixed-Frequency Data (the Case of the Russian Stock Market in 2016-2020)

3.1 Introduction

In this chapter, we step away from unconditional joint distributions construction and assess gains of high-dimensional and mixed-frequency settings in realized volatility forecasting. The results of this research were published in CERGE-EI Working Paper Series (Pyrlik et al., 2021).

Stock market volatility is known to be a measure of the dispersion of stock returns, and it is commonly used to assess the riskiness of an asset. For the majority of asset and investment risk management problems, volatility is one of the most important and irreplaceable characteristics of assets. Hence, forecasting stock returns volatility has been popular among financial market researchers.

Measuring volatility of returns is performed in various ways, with the *realized volatility* (RV) approach popular among practitioners and researchers. It has proven to be a preferred volatility evaluation technique due to its natural use of more information from high-frequent market data. Andersen and Bollerslev (1997) were the first to show that realized volatility forecasts outperform the predictions of numerous alternative approaches.

The concept of realized volatility was first introduced by Andersen and Bollerslev (1998) and is defined as the sum of squared intraday stock returns. An important advantage of the RV approach is that this volatility measure is observed directly from data, unlike the definitions of market volatility based on latent volatility models such as stochastic volatility (SV) or generalized autoregressive conditionally heteroskedastic models (GARCH).

Modeling and forecasting realized volatility can be done via several approaches, such as the heterogeneous autoregressive realized volatility (HAR-RV) model or multiplicative error model (MEM). Alternative approaches to modeling volatility are often compared. Andersen

et al. (2003) show that HAR-RV is superior to SV models and GARCH. Hence, we use HAR-RV as the benchmark in our research.

In the literature, it is common to apply various modifications to the models to improve the quality of the forecasts. The method most commonly used and one that has proven relatively reliable is inclusion of exogenous variables that contain valuable information about the dynamics of volatility, the asset at hand or even the market or the economy in general. Based on the literature, the most commonly used variables to explain the dynamics of volatility and to improve the predictive power of the models can be divided into the following groups.

Financial market variables are one of the most extensive and prominent sets of indicators in the literature. Lagged equity market returns are often shown to predict volatility. For example, a well known stylized fact on most markets is that, if market returns are negative, volatility increases (Christiansen et al., 2012; Nonejad, 2017). The earning-price ratio is an important indicator of a firm's well-being and value, so changes in the ratio potentially predict the stock returns volatility. When the earning-price ratio decreases, it is likely to indicate poor current and future performance of the firm, and hence a higher level of stock returns volatility in the future (Christiansen et al., 2012; Nonejad, 2017). Similarly, the dividend-price ratio can capture changes in stock returns volatility through the channel of investment productivity. When this ratio decreases, stock returns volatility is expected to increase (Christiansen et al., 2012; Nonejad, 2017). Long-term bond returns are considered to carry higher risks than short-term ones, and thus have higher interest rates. Variations of these quantities can be used to proxy investor attitude towards risk (Nonejad, 2017; Audrino et al., 2020).

Market liquidity is another important indicator that provides information about stocks returns and their volatility. An increase in liquidity is expected to indicate an increase in the level of market participants activity in the market. A significant change in activity typically leads to changes in price levels, returns, and volatility. As liquidity is not directly observable, a variety of indicators have been developed to capture it (such as Amihud, Roll, and High-Low). According to Będowska-Sójka and Kliber (2019), there is a significant relationship between volatility and liquidity, but the sign of the correlation can differ depending on the liquidity

proxy. The authors conclude that the most relevant approximation of liquidity is High-Low, as this measure unilaterally influences volatility. Xu et al. (2019) show that there is a non-linear dependence between liquidity and volatility with persistent influence of the former on the latter. This research also exhibits that High-Low liquidity proxies are the most influential in realized volatility forecasting.

Further, the number of daily transactions may carry significant information for movements of stock volatility. Some studies confirm this effect, while others state that it does not exist. For example, Shahzad et al. (2014) show that the number of trades in a day is a more significant predictor of volatility than average daily volume. Moreover, they demonstrate that the number of individual trades is a more important predictor than the number of trades by institutional market participants. A possible explanation is that individuals' actions represent a noise term (because they possess less reliable information about the market than organizations), which, in certain time periods, can lead to abnormally high volatility. Wang et al. (2015) also confirm the existence of the trading volume effect and point out that, the longer the forecasting horizon is, the lower the influence is. As for a contrary view of this effect, Todorova and Souček (2014) show that, for the German market, the trading volume of stock does not include any significant information for explaining realized volatility. It is worth noting that this result was achieved both in-sample and out-of-sample.

Stock volatility is also known to be significantly time-dependent. Hence, incorporation of day-of-the-week, weekend, and holiday effects is of great importance for precise forecasting of stock volatility. Many authors focus specifically on the effects of non-trading days. Martens et al. (2009) claim that stock volatility is usually higher after holidays, and on Christmas, half of its regular level. As shown by Wang and Hsiao (2010), weekday holidays increase the volatility of the S&P 100 and FTSE 100, while half-day trading periods decrease it. Diaz-Mendoza and Pardo (2020) find that volatility significantly decreases on the first day after a holiday or weekend, but after a long holiday, volatility either rises or remains the same.

Similarly, overnight and lunch-break periods are relevant for forecasting returns volatility, because during these periods important information on trade or macroeconomic news may

arrive. According to Wang et al. (2015), these non-trading periods significantly influence volatility. Moreover, they state that the leverage effect is captured, as the volatility rises higher after negative shocks to returns. The same results are achieved by Todorova and Souček (2014) and Zhu et al. (2017), who claim that the effects of overnight returns are higher than those of lunch-break returns.

In addition to the calendar effects, expiration-day effects of related derivatives have been thoroughly investigated. These effects measure how futures or option contract trading close to an expiration day may influence the underlying stock returns and volatility. This has been studied for stock markets in various countries, and the results are drastically different. Bollen and Whaley (1999) use Chinese data and do not discover statistically significant difference in stock volatility on expiration and non-expiration days of the derivatives. A similar result is achieved by Xu (2014) using Swedish data. However, Arago and Fernandez (2002) conclude that, for the Spanish market, volatility is significantly higher during a week with an expiration day. Chou et al. (2006) arrive at the same conclusion in the case of the Taiwanese market.

The inclusion of a variety of macroeconomic indicators is justified, because the overall economic environment influences the well-being of the corporate sector and thus the volatility in the market. The most frequently used proxies are CPI, industrial production growth, and GDP growth (Wongbangpo and Sharma, 2002; Christiansen et al., 2012; Paye, 2012; Nonejad, 2017; Audrino et al., 2020; Fang et al., 2020; Thampanya et al., 2020). It is important to note that, when macroeconomic variables are used in combination with financial indicators, most of the time, the former appear to be insignificant. Housing starts is one of the few indicators that has proven to influence volatility (Audrino et al., 2020; Fang et al., 2020). A possible mechanism behind this effect is that the more new houses are built, the more the credit market expands Fang et al. (2020). T-bill rates are often used as predictors of market volatility. If the economy is unstable, then T-bill rates generally tend to decrease, while volatility commonly increases. These variables are used to proxy the steadiness of the current economic situation (Christiansen et al., 2012; Nonejad, 2017; Audrino et al., 2020).

Not only does the domestic market affect stock volatility, but so do spillover effects from adjacent or global financial markets. These spillover effects represent the influence of foreign

or adjacent markets on a local market. Balli et al. (2015) show that one of the important representations of the spillover effect is the trading volume of goods between developed and emerging markets. They also demonstrate that spillovers from the US are higher than those from Europe or Japan. Martens et al. (2009) illustrate that RV is higher on news announcement dates. Similarly, Wang and Hsiao (2010) demonstrate that for the Taiwanese market, weekend days raise the volatility of stocks, because, typically, a considerable amount of macroeconomic news is issued on Fridays in the US.

Further, the oil market appears to be closely connected to stock prices, which is a manifestation of spillover between adjacent markets. According to Kang et al. (2015) negative shocks in oil production lead to positive shocks in stock returns volatility. Similarly, an increase in demand for oil translates into a decrease in volatility. Luo and Qin (2017) state that oil price shocks positively influence returns on the Chinese stock market, as a rise in the oil price is a sign of an upturn in the economy.

Changing the functional form of the regression is another approach that is often used to improve both the explanatory power, and the predictive performance of the models. On the one hand, models like HAR-RV or MEM are commonly said to exhibit a relevant level of interpretability. On the other hand, they are not guaranteed to deliver reasonable forecasting power for either short or long time horizons, due to their limited ability to capture effects that are more complicated than the linear correlations between the volatility dynamics and the explanatory variables. However, machine learning (ML) algorithms are specifically known for highly accurate predictions, due to their ability to capture various non-linear patterns in the relationships between the variables. Recently, much attention has been focused on investigating the applicability of ML in forecasting returns and their volatility.

Ingle and Deshmukh (2021) implement several types of models to predict closing prices of stocks: Generalized Linear Model (GLM), Gradient Boosting Model (GBM), and several types of neural networks in combination with machine learning methods. The results show that GLM displays the highest level of forecasting accuracy, followed by ensemble models and deep learning networks.

Hamid and Iqbal (2004) use a three-layered neural network to forecast the volatility of S&P

500 futures and show that the predictions significantly outperform the benchmark. Further, Parisi et al. (2008) research changes in the market price of gold and find that the best performance, in-sample and out-of-sample, is delivered by a rolling neural network. Ding et al. (2015) investigate potential improvements in predictions for S&P 500, and show that forecasts from a deep convolutional neural network appear 6% better than those from the baseline model.

Long Short-Term Memory (LSTM) is another machine learning method that has been gaining popularity among researchers and practitioners. Xiong et al. (2015) compare the S&P 500 returns volatility forecasts from GARCH, Lasso regression, Ridge regression, and LSTM. The results show that the LSTM forecasts significantly outperform its competitors. Another notable study is Liu (2019), which shows that a combination of recurrent neural networks with LSTM significantly outperforms GARCH in forecasting the returns volatility of S&P 500 and AAPL.

Combining multiple models into one appears to be an effective and, hence, popular approach to applying deep learning algorithms. For example, Kristjanpoller and Minutolo (2015) use artificial neural networks to combine GARCH-based forecasts of the gold price returns variance, and achieve a sound reduction in the mean average percentage error. Vidal and Kristjanpoller (2020) combine LSTM and convolutional neural networks for gold price returns volatility forecasting, achieving a significantly better predictions than GARCH or LSTM alone.

An important feature of the current literature on volatility forecasting, using either traditional approaches or ML, is the scope of the markets under analysis. Overall, most research focuses on the US, Chinese and European markets. Few studies consider emerging markets, and even fewer consider Russia (Aganin et al., 2017; Nagapetyan et al., 2019; Fantazzini and Shangina, 2019; Bazhenov and Fantazzini, 2019; Aganin, 2020). To the best of our knowledge, no research on the Russian stock market has gone beyond GARCH-type or HAR-RV-type of methodology in the analysis of returns volatility. Hence, our main goal is to contribute to the existing literature by performing a comparative analysis of several approaches to forecasting RV in the context of the Russian stock market, including the HAR-RV and ML approaches.

We first aim to identify the extent to which ML is more suitable for RV forecasting than the benchmark HAR-RV on the Russian stock market. Secondly, we seek to learn, what information is significant for the Russian stock market RV forecasting, and how it is different from the situation on international markets. To achieve these goals, we extract an extensive dataset for the Russian stock market and compare the out-of-sample performance of the HAR-RV and 4 ML algorithms (Lasso, Random Forest, Gradient Boosting, and Long Short-Term Memory) in returns RV forecasting for selected top stocks in the market. We find that both the HAR-RV and ML approaches provide us with reasonable predictive power in terms of RMSE of RV in a rolling forecasting scheme, with the ML generally outperforming the benchmark when a reasonable set of exogenous features are included. In particular, Lasso regression appears to deliver a convenient combination of easy implementation and forecasts precision. More complicated algorithms (Random Forest, Gradient Boosting, Long Short-Term Memory) are very promising, but we show that, to benefit from them, they require fine-tuning and frequent re-training, which is a computationally demanding task.

The rest of the chapter is organized as follows. Section 3.2 introduces the methodology of the benchmark HAR-RV model and ML algorithms used in this research, and the data splitting and forecasting schemes we choose. In Section 3.3, we describe the dataset and proceed to our exploratory data analysis in Section 3.3.2. In Section 3.4, we describe the modeling technique and analyze the results in Section 3.5. Section 3.6 offers a discussion of our main findings and limitations of the research, and Section 3.7 concludes. Supplementary aids on the data and results are collected in an online appendix.

3.2 Methodology

3.2.1 The benchmark model

The benchmark model in our study is HAR-RV of Corsi (2009). The main idea of the approach is to use high-frequency data to obtain more accurate forecasts of volatility based

on daily, weekly, and monthly RV. The notation of the model is:

$$RV_{t+1d}^{(d)} = \alpha + \beta^{(d)} \cdot RV_t^{(d)} + \beta^{(w)} \cdot RV_t^{(w)} + \beta^{(m)} \cdot RV_t^{(m)} + \omega_{t+1d}, \tag{3.1}$$

where weekly realized volatility, for example, is given by

$$RV_t^{(w)} = \frac{1}{5} \cdot (RV_t^{(d)} + RV_{t-1d}^{(d)} + ... + RV_{t-4d}^{(d)}). \tag{3.2}$$

Inclusion of additional regressors to the model is straightforward:

$$RV_{t+1d}^{(d)} = \alpha + \beta^{(d)} \cdot RV_t^{(d)} + \beta^{(w)} \cdot RV_t^{(w)} + \beta^{(m)} \cdot RV_t^{(m)} + \sum_i \beta_i \cdot x_{it} + \omega_{t+1d}, \tag{3.3}$$

where x_{it} is an additional explanatory variable i at moment t.

Estimation of the model is typically performed via OLS. Newey-West robust standard errors are used to retain consistency of estimates with heteroskedicity and autocorrelation of the error term. When the extended version (3.3) of the model is used, a model selection technique is required for an in-sample based selection of the optimal combination of explanatory variables. We estimate several specifications of the model with differing additional explanatory variables and select the one that minimizes AIC from those without significant residual autocorrelation.

3.2.2 Machine learning algorithms

Here, we briefly describe the ML algorithms we use and compare to the benchmark HAR-RV. These are: Lasso regression (Lasso), Gradient Boosting (GB), Random Forest (RF), and Long Short-Term Memory (LSTM). Below, we cover distinct features of each algorithm, with a fuller description and explanations of the mechanisms given in Appendix E, page 129.

The main feature of Lasso is regularization on weights of linear regression with zeroing of extreme-value coefficients. This algorithm is typically good at dealing with overfitting that may occur as a result of either a relatively small sample size or too many collinear regressors. Lasso uses the only hyperparameter, which is the penalty for the degree of collinearity. Hence, training this algorithm is fast.

GB is an ensemble algorithm with a consequent learning of regression trees, while RF is another ensemble algorithm that uses parallel learning of regression trees. Due to the consequent structure, GB is capable of accurate capturing of dependencies in the data, but is prone to overfitting. RF, on the other hand, is more robust to overfitting. However, both algorithms are considered well suited for feature selection and coping with multicollinearity. When some features appear highly collinear, the trees will avoid using them together for the sake of greater information gain. These algorithms classify some variables as the most/least significant, depending on their inputs to information gain.

As for neural networks, LSTM is a type of RNN that works with sequences of variables. Due to its recurrent structure, the algorithm can capture autoregressive dependencies, which makes it particularly useful in the tasks of time series forecasting. In contrast with other networks, LSTM is designed to work better with longer sequences of data. The architecture of LSTM is tunable, which makes the algorithm flexible for different data types and tasks. The quality of this algorithm also depends on the learning approach. Hence, such hyperparameters as batch size, learning rate, number of epochs, and type of the optimizer should also be tuned. LSTM, thus, is the most complicated and computationally challenging algorithm among those used in our study.

3.2.3 Partitioning the data and training the models

To train and evaluate our models, we perform a rolling scheme with out-of-sample validation and testing. We sequentially divide our dataset into training, validation, and test sub-samples. Each training sample includes information over a two-year period, and the validation and test samples are the following two quarters (one each). We roll forward by one quarter at a time. We use RMSE to measure the quality of our forecasts.

The validation parts are used to select the hyperparameters of the models (when there are any). The choice of the hyperparameters is made via grid search over excessive sets of possible values of the parameters. The optimal combinations are chosen to minimize the validation RMSE.

3.3 Data

3.3.1 Groups of variables

We extract historical data on the stock prices of the top 9 companies of Moscow Exchange index from the 1st of January 2016 to the 31st of December 2020 at 5-minute frequency[13]. We use the data to calculate daily, weekly and monthly stock realized volatility for each company.

Following the results of previous studies, we add a variety of explanatory variables to our dataset.

- To give an alternative for the T-bill rate, daily values of the Russian Government Bond Index (RGBI) are included. Because returns on market portfolios cannot be reported directly, daily log-returns on the stock market index RTSI were taken as a proxy. An important characteristic for this proxy is high level of diversification; RTSI is a composite index with the most liquid Russian stocks[14].

- To control for changes in the economic environment and macroeconomic circumstances, we added the dynamics of GDP (quarterly), CPI (monthly), and dwellings commenced (monthly)[15]. From the same source, we obtained a few financial performance indicators, specifically, the dynamics of dividend price ratio and earning price ratio for each of the 9 companies (monthly). It is important to note that for POLYUS (www.polyus.com), the major part of these variables do not appear to be available; hence, we omit them from the specifications for POLYUS.

- We included exports and imports to/from the USA from/to Russia, using the data from the census.gov WebSite. In the literature, the exports and imports to/from the USA are classified as the spillover effect. However, in our research, we include them into the group of macroeconomic indicators. This is due to the frequency of these variables,

[13]The data is open and can be obtained directly from the stock exchange website or another service; we obtain the data using the stock prices historical data exports feature of FINAM, www.finam.ru

[14]The data on RGBI and RTSI are available at www.finam.ru

[15]These data can be obtained from the Refinitiv Eikon (Thompson Reuters) https://eikon.thomsonreuters.com

and the fact that the mechanism of the spillovers is typically explained in terms of macroeconomic theory, for example, Balli et al. (2015).

- We calculate several market liquidity indicators: High-Low, Amihud, and Roll, following the approach of Będowska-Sójka and Kliber (2019). However, we end up including only the High-Low metrics into our models, since it showed the most effect on stock volatility in the literature.

- We account for the holiday effect, the weekend effect, and the Friday effect by including the respective dummy variables. To try to capture an eponymous effect, we add the overnight returns to the sets of variables, following the approach of Wang et al. (2015).

- Finally, we included the realized volatility of S&P 500 and Brent oil price to reflect spillover effects from the global stock and crude oil markets.

It is worth noting that the companies in our study represent various sectors of the economy: banking, mining, retail, and oil and gas. Literature shows that spillovers between sectors are of great importance. As presented by Hammoudeh et al. (2009), three main sectors (industrial, service, and banking) of GCC economies demonstrate volatility spillovers. Chen et al. (2019) confirmed results of the previous paper and showed that consumer discretionary, industrial, and health sectors generate the largest spillovers. The US stock market also features cross-volatility between sectors, as shown by Mensi et al. (2020). They demonstrated that consumer services and goods sectors produces the largest amount of volatility, while material sectors produce the least.

Due to the industrial specificity of the Russian economy, it happens that most companies chosen for our research belong to the oil and gas sector. Hence, we do not expect to see much evidence of volatility spillovers between sectors. However, this is a field for future research.

We now describe the specificity of the data, necessary transformations of the variables, and creation of additional indicators, when required. To avoid negative forecasts of realized volatility, we apply the natural logarithm to the dependent variable and its lagged values. We shift to the growth rates of the low-frequent variables to introduce more variation in our data and achieve stationarity. We include the lagged series of the main variables into our

datasets. As a result, we divide all our variables into 5 groups (see Table 1) to investigate additional predictive power that each group of variables brings to a certain model.

Table 1: Groups of explanatory variables included into models

Group	**Variables**
Basic	log RV, log RV weekly, log RV monthly
Overnight and calendar effects	is after the weekend, is after a holiday, is Friday, overnight returns
Financial effects	growth rate of dividend price ratio†, growth rate of earning price ratio†, High-Low, log-returns of RTSI, RGBI
Spillovers	log RV of S&P, log RV of Brent
Macro indicators	growth rate of import/exports from/to the USA†, growth rate of CPI†, growth rate of housing starts†, growth rate of GDP†

† - low frequency variable

We then construct 5 specifications of all implemented models with the consecutive addition of these groups of variables and 5 specifications with lags of variables. We have some missing values in the data. To keep the datasets as complete as possible, for each particular company, we omit variables that are missing at rates 30% or more in the training samples. Less frequent missing values are replaced with the latest known values of the same variable.

3.3.2 Exploratory data analysis

We conduct a preliminary data analysis to identify general patterns within the data, to detect possible effects of extreme events, and to evaluate the overall relationships between the variables.

Firstly, we consider the dynamics of the logarithm of realized volatility to investigate changes throughout the period; see Figure 1 for an example (the rest of the figures are in Appendix). Overall, our dependent variable is a typical time series of this kind: a volatile and possibly heteroskedastic series, yet most likely with a stable longer-run average level and range. Visually, there are specific differences across companies and some common patterns. For example, all the series show significant short-run increases in volatility in the first half of 2018, and in the first half of 2020, there is an obvious and very sound change in the average

volatility level. The shifts in 2018 are likely to be the result of GDP growth deceleration due to sanctions policies of foreign countries, and depreciation of the national currency. The shifts in 2020 are obvious consequences of the COVID-19 crisis and of the oil market shocks. We address sampling around those periods with caution, yet we expect a loss in the predictive power of the models anyway.

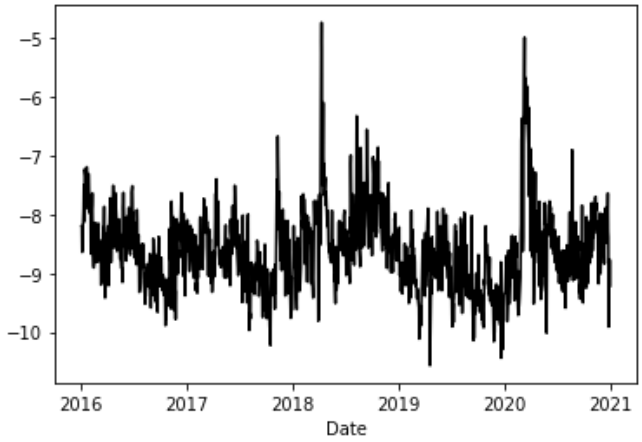


Figure 1: Dynamics of logarithm of realized volatility of returns, SBERBANK

Secondly, as we aim to capture changes in mean of our dependent variables that occur over time, we consider the distribution of the average level of the dependent variable across weekdays. For all companies, Thursdays feature the highest average volatility level. We also observe that, for all companies, except GAZPROM and POLYUS, RV is the lowest on Mondays on average. For GAZPROM and POLYUS, Friday features the lowest RV. These findings line up with the literature. We include the corresponding dummies into our datasets.

Next, we select variables that are company-specific according to the information that we want to account for in our models. We present and analyze the descriptive statistics, including sample averages, standard deviations, and autocorrelations. For all companies, RV and High-Low proxy of liquidity exhibit persistent significant autocorrelation. Growth rates of dividend price ratio and earning price ratio show only a few significant autocorrelation terms, and this order appears to be company-specific.

Finally, to examine correlations between the dependent variable and some potential explanatory variables, we examine correlation-scatterplot diagrams (in Appendix). For each company,

we select the following indicators: logarithm of stock realized volatility (the dependent variable), logarithm of realized volatility of S&P 500 (captures potential spillovers from the global market), High-Low liquidity proxy (liquidity spillovers), RGBI (proxy of the economy steadiness), and the log-returns of RTSI (the local market spillovers). The results are rather similar across the companies. The log-RV is significantly correlated with the other variables; it approves the inclusion of these indicators in the models. However, the there are obvious non-linear dependencies present, hence, machine learning algorithms are naturally expected to perform well.

Overall, the RV on the Russian stock market features properties typically outlined in the literature. Also, this measure appears to be relatively similar across the companies with drastic changes of values early in 2018 and 2020. We also find weekday effects in the average level of RV. Further, we observe sound correlation between RV and various indicators that we intend to use in the models as explanatory features. However, these dependencies are ambiguous. We need to conduct a thorough search for the optimal forecasting specification in the case of each particular asset and over a particular period of time.

3.4 Modeling technique

We aim to forecast realized volatility on the Russian stock market as precisely as possible, and to identify the effects of potentially informative variables on the volatility dynamics. We perform our analysis iteratively, going over different predictive algorithms combined with different groups of explanatory variables. Each such combination is estimated for each of the assets in our datasets on rolling samples. The performance of the different models is then compared across time, across sets of explanatory variables, across algorithms, and across assets.

For each dataset we perform the following sequence of steps.

1. We split the data across time into 11 samples, using the rolling window scheme, rolling forward by one quarter at a time. Each sub-sample consists of 10 consecutive quarters.

    (a) The first two years (8 quarters) are used to train the models.

(b) The next one quarter is a validation sample, used to tune the hyperparameters of the models.

(c) The succeeding one quarter is the testing sample, used to assess the models' performance via comparing the RMSE of the forecasts of log-RV.

2. We construct 40 different model specifications, combining different model types with sets of explanatory variables.

   (a) HAR-RV is our benchmark model, and our ML algorithms are: Lasso, RF, GB, and LSTM.

   (b) We split all the explanatory variables into 5 groups (see Table 1 in Section 3.3), and for every model specification we add groups consecutively.

   (c) Each of the algorithms is trained on the sets of variables either including lagged values or not, except in the case of HAR-RV and LSTM.

3. All the models are trained in terms of minimization of RMSE. The hyperparameters of the models are tuned via a grid search to minimize RMSE, too, but on the validation sample.

4. We choose RMSE as the main measure of the predictive power, too.

5. In addition to plain comparison of the RMSE of different models, we consider relative win-rates of different specifications by algorithm and by set of variables. We track the velocity of the accumulation of the squared sum of forecast errors on the testing samples, to gather insights about which models and which sets of variables should be preferred, and how they can be further improved.

3.5 Results

3.5.1 HAR-RV and the explanatory power of the variables

The full results of the estimation of the benchmark HAR-RV regressions are available in Appendix 14, page 91. Overall, the most commonly selected statistically significant variables

are the lagged daily, weekly, and monthly log-RV, the liquidity proxy High-Low, and the log-RV of S&P 500. Regarding the other variables, growth rate of exports and log-returns of RTSI appear to be significant for some firms (3 of 9). The variables that are not significant in any of the regressions are from the group of calendar effects. The signs of the significant coefficients coincide with those found in other studies which considered similar effects.

Regarding the global market effects, the result is that, the higher the S&P 500 RV is, the higher the stock realized volatility of a Russian company tends to be. This confirms the existence of spillover effect from the global market. The liquidity proxy High-Low also has a reasonable sign: the higher the liquidity of a stock is, the lower is its volatility. As for the basic variables, their signs make sense as well, because the higher the weekly or monthly volatility is, the higher the value of the dependent variable is. The log-returns on RTSI (the local market effect) show a negative effect, similarly to the findings of Christiansen et al. (2012).

Though the calendar effects are not significant, the signs of their coefficients also coincide with those typically found in the literature. Similarly, for example, to the results of Diaz-Mendoza and Pardo (2020) and Todorova and Souček (2014), we find that, in the Russian market, too, volatility decreases after a holiday or a weekend, and due to high overnight returns.

We run Breusch-Godfrey LM tests for residual autocorrelation, and find that the results vary across the firms. There are companies for which there is significant residual autocorrelation in all the specifications (ROSNEFT, NORNICKEL, POLYUS, and MAGNIT). For the other firms, the residual correlation vanishes with inclusion of rather few additional variables (GAZPROM, LUKOIL and NOVATEK). Finally, for SBERBANK and POLYMETAL there are no signs of residual autocorrelation in all specifications of HAR-RV.

We compare the AICa of the specifications, and find that for most companies the specification of HAR-RV that includes spillover effects appears to be the best. This pattern is violated only for LUKOIL, for which regression with inclusion of all variables should be chosen. In general, this result has proven that the selected explanatory variables contain valuable information for explanation of stock realized volatility.

3.5.2 ML and predictive power of the models

To report results of machine learning algorithms, we select top-1 models of each type in terms of average RMSE, and put them on one graph for each company (see Appendix F.1, page 132). The most distinct features for all figures are peaks in RMSE early in 2020 and in different quarters of 2018 and 2019. The most reliable explanation, in our opinion, is the market shock from COVID-19 early in 2020, and the oil market shock in the same time. For the 2018-2019 shocks, there could be multiple reasons, most likely including deceleration in growth of GDP due to sanctions policies of foreign countries, pension reform, and depreciation of the national currency in 2018 and 2019. However, in the quarter after those peaks, the RMSE lowers significantly, which indicates that the proposed models adjust to the new information, process it, and can regain their predictive power.

Further, for most of the companies, GB and RF algorithms appear to be the weakest. GB appears to be the worst overall in most cases (across time, across specifications, and across assets). This suggests that consequent learning of regression trees might not be the best for forecasting stock realized volatility. However, unlike the others, RF models have the lowest RMSE for all test periods for POLYUS. In turn, Lasso and HAR-RV appear to be the best models, replacing each other in the leading position in different test quarters. The benchmark model is chosen in its basic specification most of the time, while Lasso performs better with inclusion of all types of variables into the model. The model specifications without lags of variables are chosen more often, meaning that lags do not lend much forecasting power into the algorithms.

To understand the predictive capabilities of different models better, we considered top-3 specifications for each class of ML algorithms for each dataset (and the top-1 benchmark model specification for comparison). The predictive performance measures and description of the specifications are in Appendix F.2, page 137. As in the previous step, Lasso and HAR-RV deliver the lowest average RMSE on the testing samples, followed by RF and GB. For LSTM, most notably, with the inclusion of extra variables, an increase in RMSE is much higher than for the other models. We believe that the poor performance of LSTM is a sign of overfitting. Nevertheless, it should be noticed that any model can deliver the lowest RMSE

in a particular quarter. Thus, it is important that the majority of top-3 specifications of each model are based on variables without lags. Moreover, we conclude that even though top-1 specifications can be based on the basic variables only, among the top-3 specifications there is commonly at least one model with addition of extra variables, which does not perform significantly differently across the top-3 options. This proves that various groups of variables indeed carry valuable information about volatility and are important for better forecasting. To confirm this claim further, we repeat the analysis of the top-3 specifications, excluding the influence of the 1st and 2nd quarters of 2020. The results are comparable to those from the previous step.

Since any algorithm can perform best in particular test quarters, we continue analysis of the results and compare the win-rate of the models by class, showing the number of cases among the firms for which a particular class of models appears to be the best in each particular quarter and on average overall. See Table 2. The most frequent winner is Lasso, followed by HAR-RV and RF. Hence, this result overlaps previous findings for benchmark model and Lasso. However, the result for RF demonstrates that, though RF does not appear as the top-1 model, it is still a powerful algorithm. LSTM and GB have the lowest win-rates. However, the two periods when either LSTM or RF show the highest win-rate are the 3d quarter of 2018, and the 3d quarter of 2020, right after the periods with abnormally high volatility for most of the companies. This suggests that these particular algorithms can adjust faster to changes in the patterns and absorb new arriving information better than the other models. If so, it is worth an attempt to improve their performance by more frequent re-training (more on this in the conclusion).

3.5.3 Prediction-based importance of the variables

Because Lasso is among the best model classes in terms of prediction across both time periods and assets, we are able to determine which variables are the most significant. We point out two groups of relatively significant variables: those that were sustainably chosen by the algorithm, and those that were impermanently, but frequently chosen. The groups are presented in tables in Appendix F.3, page 143, and Table 3 below summarizes the results

Table 2: Total win-rate of models by class and testing sample period

Period	**Models**				
	HAR-RV	*Lasso*	*LSTM*	*RF*	*GB*
2018Q2	2	**4**	0	1	3
2018Q3	1	2	1	**3**	2
2018Q4	2	**6**	0	1	0
2019Q1	**3**	2	1	2	1
2019Q2	1	2	2	2	2
2019Q3	**4**	3	0	2	0
2019Q4	2	**6**	0	1	0
2020Q1	2	**5**	0	2	0
2020Q2	**4**	**4**	0	1	0
2020Q3	2	1	**3**	**3**	0
2020Q4	**3**	**3**	0	2	1
Average	2.36	**3.45**	0.64	1.82	0.82

across all firms.

Table 3: Best overall variables, chosen by Lasso

	Group of variables
Sustainably chosen	Log RV, log RV weekly, log RV monthly, is after weekend, is Friday
Frequently chosen	Log-RV of S&P, log-RV of Brent, growth rates of imports, growth rates of exports, growth rates of GDP, growth rates of CPI, overnight returns, RGBI, earning price ratios, dividend price ratios, growth rates of housing starts, High-Low

According to these results, the first group includes basic HAR-RV model variables: logarithms of daily, weekly, and monthly realized volatility. It happens because the HAR-RV gives an accurate description of the autoregressive process of RV, and with so few variables the Lasso must be very close to the baseline linear model. Note that, even though the calendar effects are often insignificant in-sample, they are rather persistently chosen to improve prediction by Lasso (e.g., the Friday effect).

The second group contains less frequently chosen variables, including indicators of spillover effects (log-RV of S&P 500 and Brent oil price returns). In many cases, macroeconomic factors including growth rates of GDP, CPI, imports and exports, and housing starts are significant. Moreover, financial indicators including growth rates of earning price ratios, growth rates of dividend price ratios, growth rates of housing starts, High-Low proxy of liquidity, and RGBI

are important in forecasting realized volatility for most companies. Lastly, overnight returns were frequently chosen by Lasso. Similarly to the calendar effects, many of these variables are not detected as carriers of significant explanatory power in-sample by the benchmark.

Compared to the results obtained by the benchmark model, many more variables from various groups are chosen by Lasso. However, logs of daily, weekly, and monthly RV and High-Low are chosen by both algorithms.

Our results demonstrate that linear models are more suitable for RV forecasting than more complicated machine learning algorithms, at least in our framework. However, Lasso provides more accurate forecasts than HAR-RV. Importantly, in terms of predictive power optimization, Lasso tends to choose more variables as being valuable, while the benchmark model works the best on the basic sets of regressors.

There are multiple reasons for the relative failure of GB, RF, and LSTM in the task of RV forecasting. We believe that the most crucial source of high prediction numbers of errors by these algorithms is overfitting. Another possible reason for the failure of these algorithms is re-training that is not frequent enough. In both cases, the underlying reason must be the nature of the volatility process itself, as it is essentially noisy. It is less likely but possible that the tree structure used by GB and RF may be unsuitable for forecasting time series such as RV. Lastly, LSTM is the closest to HAR-RV and Lasso in terms of predictive power, but the problem of overfitting is likely to have escalated for this model.

3.6 Discussion

3.6.1 Applications of the results

There are several ways the results of this study can be implemented. Firstly, we were able to identify the most suitable model for forecasting realized volatility on the Russian stock market, so researchers and investors who want to study this topic or trade on the market can use the model. Secondly, if researchers or investors want to build other models for forecasting stock volatility, they can use our findings on the significant predictors of realized volatility. Thirdly, with help of our results, traders can quantitatively assess the future short-run risks of an asset

on the Russian stock market. Lastly, as realized volatility is important for optimal portfolio allocation, our results can be used by portfolio investors to improve their (re)allocation decisions.

3.6.2 Limitations of the study

We encountered the impossibility of acquiring some data. Although we included variables related to calendar effects, spillover effects, and financial and macroeconomic effects, many factors that can influence RV could not be taken into account. The most obvious reason is unavailability of data. For instance, investors sentiment is expected to be an important predictor of RV, yet we have not yet managed to access suitable structured or unstructured data that would have been convenient to use in our research.

Another challenge is the computational capacity requirements necessary for appropriately fine and frequent tuning of the models, particularly, the computationally heavy ML algorithms (RF, LSTM). We briefly studied the velocity with which the sum of the squared prediction error is accumulated by different models within a particular testing quarter. Figure 2 shows the accumulation process for SBERBANK volatility top-1 by-class predictors during the 1st quarter of 2019. The overall winning model for this firm and this period was Lasso (it shows the lowest accumulated sum at the end of the quarter, day 60, see the left panel of Figure 2). However, the superiority of this model is not stable within the period. Obviously, LSTM and GB have higher values of the squared error to begin with. This supports our intuition about the overfitting problem. However, the other three models start off rather close to each other in the beginning of the quarter, with RF being a sound leader for several days (see the right panel of Figure 2). RF becomes outdated rather quickly (after approximately 6 days), and does not recover throughout the rest of the testing sample. Moreover, the accumulation of the sum of the squared error occurs, on average, with increasing rates for all the models, with dramatic increases in some periods, which possibly signal arrivals of new information not yet accounted for by the trained models. These observations support the idea that more frequent re-training of the models might significantly improve predictive power. Even though this seems a rather obvious path to take, we leave it for future research,

as it is too computationally demanding, particularly in the case of the GB, RF, and LSTM algorithms.

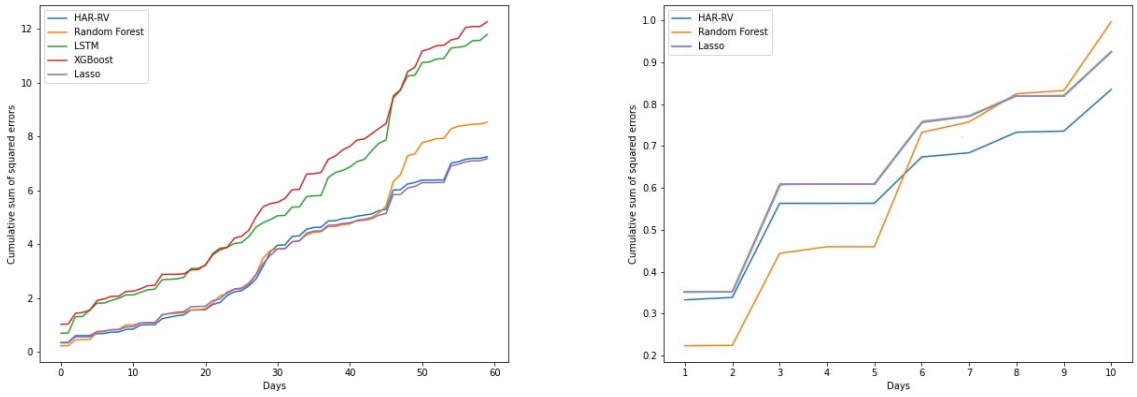


Figure 2: Top-1 predictors cumulative sum of squared errors, SBERBANK, 2019Q1

3.7 Conclusion

We aim to employ data on the Russian stock market and to compare the suitability of the benchmark HAR-RV with several ML algorithms (Lasso, Random Forest, Gradient Boosting, Long Short-Term Memory) in the task of forecasting daily RV of selected top stocks on the Russian stock market. We further seek to identify the most valuable factors for explaining the dynamics and forecasting the future values of the RV.

We collect a novel and extensive dataset for the top-9 Russian companies based on MICEX, consisting of our variable of interest and various groups of additional variables, including calendar effects, financial variables, spillover effects and macroeconomic variables. For each of the models, we constructed a number of specifications based on either HAR-RV or ML algorithms that are trained on various sets of explanatory variables.

The results show that Gradient Boosting, Random Forest, and LSTM did not appear to perform well in the forecasting task. The best performing models were Lasso and HAR-RV. From Lasso, we were able to highlight the most significant factors for forecasting the RV. The variables that showed the most effect on future RV across all companies are: logarithm of daily, weekly, and monthly realized volatility, High-Low proxy of liquidity, calendar effects,

and some macroeconomic variables and financial market and spillover effects, including as, for example, logarithm of realized volatility of S&P, growth rates of CPI and GDP, growth rates of earning price ratio and dividend price ratio.

We also find that, once trained, the specifications become outdated rather quickly. Their predictive performance could be improved by finer tuning and more frequent re-training. This is a computationally heavy task, which could be addressed in future research. Furthermore, as most companies in our study are from the oil and gas sector (because of the industry specificity of the Russian economy), spillovers between sectors could not be investigated fully, opening another promising direction for further research.

Conclusion

In conclusion, our research delves into the intricacies of high-dimensional statistical analysis, exploring methodologies and applications across three key chapters. In Chapter 1, we addressed the estimation challenges of Gaussian and t copulas in ultra-high dimensions. Leveraging large covariance matrix shrinkage estimators, our approach demonstrated efficacy in handling up to thousands of variables with considerably reduced sample lengths. The findings showcased not only the precision in estimating copula matrix parameters but also the practical applications, particularly in portfolio allocation. Despite acknowledged limitations in capturing all data properties, Gaussian and t copulas emerged as valuable tools in various applications, serving either as primary dependence models or essential benchmarks.

Chapter 2 introduced a novel high-dimensional approach to estimating the skew-t copula, surpassing existing copula size limitations. The two-step algorithm, incorporating simulated method of moments and analytical non-linear shrinkage, exhibited robustness even with limited observations. The application of the skew-t copula in a dynamic portfolio allocation exercise underscored its superiority, particularly in accommodating tail dependence among asset pairs. The results emphasized its crucial role in constructing portfolios that outperform alternative models, thus contributing significantly to the literature on copulas and portfolio optimization.

Shifting focus to Chapter 3, we extended our exploration into the realm of machine learning for forecasting daily realized volatility on the Russian stock market. The study identified the most suitable model, with Lasso and HAR-RV emerging as the top performers. The detailed analysis of predictors provided valuable insights for researchers, investors, and traders seeking to understand and navigate short-run risks on the Russian stock market. The limitations, including data unavailability and computational challenges, signal the need for future research to overcome these obstacles and enhance the performance of predictive models.

In essence, our work represents a multifaceted contribution to high-dimensional statistical methods, copulas, and machine learning applications in the financial domain. The demonstrated benefits in portfolio allocation, tail dependence modeling, and volatility forecasting

underscore the practical relevance of our findings. This opens avenues for future research, where refinements in methodologies, addressing computational challenges, and exploring sectoral spillovers can further enhance the applicability and effectiveness of these statistical and machine learning tools in financial analysis.

Summary

In summary, our research tackles challenges in high-dimensional statistical analysis. In Chapter 1, we employ large covariance matrix shrinkage estimators for precise estimation of Gaussian and t copulas in ultra-high dimensions. Results highlight their effectiveness in portfolio allocation. Chapter 2 introduces a novel approach for high-dimensional skew-t copula estimation, showcasing robustness and superiority in dynamic portfolio allocation. In Chapter 3, we leverage machine learning for daily realized volatility forecasting on the Russian stock market, with Lasso and HAR-RV as top performers. Our findings offer practical insights for risk assessment and portfolio optimization. Overall, the research contributes to advancing methodologies in high-dimensional statistical analysis, copulas, and machine learning applications in finance.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics 44*(2), 182–198.

Aganin, A. et al. (2017). Forecast comparison of volatility models on russian stock market. *Applied Econometrics 48*, 63–84.

Aganin, A. D. (2020). Russian stock index volatility: Oil and sanctions. *Voprosy Ekonomiki (2)*, 86–100.

Anatolyev, S., R. Khabibullin, and A. Prokhorov (2018). Estimating asymmetric dynamic distributions in high dimensions". *Asymmetric Dependence in Finance: Diversification, Correlation and Portfolio Management in Market Downturns, 1*, 169–197.

Anatolyev, S. and V. Pyrlik (2022). Copula shrinkage and portfolio allocation in ultra-high dimensions. *Journal of Economic Dynamics and Control 143*, 104508.

Andersen, T. G. and T. Bollerslev (1997). Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *The journal of Finance 52 (3)*(3), 975–1005.

Andersen, T. G. and T. Bollerslev (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, 885–905.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71 (2)*(2), 579–625.

Arago, V. and A. Fernandez (2002). Expiration and maturity effect: empirical evidence from the spanish spot and futures stock index. *Applied Economics 34 (13)*(13), 1617–1626.

Atalay, F. and A. E. Tercan (2017). Coal resource estimation using gaussian copula. *International Journal of Coal Geology 175*, 1–9.

Audrino, F., F. Sigrist, and D. Ballinari (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting 36 (2)*(2), 334–357.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 367–389.

Balli, F., H. R. Hajhoj, S. A. Basher, and H. B. Ghassan (2015). An analysis of returns and volatility spillovers and their determinants in emerging asian and middle eastern countries. *International Review of Economics & Finance 39*, 311–325.

Bates, D. and M. Maechler (2019). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-17.

Bazhenov, T. and D. Fantazzini (2019). Forecasting realized volatility of russian stocks using google trends and implied volatility. *Russian Journal of Industrial Economics 12 (1)*(1), 79–88.

Będowska-Sójka, B. and A. Kliber (2019). The causality between liquidity and volatility in the polish stock market. *Finance Research Letters 30*, 110–115.

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2017). Julia: A fresh approach to numerical computing. *SIAM Review 59*(1), 65–98.

Bollen, N. P. and R. E. Whaley (1999). Do expirations of hang seng index derivatives affect stock market volatility? *Pacific-Basin Finance Journal 7 (5)*(5), 453–470.

Brechmann, E. C. and C. Czado (2013). Risk management with high-dimensional vine copulas: An analysis of the euro stoxx 50. *Statistics & Risk Modeling 30*(4), 307–342.

Brechmann, E. C., C. Czado, and K. Aas (2012). Truncated regular vines in high dimensions with application to financial data. *Canadian Journal of Statistics 40*(1), 68–85.

Breiman, L. (2001). Random forests. *Machine learning 45 (1)*(1), 5–32.

Broda, S. A. and M. S. Paolella (2020). Archmodels. jl: Estimating arch models in julia. *Jl: Estimating Arch Models in Julia (March 9, 2020)*.

Chen, Y., W. Li, and F. Qu (2019). Dynamic asymmetric spillovers and volatility interdependence on china's stock market. *Physica A: Statistical Mechanics and its Applications 523*, 825–838.

Chou, H. C., W. N. Chen, and D. H. Chen (2006). The expiration effects of stock-index derivatives: Empirical evidence from the taiwan futures exchange. *Emerging Markets Finance and Trade 42 (5)*(5), 81–102.

Christiansen, C., M. Schmeling, and A. Schrimpf (2012). A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics 27 (6)*(6), 956–977.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics 7 (2)*(2), 174–196.

Czado, C., E. C. Brechmann, and L. Gruber (2013). Selection of vine copulas. In *Copulae in Mathematical and Quantitative Finance*, pp. 17–37. Springer.

Daniels, M. J. and R. E. Kass (2001). Shrinkage estimators for covariance matrices. *Biometrics 57*(4), 1173–1184.

Daul, S., E. G. De Giorgi, F. Lindskog, and A. McNeil (2003). The grouped t-copula with an application to credit risk. *Available at SSRN 1358956*.

De Leon, A. R. and K. C. Chough (2013). *Analysis of Mixed Data: Methods & Applications*. CRC Press.

De Nard, G., R. F. Engle, O. Ledoit, and M. Wolf (2020). Large dynamic covariance matrices: enhancements based on intraday data. *University of Zurich, Department of Economics, Working Paper* (356).

De Nard, G., O. Ledoit, and M. Wolf (2018). Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics*.

Demarta, S. and A. J. McNeil (2005). The t copula and related copulas. *International Statistical Review 73*(1), 111–129.

DeMiguel, V., L. Garlappi, and R. Uppal (2007). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies 22*(5), 1915–1953.

Diaz-Mendoza, A.-C. and A. Pardo (2020). Holidays, weekends and range-based volatility. *The North American Journal of Economics and Finance 52*, 101–124.

Ding, X., Y. Zhang, T. Liu, and J. Duan (2015). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis 59*, 52–69.

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

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*(1), 186–197.

Fang, T., T.-H. Lee, and Z. Su (2020). Predicting the long-term stock market volatility: A garch-midas model with variable selection. *Journal of Empirical Finance 58*, 36–49.

Fantazzini, D. and T. Shangina (2019). The importance of being informed: forecasting market risk measures for the russian rts index future using online data and implied volatility over two decades. *Applied Econometrics 3 (55)*.

Filzmoser, P., H. Fritz, and K. Kalcher (2018). *pcaPP: Robust PCA by Projection Pursuit*. R package version 1.9-73.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Fu, L. and Y.-G. Wang (2016). Efficient parameter estimation via gaussian copulas for quantile regression with longitudinal data. *Journal of Multivariate Analysis 143*, 492–502.

Guidolin, M. and A. Timmermann (2008). International asset allocation under regime switching, skew, and kurtosis preferences. *The Review of Financial Studies 21*(2), 889–935.

Haff, L. (1980). Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 586–597.

Hamid, S. A. and Z. Iqbal (2004). Using neural networks for forecasting volatility of s&p 500 index futures prices. *Journal of Business Research 57 (10)*(10), 1116–1125.

Hammoudeh, S. M., Y. Yuan, and M. McAleer (2009). Shock and volatility spillovers among equity sectors of the gulf arab stock markets. *The Quarterly Review of Economics and Finance 49 (3)*(3), 829–842.

Han, Y., P. Li, and Y. Xia (2017). Dynamic robust portfolio selection with copulas. *Finance Research Letters 21*, 190–200.

Harvey, C. R., J. C. Liechty, M. W. Liechty, and P. Müller (2010). Portfolio selection with higher moments. *Quantitative Finance 10*(5), 469–485.

He, Y., L. Zhang, J. Ji, and X. Zhang (2019). Robust feature screening for elliptical copula regression model. *Journal of Multivariate Analysis 173*, 568–582.

He, Y., X. Zhang, and L. Zhang (2018). Variable selection for high dimensional gaussian copula regression model: An adaptive hypothesis testing procedure. *Computational Statistics & Data Analysis 124*, 132–150.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9 (8)*(8), 1735–1780.

Hofert, M., I. Kojadinovic, M. Maechler, and J. Yan (2018). *copula: Multivariate Dependence with Copulas*. R package version 0.999-19.1.

Hofert, M., M. Mächler, and A. J. Mcneil (2012). Likelihood inference for archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis 110*, 133–150.

Hörmann, W. and H. Sak (2010). t-copula generation for control variates. *Mathematics and Computers in Simulation 81*(4), 782–790.

Huang, J.-J., K.-J. Lee, H. Liang, and W.-F. Lin (2009). Estimating value at risk of portfolio by conditional copula-garch method. *Insurance: Mathematics and Economics 45*(3), 315–324.

Huang, J. Z., N. Liu, M. Pourahmadi, and L. Liu (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika 93*(1), 85–98.

Ingle, V. and S. Deshmukh (2021). Ensemble deep learning framework for stock market data prediction (edlf-dp). *Global Transitions Proceedings 2 (1)*, 47–66.

Ivan Kojadinovic and Jun Yan (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software 34*(9), 1–20.

Jeff Bezanson, Stefan Karpinski, V. B. S. and other contributors (2022). Distributed.jl.

Jun Yan (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software 21*(4), 1–21.

Kang, W., R. A. Ratti, and K. H. Yoon (2015). The impact of oil price shocks on the stock market return and volatility relationship. *Journal of International Financial Markets, Institutions and Money 34*, 41–54.

Karmakar, M. (2017). Dependence structure and portfolio risk in indian foreign exchange market: A garch-evt-copula approach. *The Quarterly Review of Economics and Finance 64*, 275–291.

Kojadinovic, I. and J. Yan (2010). Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance: Mathematics and Economics 47*(1), 52–63.

Kollo, T. and G. Pettere (2010). Parameter estimation and application of the multivariate skew t-copula. In *Copula Theory and its Applications*, pp. 289–298. Springer.

Kolm, P. N., R. Tütüncü, and F. J. Fabozzi (2014). 60 years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research 234*(2), 356–371.

Kristjanpoller, W. and M. C. Minutolo (2015). Gold price volatility: A forecasting approach using the artificial neural network–garch model. *Expert systems with applications 42 (20)*(20), 7245–7251.

Kwak, M. (2017). Estimation and inference on the joint conditional distribution for bivariate longitudinal data using gaussian copula. *Journal of the Korean Statistical Society 46*(3), 349–364.

Ledoit, O. and S. Péché (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields 151*(1-2), 233–264.

Ledoit, O. and M. Wolf (2004a). Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management 30*(4), 110–119.

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

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

Ledoit, O. and M. Wolf (2017a). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *The Review of Financial Studies 30*(12), 4349–4388.

Ledoit, O. and M. Wolf (2017b). Numerical implementation of the QuEST function. *Computational Statistics & Data Analysis 115*, 199–223.

Ledoit, O. and M. Wolf (2019). Quadratic shrinkage for large covariance matrices. *University of Zurich, Departmenf of Economics, Working Paper* (335).

Ledoit, O. and M. Wolf (2022). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics 20*(1), 187–218.

Ledoit, O., M. Wolf, et al. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics 48*(5), 3043–3065.

Li, C., Y. Huang, and Y. Xue (2019). Dependence structure of gabor wavelets based on copula for face recognition. *Expert Systems with Applications 137*, 453–470.

Li, C., Y. Huang, and L. Zhu (2017). Color texture image retrieval based on gaussian copula models of gabor wavelets. *Pattern Recognition 64*, 118–129.

Li, F., J. Zhou, and C. Liu (2018). Statistical modelling of extreme storms using copulas: A comparison study. *Coastal Engineering 142*, 52–61.

Liu, Y. (2019). Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. *Expert Systems with Applications 132*, 99–109.

Lourme, A. and F. Maurer (2017). Testing the gaussian and student's t copulas in a risk management framework. *Economic Modelling 67*, 203–214.

Luo, X. and S. Qin (2017). Oil price uncertainty and chinese stock returns: New evidence from the oil volatility index. *Finance Research Letters 20*, 29–34.

Lyon, S. (2022). Plotlyjs.jl.

Marius Hofert and Martin Mächler (2011). Nested archimedean copulas meet R: The nacopula package. *Journal of Statistical Software 39*(9), 1–20.

Martens, M., D. Van Dijk, and M. De Pooter (2009). Forecasting s&p 500 volatility: Long memory, level shifts, leverage effects, day-of-the-week seasonality, and macroeconomic announcements. *International Journal of Forecasting 25 (2)*(2), 282–303.

Mensi, W., R. Nekhili, X. V. Vo, T. Suleman, and S. H. Kang (2020). Asymmetric volatility connectedness among us stock sectors. *The North American Journal of Economics and Finance 56*, 101327.

Mersmann, O. (2019). *microbenchmark: Accurate Timing Functions*. R package version 1.4-7.

Michaud, R. O. and R. O. Michaud (2008). *Efficient asset management: a practical guide to stock portfolio optimization and asset allocation*. Oxford University Press.

Mogensen, P. and A. Riseth (2018). Optim: A mathematical optimization package for julia. *Journal of Open Source Software 3*(24).

Müller, D. and C. Czado (2017). Selection of sparse vine copulas in high dimensions with the lasso. *arXiv preprint arXiv:1705.05877*.

Müller, D. and C. Czado (2019). Dependence modeling in ultra high dimensions with vine copulas and the graphical lasso. *Computational Statistics & Data Analysis*.

Nagapetyan, A. et al. (2019). Precondition stock and stock indices volatility modeling based on market diversification potential: Evidence from russian market. *Applied Econometrics 4 (56)*, 45–61.

Ning, C. (2010). Dependence structure between the equity market and the foreign exchange market–a copula approach. *Journal of International Money and Finance 29*(5), 743–759.

Nonejad, N. (2017). Forecasting aggregate stock market volatility using financial and macroeconomic predictors: Which models forecast best, when and why? *Journal of Empirical Finance 42*, 131–154.

Novomestky, F. (2012). *matrixcalc: Collection of functions for matrix calculations*. R package version 1.0-3.

Oh, D. H. and A. J. Patton (2013). Simulated method of moments estimation for copula-based multivariate models. *Journal of the American Statistical Association 108*(502), 689–700.

Oh, D. H. and A. J. Patton (2016). High-dimensional copula-based distributions with mixed frequency data. *Journal of Econometrics 193*(2), 349–366.

Oh, D. H. and A. J. Patton (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics 35*(1), 139–154.

Parisi, A., F. Parisi, and D. Díaz (2008). Forecasting gold price changes: Rolling and recursive neural network models. *Journal of Multinational financial management 18 (5)*(5), 477–487.

Patton, A. J. (2009). Copula–based models for financial time series. In *Handbook of Financial Time Series*, pp. 767–785. Springer.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis 110*, 4–18.

Patton, A. J. (2013). Copula methods for forecasting multivariate time series. In *Handbook of Economic Forecasting*, Volume 2, pp. 899–960. Elsevier.

Paye, B. S. (2012). 'déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics 106 (3)*(3), 527–546.

Pyrlik, V., A. Leonova, and P. Elizarov (2021). Forecasting realized volatility using machine learning and mixed-frequency data (the case of the russian stock market). *CERGE-EI Working Paper Series* (713).

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramprasad, P. (2016). *nlshrink: Non-Linear Shrinkage Estimation of Population Eigenvalues and Covariance Matrices*. R package version 1.0.1.

Schindler, D. and C. Jung (2018). Copula-based estimation of directional wind energy yield: A case study from germany. *Energy Conversion and Management 169*, 359–370.

Shahzad, H., H. N. Duong, P. S. Kalev, and H. Singh (2014). Trading volume, realized volatility and jumps in the australian stock market. *Journal of International Financial Markets, Institutions and Money 31*, 414 – 430.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Smith, M. S. (2021). Implicit copulas: An overview. *Econometrics and Statistics*.

Smith, M. S., Q. Gan, and R. J. Kohn (2012). Modelling dependence using skew t copulas: Bayesian inference and applications. *Journal of Applied Econometrics 27*(3), 500–522.

Sukcharoen, K., T. Zohrabyan, D. Leatham, and X. Wu (2014). Interdependence of oil prices and stock market indices: A copula approach. *Energy Economics 44*, 331–339.

Thampanya, N., J. Wu, M. A. Nasir, and J. Liu (2020). Fundamental and behavioural determinants of stock return volatility in asean-5 countries. *Journal of International Financial Markets, Institutions and Money 65*, 101193.

Todorova, N. and M. Souček (2014). The impact of trading volume, number of trades and overnight returns on forecasting the daily realized range. *Economic modelling 36*, 332–340.

Valle, D. and D. Kaplan (2019). Quantifying the impacts of dams on riverine hydrology under non-stationary conditions using incomplete data and gaussian copula models. *Science of The Total Environment 677*, 599–611.

van Binsbergen, J. H. and M. W. Brandt (2007). Solving dynamic portfolio choice problems by recursing on optimized portfolio weights or on the value function? *Computational Economics 29*(3-4), 355–367.

Van de Vyver, H. and J. Van den Bergh (2018). The gaussian copula model for the joint deficit index for droughts. *Journal of Hydrology 561*, 987–999.

Vidal, A. and W. Kristjanpoller (2020). Gold volatility prediction using a cnn-lstm approach. *Expert Systems with Applications 157*, 113481.

Wang, X., C. Wu, and W. Xu (2015). Volatility forecasting: The role of lunch-break returns, overnight returns, trading volume and leverage effects. *International Journal of Forecasting 31 (3)*(3), 609–619.

Wang, Y.-H. and Y.-J. Hsiao (2010). The impact of non-trading periods on the measurement of volatility. *Review of Pacific Basin Financial Markets and Policies 13 (04)*(04), 607–620.

Wei, T. and V. Simko (2017). *R package "corrplot": Visualization of a Correlation Matrix*. (Version 0.84).

Wen, X., Y. Wei, and D. Huang (2012). Measuring contagion between energy market and stock market during financial crisis: A copula approach. *Energy Economics 34*(5), 1435–1446.

Weston, S. (2019a). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.15.

Weston, S. (2019b). *foreach: Provides Foreach Looping Construct*. R package version 1.4.7.

Wong, F., C. K. Carter, and R. Kohn (2003). Efficient estimation of covariance selection models. *Biometrika 90*(4), 809–830.

Wongbangpo, P. and S. C. Sharma (2002). Stock market and macroeconomic fundamental dynamic interactions: Asean-5 countries. *Journal of Asian Economics 13 (1)*(1), 27–51.

Xiong, R., E. P. Nichols, and Y. Shen (2015). Deep learning stock volatility with google domestic trends. *arXiv preprint arXiv:1512.04916*.

Xu, C. (2014). Expiration-day effects of stock and index futures and options in sweden: The return of the witches. *Journal of futures markets 34 (9)*(9), 868–882.

Xu, Y., D. Huang, F. Ma, and G. Qiao (2019). Liquidity and realized range-based volatility forecasting: Evidence from china. *Physica A: Statistical Mechanics and its Applications 525*, 1102–1113.

Yahoo Finance (2022). Euro stoxx 50 index (^STOXX50E). Retrieved September, 2022.

Yoshiba, T. (2018). Maximum likelihood estimation of skew-t copulas with its applications to stock returns. *Journal of Statistical Computation and Simulation 88*(13), 2489–2506.

Zhu, X., H. Zhang, and M. Zhong (2017). Volatility forecasting using high frequency data: The role of after-hours information and leverage effects. *Resources Policy 54*, 58–70.

Zimmer, D. M. (2012). The role of copulas in the housing crisis. *Review of Economics and Statistics 94*(2), 607–620.

Zorgati, I., F. Lakhal, and E. Zaabi (2019). Financial contagion in the subprime crisis context: A copula approach. *The North American Journal of Economics and Finance 47*, 269–282.

List of Appendices

Tables

Table 4: Mean (s.d.) time of evaluation of estimators of P, milliseconds

p	p/n	identity true P				arbitrary true P			
		smpl	i-τ	LSh	NLSh	smpl	i-τ	LSh	NLSh
10	$\frac{1}{2}$	0.032	0.333	0.127	263.653	0.031	0.331	0.125	3565.683
		(0.008)	(0.039)	(0.023)	(14.560)	(0.007)	(0.039)	(0.021)	(109.259)
	1	0.031	0.306	0.101	1178.981	0.031	0.327	0.101	3506.683
		(0.008)	(0.034)	(0.017)	(29.304)	(0.008)	(0.156)	(0.018)	(118.759)
	2	0.039	0.310	0.099	3229.914	0.031	0.423	0.096	558.311
		(0.083)	(0.113)	(0.117)	(120.781)	(0.009)	(0.348)	(0.091)	(16.000)
100	$\frac{1}{2}$	0.979	49.784	7.263	132.892	0.972	47.311	7.443	1488.073
		(0.089)	(3.579)	(0.428)	(8.497)	(0.070)	(2.642)	(0.748)	(76.432)
	1	0.515	31.403	3.988	224.496	0.511	30.161	3.928	8047.918
		(0.048)	(3.725)	(0.591)	(11.136)	(0.043)	(2.959)	(0.602)	(76.432)
	2	0.277	22.668	2.220	7733.754	0.272	21.976	2.068	1550.691
		(0.028)	(2.480)	(0.534)	(108.521)	(0.018)	(2.916)	(0.289)	(71.998)
1000	$\frac{1}{2}$	1116.264	50367.650	15802.090	46094.440	1151.178	45576.730	16735.310	444714.600
		(150.527)	(667.872)	(2515.403)	(6107.378)	(144.601)	(469.405)	(2206.170)	(16521.080)
	1	570.525	24408.820	9265.924	98169.430	584.639	23739.780	8951.203	289481.000
		(64.136)	(412.646)	(1153.461)	(6241.310)	(96.366)	(2549.029)	(1407.731)	(43049.690)
	2	260.719	12471.040	4569.003	104647.500	252.870	11564.420	3843.262	84845.240
		(23.468)	(287.130)	(510.589)	(2857.122)	(22.259)	(290.511)	(431.792)	(1691.892)

Table 5: Descriptive statistics of selected variables, SBERBANK

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.61	0.03	-0.0	-0.35
std	0.69	0.13	0.1	0.07
ρ_1	0.73^{***}	0.03	-0.1	0.5^{***}
ρ_2	0.61^{***}	-0.14	-0.12	0.35^{***}
ρ_3	0.56^{***}	-0.06	-0.12	0.3^{***}
ρ_4	0.53^{***}	0.03	0.03	0.28^{***}
ρ_5	0.51^{***}	-0.03	-0.09	0.27^{***}
ρ_6	0.46^{***}	-0.07	0.0	0.23^{***}
ρ_7	0.44^{***}	0.05	0.24^{*}	0.24^{***}
ρ_8	0.43^{***}	0.31^{**}	0.07	0.24^{***}
ρ_9	0.41^{***}	0.05	-0.09	0.25^{***}
ρ_{10}	0.39^{***}	-0.09	-0.15	0.19^{***}

Table 6: Descriptive statistics of selected variables, GAZPROM

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.81	0.01	0.05	-0.33
std	0.69	0.11	0.24	0.07
ρ_1	0.66^{***}	-0.12	0.25^{*}	0.52^{***}
ρ_2	0.55^{***}	-0.22^{*}	-0.11	0.34^{***}
ρ_3	0.49^{***}	0.02	-0.0	0.27^{***}
ρ_4	0.45^{***}	-0.11	0.16	0.25^{***}
ρ_5	0.45^{***}	-0.12	-0.18	0.24^{***}
ρ_6	0.41^{***}	-0.01	-0.06	0.24^{***}
ρ_7	0.39^{***}	0.11	0.08	0.25^{***}
ρ_8	0.38^{***}	-0.08	0.1	0.23^{***}
ρ_9	0.37^{***}	0.08	-0.0	0.24^{***}
ρ_{10}	0.34^{***}	-0.07	-0.06	0.19^{***}

Table 7: Descriptive statistics of selected variables, LUKOIL

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.68	0.0	0.05	-0.34
std	0.74	0.14	0.19	0.07
ρ_1	0.73^{***}	-0.09	0.13	0.54^{***}
ρ_2	0.64^{***}	-0.07	0.19	0.38^{***}
ρ_3	0.59^{***}	-0.22^{*}	-0.11	0.35^{***}
ρ_4	0.56^{***}	0.14	0.16	0.33^{***}
ρ_5	0.53^{***}	0.11	0.31^{**}	0.31^{***}
ρ_6	0.51^{***}	0.0	0.36^{***}	0.3^{***}
ρ_7	0.5^{***}	-0.03	0.18	0.28^{***}
ρ_8	0.5^{***}	-0.03	0.09	0.3^{***}
ρ_9	0.48^{***}	-0.08	0.08	0.31^{***}
ρ_{10}	0.48^{***}	0.15	0.05	0.28^{***}

Table 8: Descriptive statistics of selected variables, NOVATEK

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.3	0.01	0.02	-0.35
std	0.71	0.12	0.15	0.07
ρ_1	0.64^{***}	-0.02	0.13	0.57^{***}
ρ_2	0.58^{***}	-0.25^{*}	-0.14	0.41^{***}
ρ_3	0.52^{***}	0.06	0.18	0.37^{***}
ρ_4	0.49^{***}	-0.01	0.29^{**}	0.37^{***}
ρ_5	0.49^{***}	-0.08	0.06	0.33^{***}
ρ_6	0.47^{***}	0.01	-0.23^{*}	0.31^{***}
ρ_7	0.46^{***}	-0.21	-0.07	0.33^{***}
ρ_8	0.46^{***}	-0.01	0.11	0.33^{***}
ρ_9	0.45^{***}	0.11	-0.09	0.31^{***}
ρ_{10}	0.44^{***}	0.03	-0.4^{***}	0.29^{***}

Table 9: Descriptive statistics of selected variables, ROSNEFT

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.63	0.0	0.06	-0.35
std	0.72	0.15	0.21	0.07
ρ_1	0.71^{***}	0.05	0.2	0.56^{***}
ρ_2	0.64^{***}	-0.11	0.08	0.42^{***}
ρ_3	0.61^{***}	-0.02	0.04	0.39^{***}
ρ_4	0.59^{***}	-0.01	0.04	0.41^{***}
ρ_5	0.57^{***}	0.14	0.03	0.37^{***}
ρ_6	0.53^{***}	0.09	0.0	0.34^{***}
ρ_7	0.53^{***}	0.01	-0.05	0.31^{***}
ρ_8	0.52^{***}	0.12	-0.04	0.34^{***}
ρ_9	0.5^{***}	-0.1	-0.13	0.31^{***}
ρ_{10}	0.46^{***}	-0.17	-0.05	0.28^{***}

Table 10: Descriptive statistics of selected variables, NORNICKEL

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.65	0.01	0.01	-0.34
std	0.66	0.15	0.17	0.07
ρ_1	0.71^{***}	0.01	0.02	0.61^{***}
ρ_2	0.61^{***}	-0.17	-0.09	0.45^{***}
ρ_3	0.55^{***}	-0.12	-0.15	0.34^{***}
ρ_4	0.47^{***}	0.08	-0.12	0.26^{***}
ρ_5	0.45^{***}	-0.12	-0.04	0.24^{***}
ρ_6	0.41^{***}	-0.1	-0.16	0.24^{***}
ρ_7	0.39^{***}	-0.08	-0.04	0.16^{***}
ρ_8	0.36^{***}	0.13	0.1	0.15^{***}
ρ_9	0.32^{***}	-0.01	0.14	0.14^{***}
ρ_{10}	0.33^{***}	-0.19	0.1	0.1^{***}

Table 11: Descriptive statistics of selected variables, POLYMETAL

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.03	0.02	0.01	-0.38
std	0.75	0.17	0.13	0.07
ρ_1	0.66^{***}	-0.22^{*}	-0.23^{*}	0.53^{***}
ρ_2	0.56^{***}	-0.16	-0.12	0.36^{***}
ρ_3	0.52^{***}	0.0	0.01	0.31^{***}
ρ_4	0.49^{***}	-0.02	-0.02	0.25^{***}
ρ_5	0.46^{***}	-0.01	-0.08	0.25^{***}
ρ_6	0.43^{***}	-0.15	-0.25^{*}	0.22^{***}
ρ_7	0.43^{***}	0.13	0.18	0.2^{***}
ρ_8	0.4^{***}	-0.09	-0.01	0.21^{***}
ρ_9	0.39^{***}	0.01	0.06	0.22^{***}
ρ_{10}	0.39^{***}	-0.04	-0.04	0.2^{***}

Table 12: Descriptive statistics of selected variables, POLYUS

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.16	-0.01	0.02	-0.37
std	0.86	0.1	0.09	0.08
ρ_1	0.6^{***}	0.19	-0.05	0.57^{***}
ρ_2	0.51^{***}	-0.27	-0.12	0.36^{***}
ρ_3	0.46^{***}	-0.04	-0.08	0.3^{***}
ρ_4	0.43^{***}	-0.06	-0.13	0.28^{***}
ρ_5	0.39^{***}	-0.23	-0.1	0.26^{***}
ρ_6	0.39^{***}	-0.27	-0.11	0.27^{***}
ρ_7	0.39^{***}	-0.16	0.17	0.26^{***}
ρ_8	0.36^{***}	0.11	-0.04	0.23^{***}
ρ_9	0.38^{***}	0.3	0.24	0.24^{***}
ρ_{10}	0.36^{***}	0.1	0.02	0.22^{***}

Table 13: Descriptive statistics of selected variables, MAGNIT

Statistics	*log RV*	*Dividend price ratio*	*Earning price ratio*	*High − Low*
Mean	-8.35	0.03	0.0	-0.36
std	0.75	0.12	0.12	0.07
ρ_1	0.6^{***}	-0.07	-0.21	0.54^{***}
ρ_2	0.51^{***}	-0.24^{*}	-0.0	0.34^{***}
ρ_3	0.45^{***}	-0.09	-0.09	0.27^{***}
ρ_4	0.4^{***}	0.19	0.13	0.21^{***}
ρ_5	0.37^{***}	0.04	-0.14	0.21^{***}
ρ_6	0.37^{***}	-0.1	-0.05	0.21^{***}
ρ_7	0.34^{***}	-0.02	0.01	0.19^{***}
ρ_8	0.35^{***}	0.08	0.2	0.25^{***}
ρ_9	0.34^{***}	-0.04	-0.09	0.2^{***}
ρ_{10}	0.31^{***}	-0.16	-0.08	0.18^{***}

Table 14: HAR-RV estimation results, SBERBANK

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.482***	0.462***	0.310***	0.310***	0.306***
	(0.045)	(0.045)	(0.053)	(0.052)	(0.050)
log RV weekly	0.322***	0.320***	0.334***	0.317***	0.316***
	(0.055)	(0.055)	(0.054)	(0.053)	(0.055)
log RV monthly	0.070	0.076*	0.080*	0.084*	0.065
	(0.046)	(0.044)	(0.043)	(0.046)	(0.049)
is after weekend		-0.010	-0.008	-0.002	-0.002
		(0.032)	(0.032)	(0.032)	(0.032)
is friday		-0.052	-0.044	-0.047	-0.048
		(0.039)	(0.045)	(0.045)	(0.044)
is after holiday		-0.025	-0.059	-0.061	-0.060
		(0.071)	(0.069)	(0.070)	(0.070)
overnight returns		-6.979***	-2.253	-2.062	-2.269
		(2.409)	(2.852)	(2.825)	(2.838)
RGBI			-0.000	-0.001	-0.001
			(0.002)	(0.002)	(0.002)
log return RTSI			-0.905	-0.673	-0.685
			(0.990)	(0.983)	(0.991)
high-low			-1.911***	-1.871***	-1.872***
			(0.290)	(0.292)	(0.290)
growth rate of dividend price ratio			0.136	0.142	0.328
			(0.165)	(0.169)	(0.232)
growth rate of earning price ratio			0.117	0.158	0.280
			(0.242)	(0.239)	(0.299)
log RV S&P				0.040***	0.045***
				(0.015)	(0.016)
log RV Brent				-0.036**	-0.043**
				(0.017)	(0.019)
growth rate of imports from USA					-0.001
					(0.045)
growth rate of exports to USA					-0.109
					(0.068)
growth rate of CPI					4.401
					(5.706)
growth rate of housing starts					-0.034
					(0.032)
growth rate of quarterly GDP					0.142
					(0.192)
Observations	1,255	1,255	1,255	1,255	1,255
RMSE	0.46	0.46	0.45	0.45	0.45
AIC	1641.59	1637.41	1600.68	1593.7	1598.42
LM test	0.97	0.97	0.87	0.95	0.95

Note: *p<0.1; **p<0.05; ***p<0.01

Table 15: HAR-RV estimation results, GAZPROM

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.400***	0.382***	0.267***	0.251***	0.251***
	(0.040)	(0.038)	(0.040)	(0.039)	(0.039)
log RV weekly	0.371***	0.374***	0.364***	0.350***	0.349***
	(0.057)	(0.057)	(0.055)	(0.052)	(0.052)
log RV monthly	0.079*	0.077*	0.085**	0.029	0.022
	(0.041)	(0.040)	(0.041)	(0.049)	(0.049)
is after weekend		0.008	-0.001	0.009	0.009
		(0.033)	(0.033)	(0.034)	(0.034)
is friday		-0.065	-0.031	-0.030	-0.031
		(0.042)	(0.049)	(0.049)	(0.048)
is after holiday		0.106	0.112	0.109	0.101
		(0.126)	(0.126)	(0.123)	(0.125)
overnight returns		-6.795**	-0.725	0.361	0.269
		(2.975)	(3.320)	(3.171)	(3.162)
RGBI			0.003	0.002	0.003
			(0.002)	(0.002)	(0.002)
log return RTSI			-1.669*	-1.421	-1.478*
			(0.894)	(0.881)	(0.883)
high-low			-1.648***	-1.632***	-1.586***
			(0.291)	(0.286)	(0.282)
growth rate of dividend price ratio			0.062	0.022	-0.020
			(0.151)	(0.152)	(0.158)
growth rate of earning price ratio			-0.003	-0.030	-0.023
			(0.069)	(0.071)	(0.071)
log RV S&P				0.048***	0.047***
				(0.017)	(0.017)
log RV Brent				0.009	0.005
				(0.020)	(0.021)
growth rate of imports from USA					-0.034
					(0.052)
growth rate of exports to USA					-0.017
					(0.062)
growth rate of CPI					4.497
					(5.634)
growth rate of housing starts					-0.051*
					(0.028)
growth rate of quarterly GDP					-0.020
					(0.166)
Observations	1,255	1,255	1,255	1,255	1,255
RMSE	0.51	0.5	0.5	0.49	0.49
AIC	1863.85	1858.09	1835.48	1825.0	1830.11
LM test	0.1	0.04	0.07	0.15	0.14

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16: HAR-RV estimation results, LUKOIL

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.432***	0.418***	0.302***	0.293***	0.286***
	(0.042)	(0.040)	(0.042)	(0.042)	(0.041)
log RV weekly	0.348***	0.342***	0.235***	0.207***	0.188***
	(0.068)	(0.068)	(0.058)	(0.057)	(0.059)
log RV monthly	0.115**	0.118**	0.218***	0.209***	0.225***
	(0.054)	(0.050)	(0.043)	(0.046)	(0.046)
is after weekend		-0.008	-0.009	0.001	-0.000
		(0.034)	(0.034)	(0.033)	(0.033)
is friday		-0.073*	-0.064	-0.064	-0.065
		(0.041)	(0.045)	(0.044)	(0.044)
is after holiday		-0.071	-0.090	-0.090	-0.096
		(0.064)	(0.062)	(0.061)	(0.060)
overnight returns		-8.096**	-0.537	0.754	0.685
		(3.203)	(3.397)	(3.335)	(3.271)
RGBI			-0.002	-0.003	-0.002
			(0.002)	(0.002)	(0.002)
log return RTSI			-2.167**	-1.990*	-1.987*
			(1.090)	(1.075)	(1.075)
high-low			-1.735***	-1.669***	-1.678***
			(0.282)	(0.278)	(0.275)
growth rate of dividend price ratio			0.605***	0.577***	0.635***
			(0.174)	(0.167)	(0.178)
growth rate of earning price ratio			0.041	0.021	0.015
			(0.060)	(0.061)	(0.062)
log RV S&P				0.043***	0.041***
				(0.015)	(0.016)
log RV Brent				0.008	0.003
				(0.020)	(0.021)
growth rate of imports from USA					0.015
					(0.047)
growth rate of exports to USA					-0.113*
					(0.066)
growth rate of CPI					0.660
					(6.302)
growth rate of housing starts					-0.067*
					(0.035)
growth rate of quarterly GDP					-0.115
					(0.155)
Observations	1,255	1,255	1,255	1,255	1,255
RMSE	0.49	0.49	0.48	0.47	0.47
AIC	1789.72	1784.22	1730.78	1721.86	1720.8
LM test	0.06	0.03	0.38	0.68	0.75

Note: *p<0.1; **p<0.05; ***p<0.01

Table 17: HAR-RV estimation results, NOVATEK

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.297***	0.290***	0.196***	0.188***	0.188***
	(0.044)	(0.044)	(0.044)	(0.043)	(0.043)
log RV weekly	0.410***	0.400***	0.362***	0.341***	0.337***
	(0.073)	(0.073)	(0.069)	(0.066)	(0.065)
log RV monthly	0.182***	0.184***	0.214***	0.173***	0.170***
	(0.044)	(0.043)	(0.045)	(0.047)	(0.050)
is after weekend		-0.019	-0.024	-0.012	-0.014
		(0.036)	(0.035)	(0.035)	(0.035)
is friday		-0.025	-0.003	-0.003	-0.004
		(0.040)	(0.046)	(0.046)	(0.045)
is after holiday		-0.120	-0.099	-0.101	-0.104
		(0.093)	(0.093)	(0.092)	(0.090)
overnight returns		-7.322*	-2.991	-1.124	-1.142
		(4.248)	(4.534)	(4.309)	(4.313)
RGBI			-0.000	-0.001	-0.001
			(0.002)	(0.002)	(0.002)
log return RTSI			-3.059**	-2.819**	-2.809**
			(1.205)	(1.168)	(1.178)
high-low			-1.543***	-1.486***	-1.427***
			(0.297)	(0.291)	(0.297)
growth rate of dividend price ratio			0.143	0.082	0.087
			(0.112)	(0.117)	(0.124)
growth rate of earning price ratio			-0.047	-0.035	-0.030
			(0.100)	(0.100)	(0.097)
log RV S&P				0.053***	0.051***
				(0.016)	(0.017)
log RV Brent				0.004	0.001
				(0.017)	(0.018)
growth rate of imports from USA					0.001
					(0.052)
growth rate of exports to USA					-0.116*
					(0.065)
growth rate of CPI					-1.007
					(6.963)
growth rate of housing starts					-0.002
					(0.036)
growth rate of quarterly GDP					-0.220
					(0.168)
Observations	1,255	1,255	1,255	1,255	1,255
RMSE	0.51	0.51	0.5	0.5	0.5
AIC	1896.76	1898.79	1873.42	1861.16	1866.26
LM test	0.0	0.0	0.0	0.11	0.14

Note: *p<0.1; **p<0.05; ***p<0.01

Table 18: HAR-RV estimation results, ROSNEFT

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.353***	0.331***	0.224***	0.222***	0.222***
	(0.037)	(0.035)	(0.039)	(0.038)	(0.038)
log RV weekly	0.454***	0.448***	0.429***	0.418***	0.414***
	(0.060)	(0.059)	(0.059)	(0.056)	(0.056)
log RV monthly	0.089*	0.094**	0.109**	0.106**	0.096*
	(0.051)	(0.047)	(0.049)	(0.051)	(0.051)
is after weekend		0.030	0.035	0.040	0.039
		(0.035)	(0.036)	(0.036)	(0.035)
is friday		-0.037	-0.008	-0.009	-0.009
		(0.038)	(0.042)	(0.042)	(0.042)
is after holiday		-0.049	-0.072	-0.072	-0.079
		(0.078)	(0.076)	(0.076)	(0.075)
overnight returns		-10.928***	-5.891*	-4.948	-4.956
		(2.648)	(3.324)	(3.407)	(3.409)
RGBI			-0.000	-0.001	-0.001
			(0.002)	(0.002)	(0.002)
log return RTSI			-1.347	-1.262	-1.325
			(1.201)	(1.194)	(1.200)
high-low			-1.592***	-1.561***	-1.534***
			(0.270)	(0.268)	(0.270)
growth rate of dividend price ratio			0.019	0.004	-0.019
			(0.123)	(0.124)	(0.121)
growth rate of earning price ratio			-0.019	-0.018	-0.006
			(0.090)	(0.088)	(0.091)
log RV S&P				0.027**	0.024*
				(0.013)	(0.014)
log RV Brent				-0.007	-0.009
				(0.016)	(0.017)
growth rate of imports from USA					-0.014
					(0.045)
growth rate of exports to USA					-0.029
					(0.059)
growth rate of CPI					2.731
					(5.554)
growth rate of housing starts					-0.011
					(0.032)
growth rate of quarterly GDP					-0.183
					(0.168)
Observations	1,255	1,255	1,255	1,255	1,255
RMSE	0.48	0.47	0.47	0.47	0.47
AIC	1714.29	1702.82	1679.21	1678.46	1685.35
LM test	0.01	0.0	0.01	0.03	0.04

Note: *p<0.1; **p<0.05; ***p<0.01

Table 19: HAR-RV estimation results, NORNICKEL

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.316***	0.312***	0.183***	0.166***	0.166***
	(0.050)	(0.049)	(0.051)	(0.050)	(0.050)
log RV weekly	0.398***	0.397***	0.370***	0.349***	0.347***
	(0.066)	(0.065)	(0.063)	(0.059)	(0.059)
log RV monthly	0.138***	0.140***	0.154***	0.126**	0.129**
	(0.053)	(0.052)	(0.052)	(0.054)	(0.053)
is after weekend		-0.046	-0.058*	-0.046	-0.046
		(0.035)	(0.034)	(0.034)	(0.034)
is friday		-0.037	-0.036	-0.034	-0.034
		(0.044)	(0.046)	(0.045)	(0.045)
is after holiday		-0.036	-0.032	-0.039	-0.039
		(0.076)	(0.072)	(0.074)	(0.074)
overnight returns		-3.807	2.378	2.665	2.476
		(3.463)	(3.353)	(3.169)	(3.158)
RGBI			0.002	0.001	0.001
			(0.002)	(0.002)	(0.002)
log return RTSI			-1.007	-0.701	-0.677
			(1.327)	(1.289)	(1.281)
high-low			-2.110***	-2.119***	-2.096***
			(0.323)	(0.319)	(0.320)
growth rate of dividend price ratio			0.141	0.135	0.144
			(0.143)	(0.140)	(0.143)
growth rate of earning price ratio			0.103	0.124	0.113
			(0.099)	(0.099)	(0.099)
log RV S&P				0.066***	0.064***
				(0.015)	(0.016)
log RV Brent				-0.020	-0.020
				(0.018)	(0.020)
growth rate of imports from USA					0.019
					(0.055)
growth rate of exports to USA					-0.045
					(0.071)
growth rate of CPI					-1.961
					(6.680)
growth rate of housing starts					-0.026
					(0.035)
growth rate of quarterly GDP					-0.124
					(0.189)
Observations	1,785	1,785	1,785	1,785	1,785
RMSE	0.51	0.51	0.49	0.49	0.49
AIC	2643.13	2645.91	2580.13	2548.1	2555.23
LM test	0.0	0.0	0.0	0.0	0.0

Note: *p<0.1; **p<0.05; ***p<0.01

Table 20: HAR-RV estimation results, POLYMETAL

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.367***	0.360***	0.264***	0.249***	0.248***
	(0.038)	(0.039)	(0.047)	(0.047)	(0.046)
log RV weekly	0.340***	0.338***	0.285***	0.269***	0.252***
	(0.067)	(0.066)	(0.061)	(0.060)	(0.061)
log RV monthly	0.179***	0.181***	0.240***	0.226***	0.252***
	(0.055)	(0.053)	(0.051)	(0.053)	(0.055)
is after weekend		-0.007	-0.003	0.009	0.008
		(0.040)	(0.041)	(0.041)	(0.040)
is friday		-0.054	-0.023	-0.023	-0.023
		(0.042)	(0.046)	(0.045)	(0.045)
is after holiday		-0.001	-0.017	-0.029	-0.016
		(0.101)	(0.099)	(0.101)	(0.100)
overnight returns		-8.337	-6.258	-5.105	-4.660
		(6.014)	(5.688)	(5.622)	(5.690)
RGBI			-0.002	-0.003*	-0.004**
			(0.002)	(0.002)	(0.002)
log return RTSI			-2.086**	-1.727*	-1.583
			(1.044)	(1.014)	(1.005)
high-low			-1.380***	-1.398***	-1.379***
			(0.338)	(0.339)	(0.339)
growth rate of dividend price ratio			0.234	0.164	0.238
			(0.226)	(0.223)	(0.221)
growth rate of earning price ratio			0.001	-0.003	0.006
			(0.222)	(0.226)	(0.240)
log RV S&P				0.049***	0.054***
				(0.017)	(0.017)
log RV Brent				0.007	0.003
				(0.024)	(0.024)
growth rate of imports from USA					0.072
					(0.055)
growth rate of exports to USA					-0.130*
					(0.075)
growth rate of CPI					-5.176
					(6.975)
growth rate of housing starts					0.002
					(0.031)
growth rate of quarterly GDP					0.095
					(0.181)
Observations	1,255	1,255	1,255	1,255	1,255
RMSE	0.54	0.54	0.53	0.53	0.53
AIC	2026.9	2030.03	2004.52	1992.57	1996.41
LM test	0.65	0.72	0.9	0.94	0.93

Note: *p<0.1; **p<0.05; ***p<0.01

Table 21: HAR-RV estimation results, POLYUS

	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.356***	0.369***	0.208***	0.200***	0.199***
	(0.038)	(0.039)	(0.059)	(0.058)	(0.058)
log RV weekly	0.333***	0.327***	0.299***	0.286***	0.262***
	(0.058)	(0.057)	(0.065)	(0.062)	(0.061)
log RV monthly	0.205***	0.203***	0.171***	0.122**	0.156**
	(0.044)	(0.045)	(0.059)	(0.062)	(0.070)
is after weekend		-0.019	-0.106**	-0.088*	-0.089*
		(0.054)	(0.054)	(0.053)	(0.053)
is friday		0.055	0.014	0.008	0.007
		(0.056)	(0.074)	(0.073)	(0.073)
is after holiday		-0.057	-0.168*	-0.202**	-0.200**
		(0.138)	(0.096)	(0.099)	(0.097)
overnight returns		9.454*	7.686	7.529	7.739
		(5.649)	(5.124)	(5.150)	(5.235)
RGBI			0.005	0.006*	0.006
			(0.003)	(0.003)	(0.004)
log return RTSI			-0.161	0.632	0.743
			(1.575)	(1.481)	(1.469)
high-low			-2.292***	-2.190***	-2.227***
			(0.479)	(0.470)	(0.475)
growth rate of dividend price ratio			-0.709*	-0.728*	-0.799*
			(0.424)	(0.420)	(0.484)
growth rate of earning price ratio			-1.005**	-0.946*	-1.103*
			(0.508)	(0.499)	(0.574)
log RV S&P				0.082***	0.078***
				(0.021)	(0.023)
log RV Brent				-0.024	-0.027
				(0.030)	(0.034)
growth rate of imports from USA					0.117
					(0.106)
growth rate of exports to USA					-0.149
					(0.109)
growth rate of CPI					-5.719
					(11.094)
growth rate of housing starts					0.035
					(0.049)
growth rate of quarterly GDP					-0.167
					(0.302)
Observations	1,256	1,256	838	838	838
RMSE	0.74	0.73	0.64	0.64	0.64
AIC	2800.92	2804.32	1668.02	1655.67	1661.63
LM test	0.01	0.01	0.05	0.07	0.1

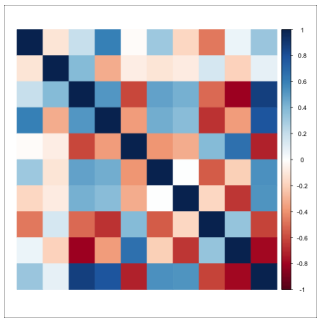
Note: *p<0.1; **p<0.05; ***p<0.01

Table 22: HAR-RV estimation results, MAGNIT

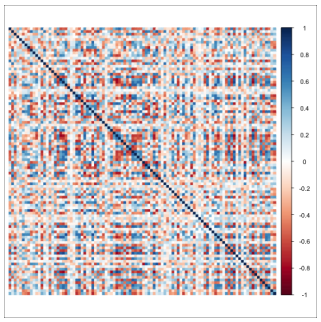
	Dependent variable: log RV_{t+1}				
	Basic	Overnight and calendar	Financial	Spillover	Macroeconomic
log RV	0.369***	0.356***	0.196***	0.192***	0.190***
	(0.043)	(0.044)	(0.044)	(0.045)	(0.045)
log RV weekly	0.336***	0.341***	0.310***	0.294***	0.276***
	(0.059)	(0.059)	(0.055)	(0.056)	(0.055)
log RV monthly	0.129**	0.130**	0.148***	0.131**	0.131**
	(0.054)	(0.055)	(0.054)	(0.057)	(0.054)
is after weekend		-0.044	-0.038	-0.029	-0.030
		(0.039)	(0.039)	(0.039)	(0.039)
is friday		-0.024	-0.028	-0.029	-0.029
		(0.043)	(0.048)	(0.047)	(0.047)
is after holiday		0.011	0.023	0.024	0.007
		(0.106)	(0.107)	(0.107)	(0.104)
overnight returns		-5.472*	-0.051	1.250	0.892
		(2.818)	(3.150)	(3.074)	(2.962)
RGBI			0.001	0.000	0.001
			(0.002)	(0.002)	(0.002)
log return RTSI			-0.124	0.048	-0.044
			(1.181)	(1.162)	(1.166)
high-low			-2.395***	-2.395***	-2.344***
			(0.314)	(0.312)	(0.311)
growth rate of dividend price ratio			-0.119	-0.085	-0.073
			(0.205)	(0.207)	(0.200)
growth rate of earning price ratio			-0.260	-0.191	-0.177
			(0.233)	(0.225)	(0.231)
log RV S&P				0.038**	0.033**
				(0.016)	(0.017)
log RV Brent				0.001	0.001
				(0.019)	(0.021)
growth rate of imports from USA					-0.008
					(0.054)
growth rate of exports to USA					-0.030
					(0.072)
growth rate of CPI					-0.269
					(6.573)
growth rate of housing starts					-0.068
					(0.043)
growth rate of quarterly GDP					-0.298
					(0.190)
Observations	1,255	1,255	1,255	1,255	1,255
RMSE	0.58	0.58	0.57	0.57	0.57
AIC	2209.2	2212.91	2174.45	2171.07	2171.79
LM test	0.0	0.0	0.01	0.05	0.08

Note: *p<0.1; **p<0.05; ***p<0.01

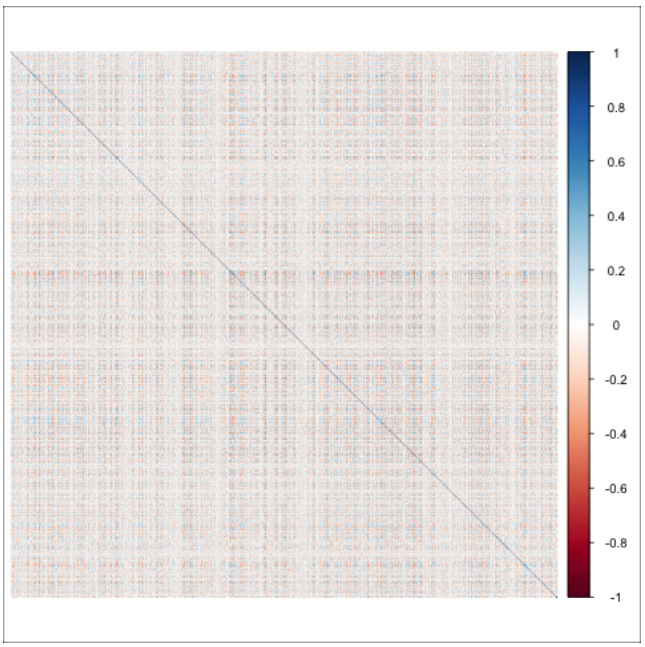
Figures



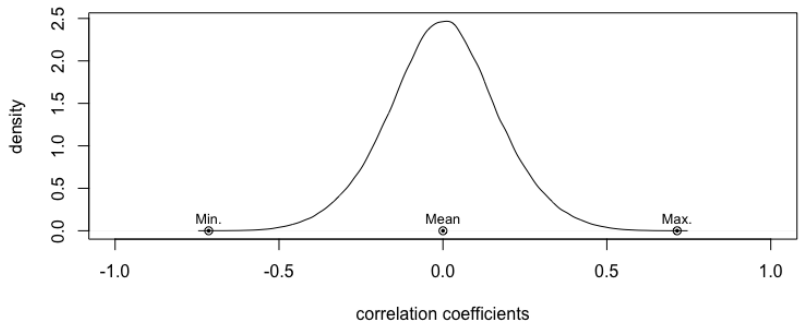
(a) $p = 10$



(b) $p = 100$

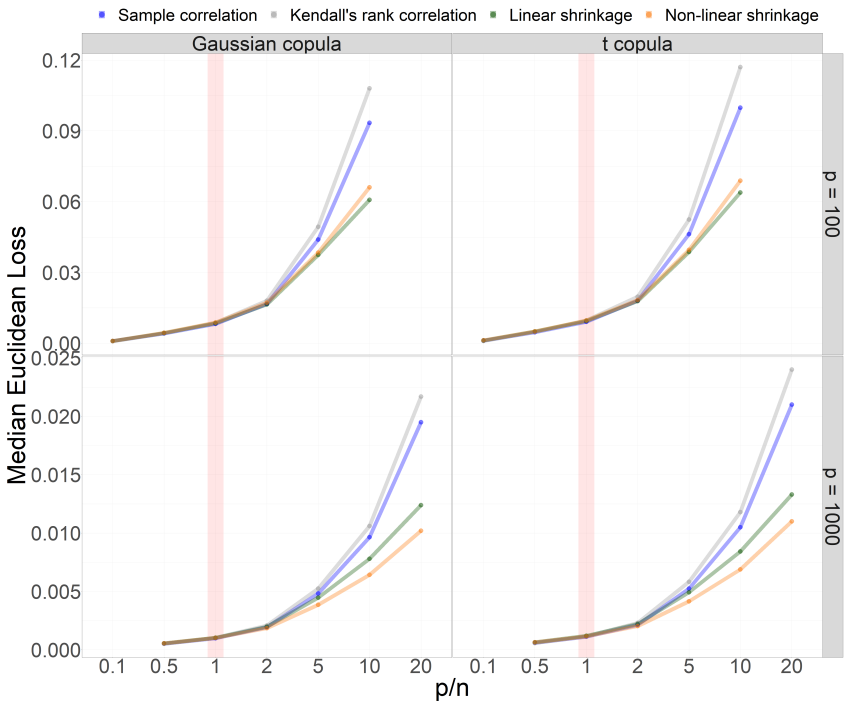


(c) $p = 1000$

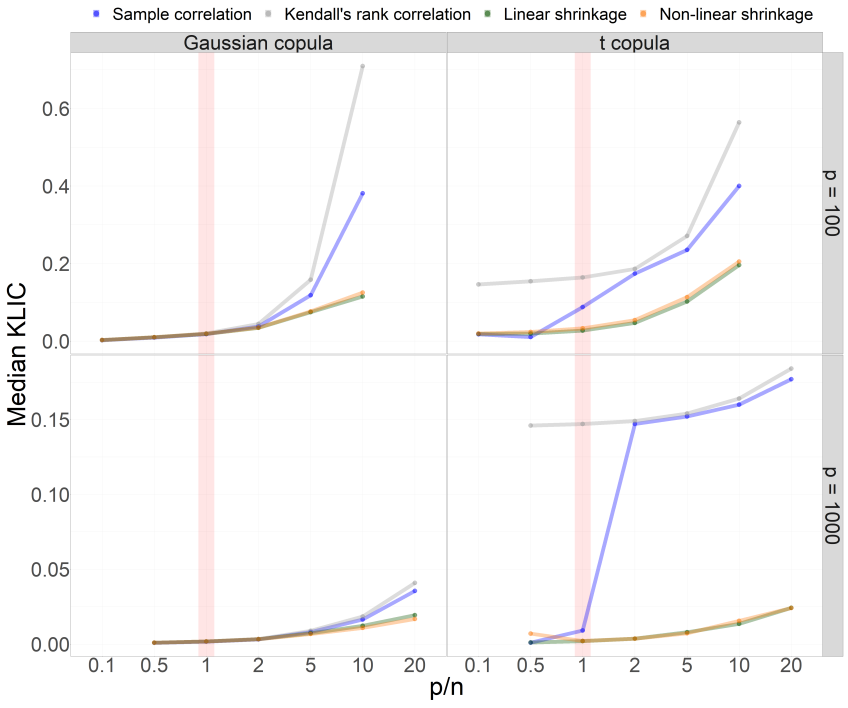


(d) distribution of correlation coefficients for the $p = 1000$ matrix

Figure 3: True correlation matrices P of arbitrary structure

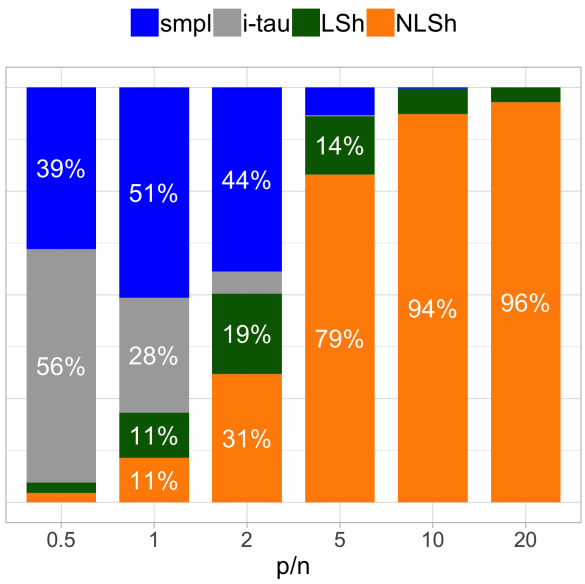


(a) Median Euclidean Loss

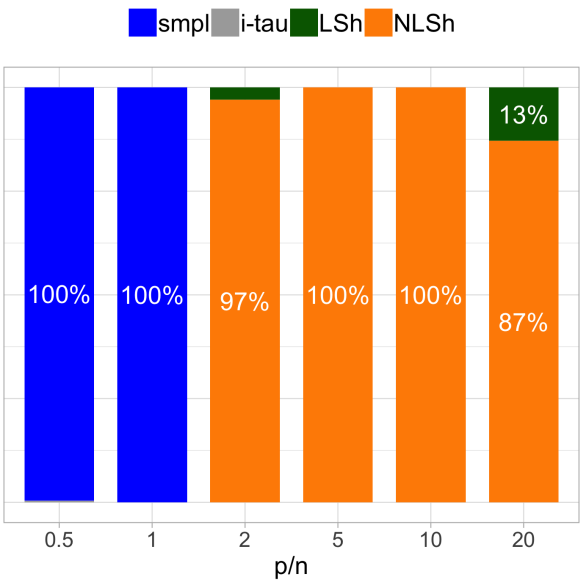


(b) Median KLIC (for t copula, the case of PMLE d.f.)

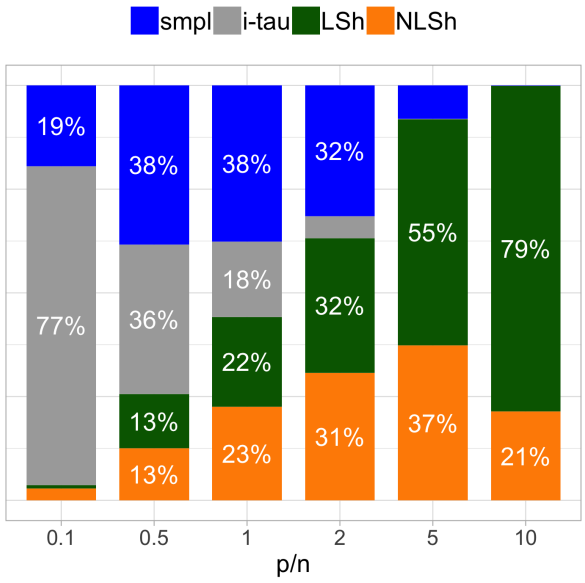
Figure 4: Median quality metrics of estimators in a selected slice of simulations



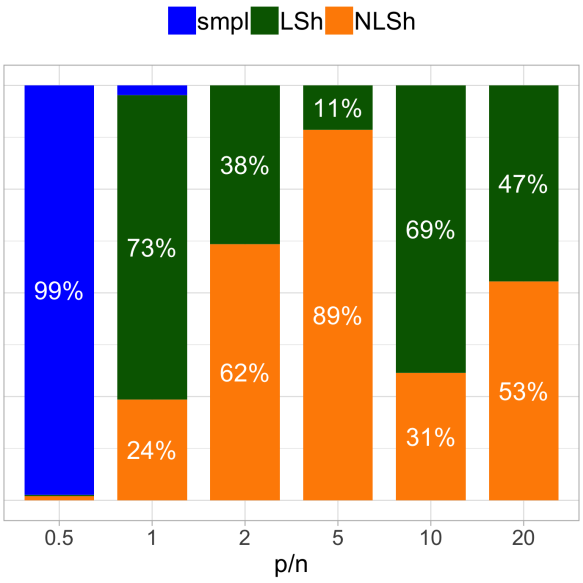
(a) $p = 1000$, Gaussian copula, arbitrary P



(b) $p = 1000$, t copula, identity P

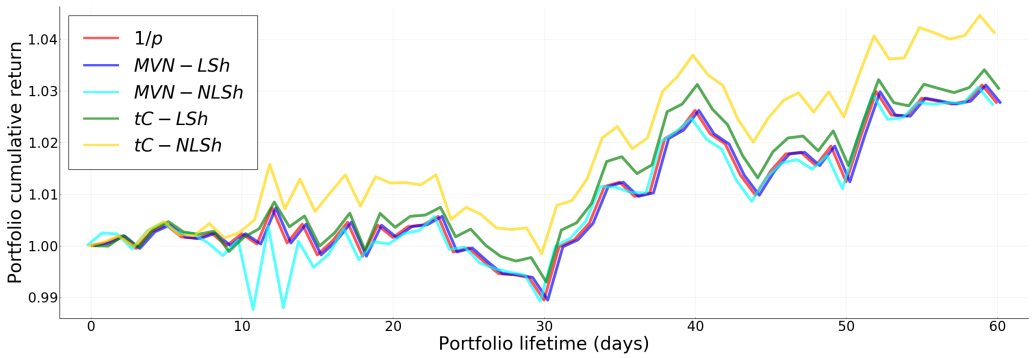


(c) $p = 100$, Gaussian copula, arbitrary P

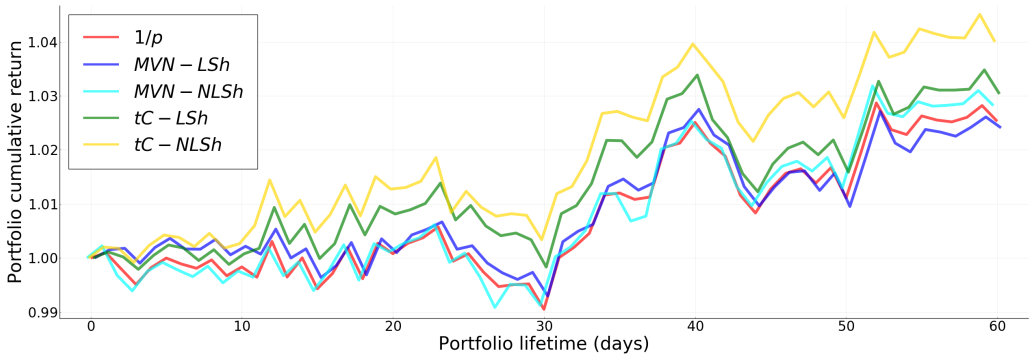


(d) $p = 1000$, t copula, arbitrary P

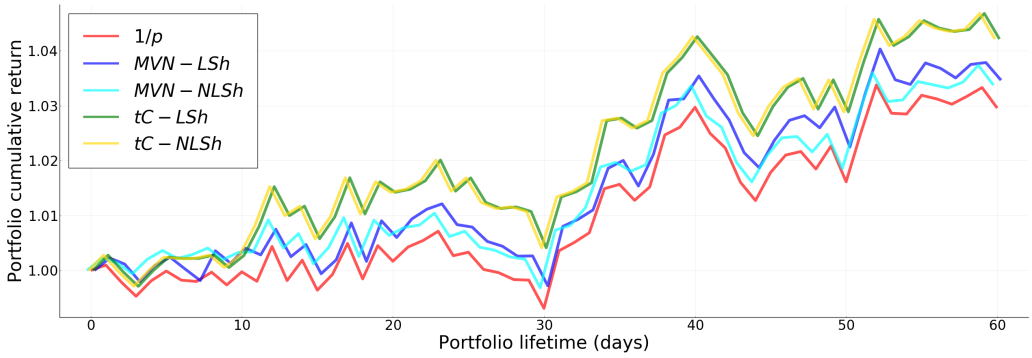
Figure 5: Shares of simulations in which each estimator returns the best KLIC



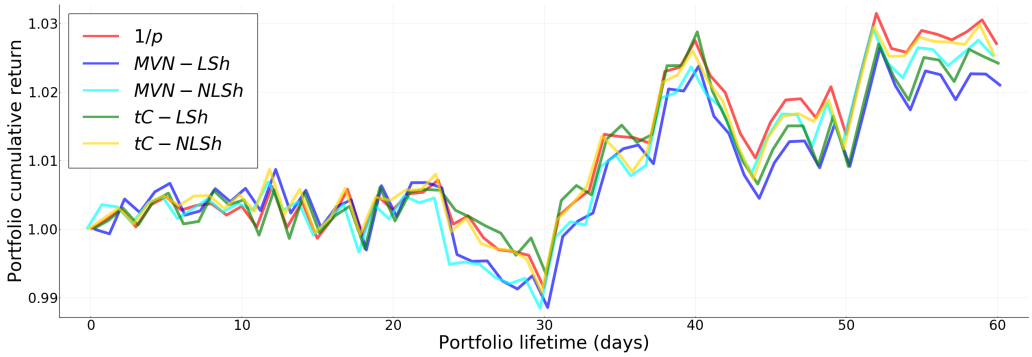
(a)



(b)



(c)



(d)

Figure 6: Examples of model-based portfolio cumulative return dynamics

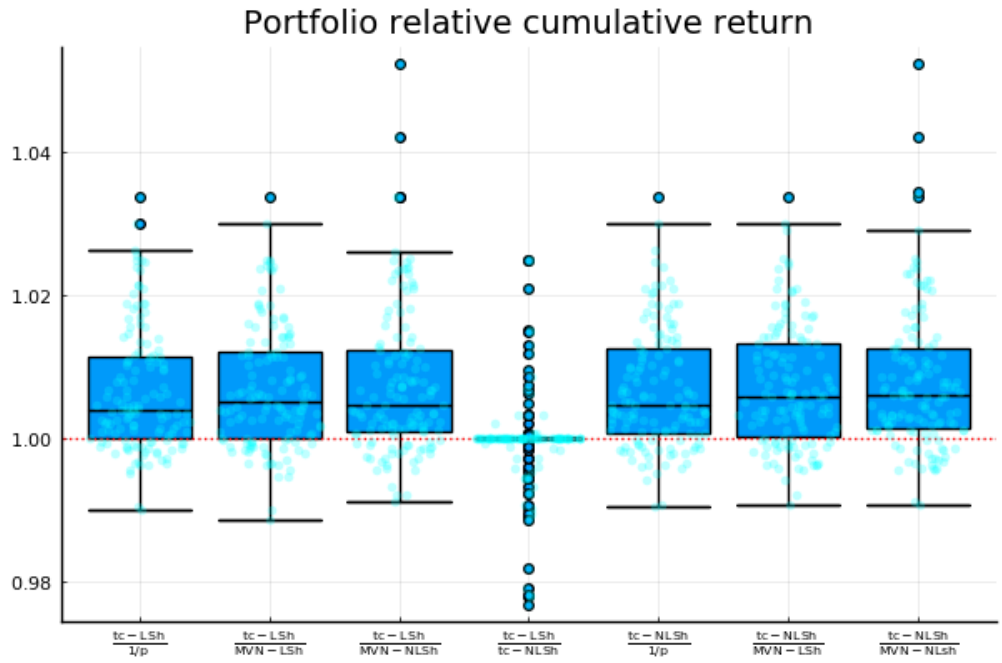


Figure 7: Relative returns of model-based portfolios across sets of assets

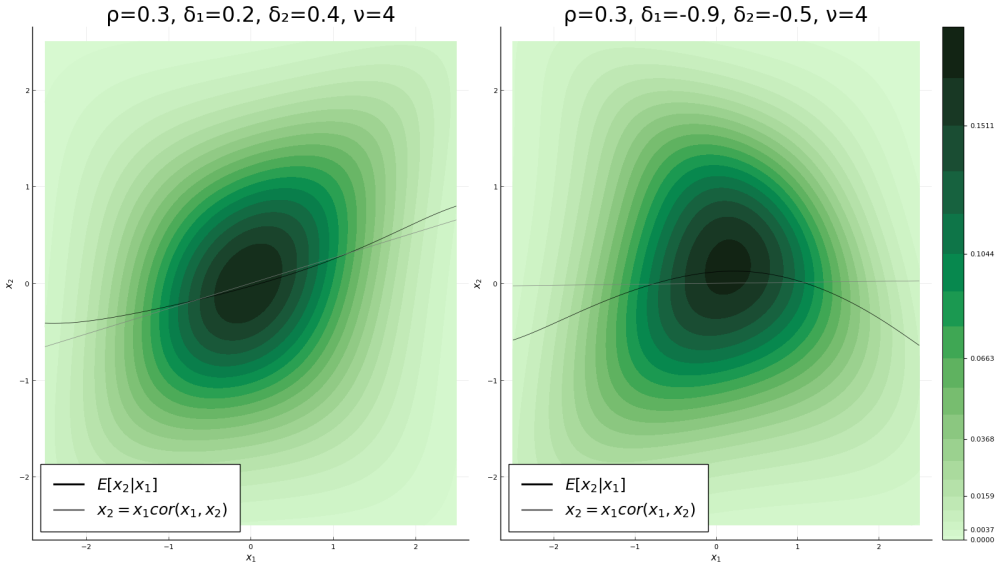


Figure 8: Density for standard normal marginals in bivariate skew-t copula

Figure 9: EUROSTOXX50 dynamics in the first half of 2022

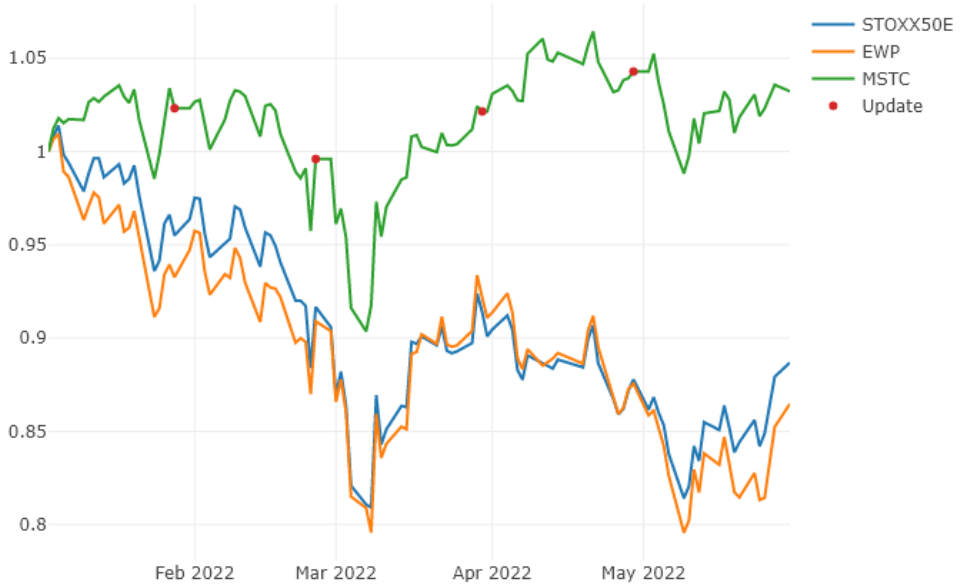


Figure 10: Skew-t copula-based portfolio value dynamics compared to the market

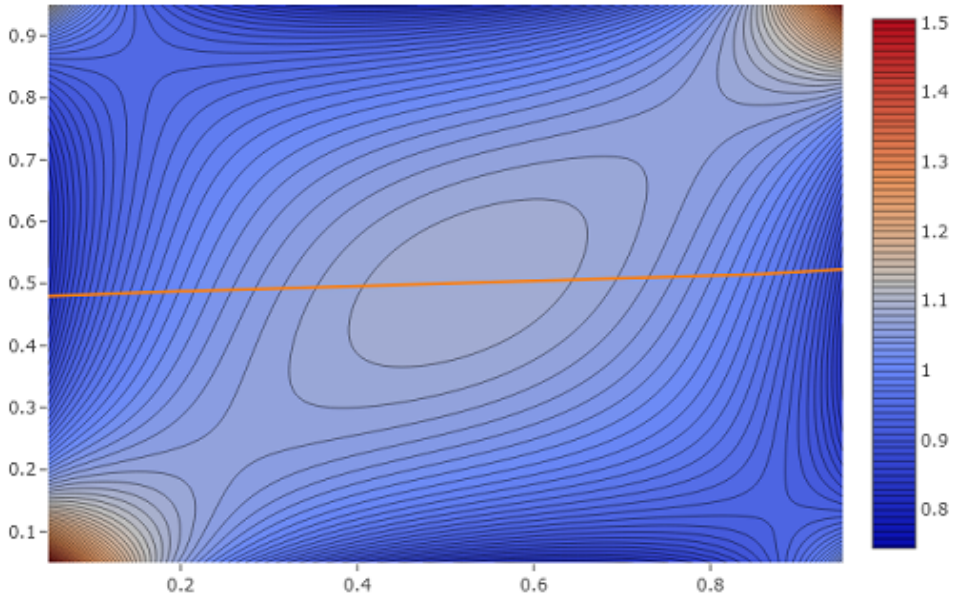


Figure 11: Skew-t copula density of Sanofi and Carrefour returns in Period 1

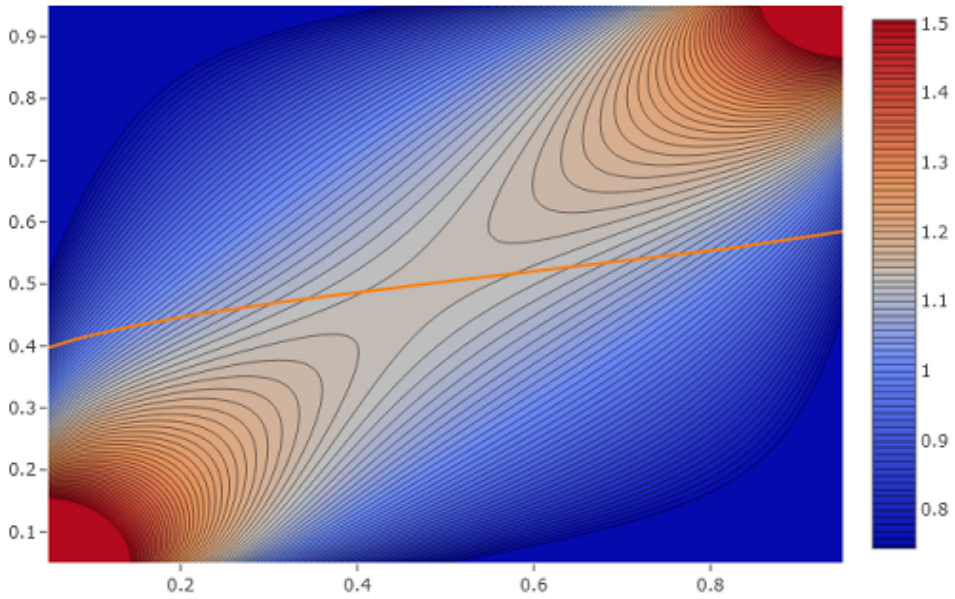


Figure 12: Skew-t copula density of Inditex and AB InBev returns in Period 2

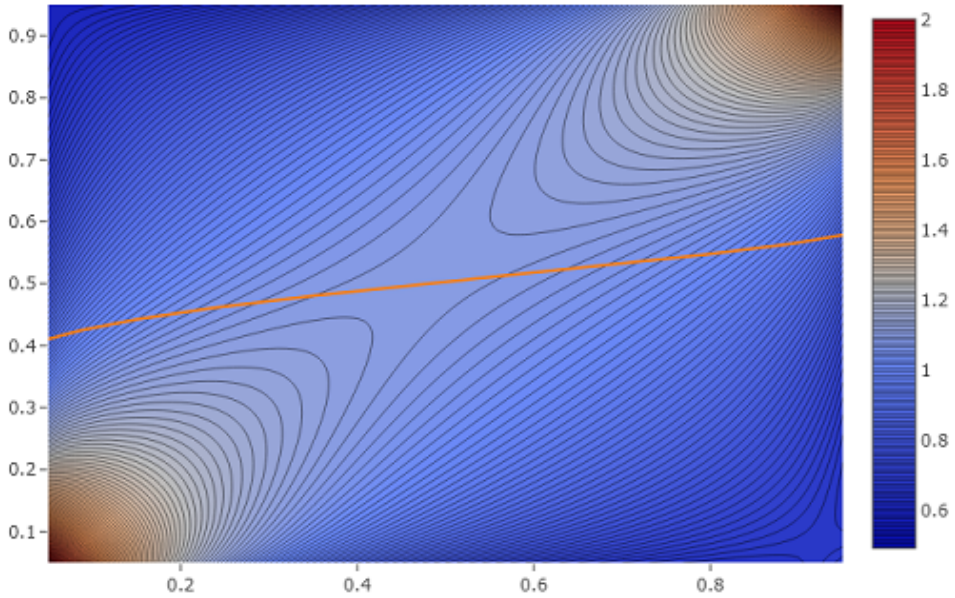


Figure 13: Skew-t copula density of Eni and Generali Group returns in Period 3

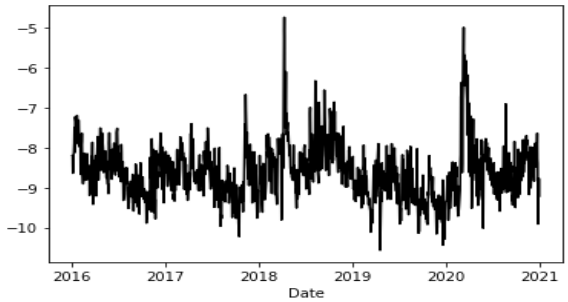


Figure 14: Dynamics of log-RV, SBERBANK

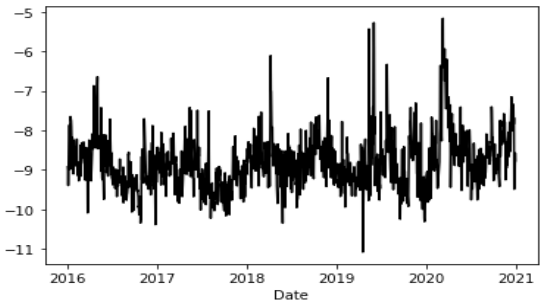


Figure 15: Dynamics of log-RV, GAZPROM

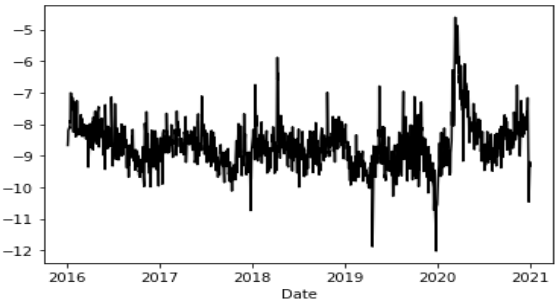


Figure 16: Dynamics of log-RV, LUKOIL

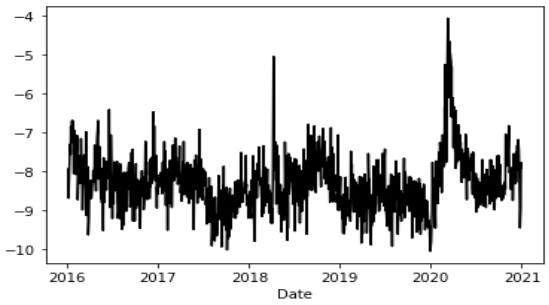


Figure 17: Dynamics of log-RV, NOVATEK

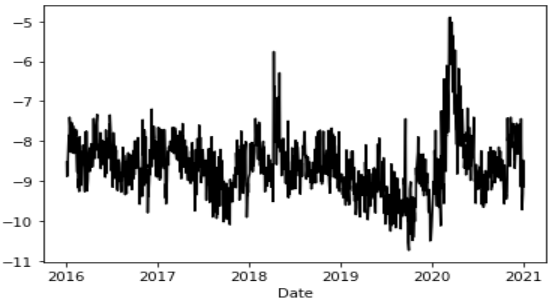


Figure 18: Dynamics of log-RV, ROSNEFT

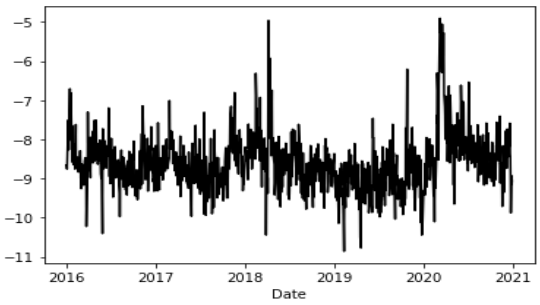


Figure 19: Dynamics of log-RV, NORNICKEL

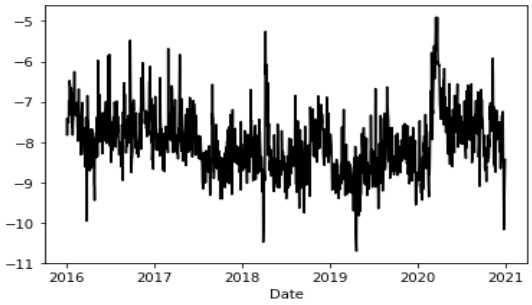


Figure 20: Dynamics of log-RV, POLYMETAL

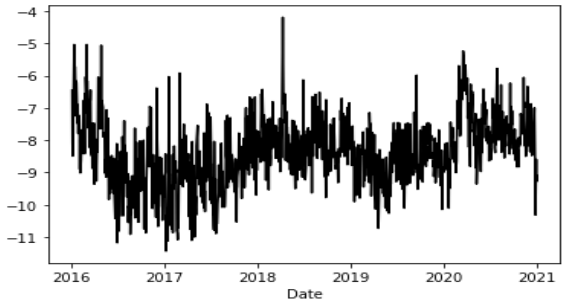


Figure 21: Dynamics of log-RV, POLYUS

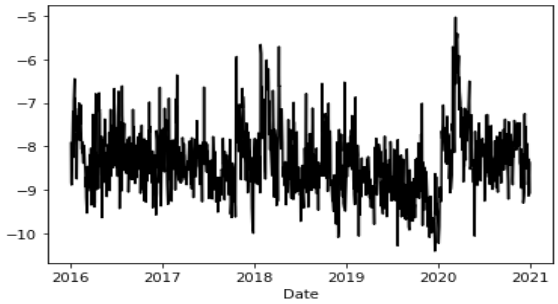


Figure 22: Dynamics of log-RV, MAGNIT

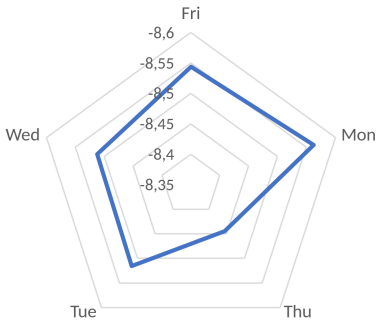


Figure 23: Average log-RV across weekdays, SBER

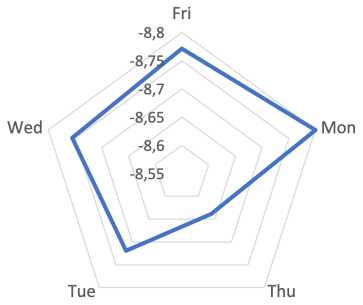


Figure 24: Average log-RV across weekdays, GAZPROM

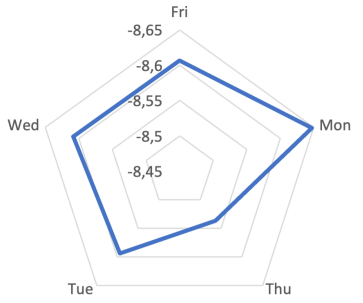


Figure 25: Average log-RV across weekdays, LUKOIL

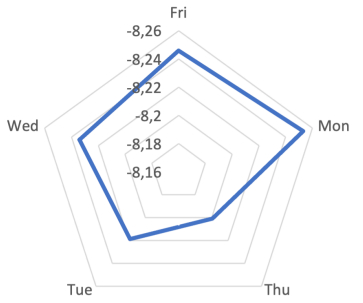


Figure 26: Average log-RV across weekdays, NOVATEK

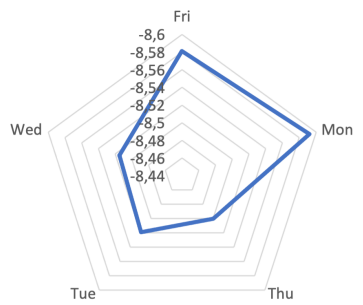


Figure 27: Average log-RV across weekdays, ROSNEFT

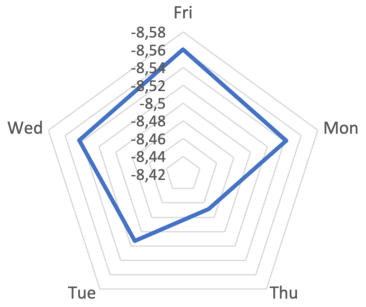


Figure 28: Average log-RV across weekdays, NORNICKEL

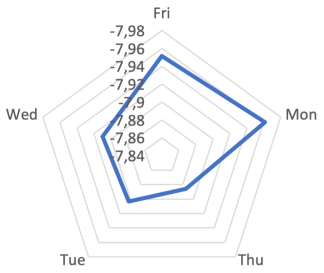


Figure 29: Average log-RV across weekdays, POLYMETAL

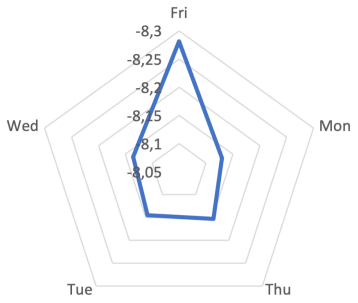


Figure 30: Average log-RV across weekdays, POLYUS

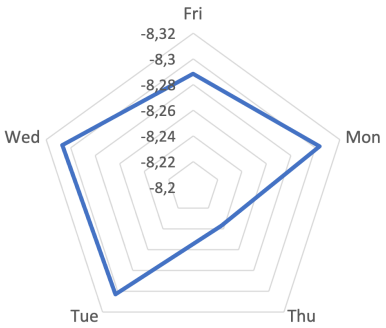


Figure 31: Average log-RV across weekdays, MAGNIT

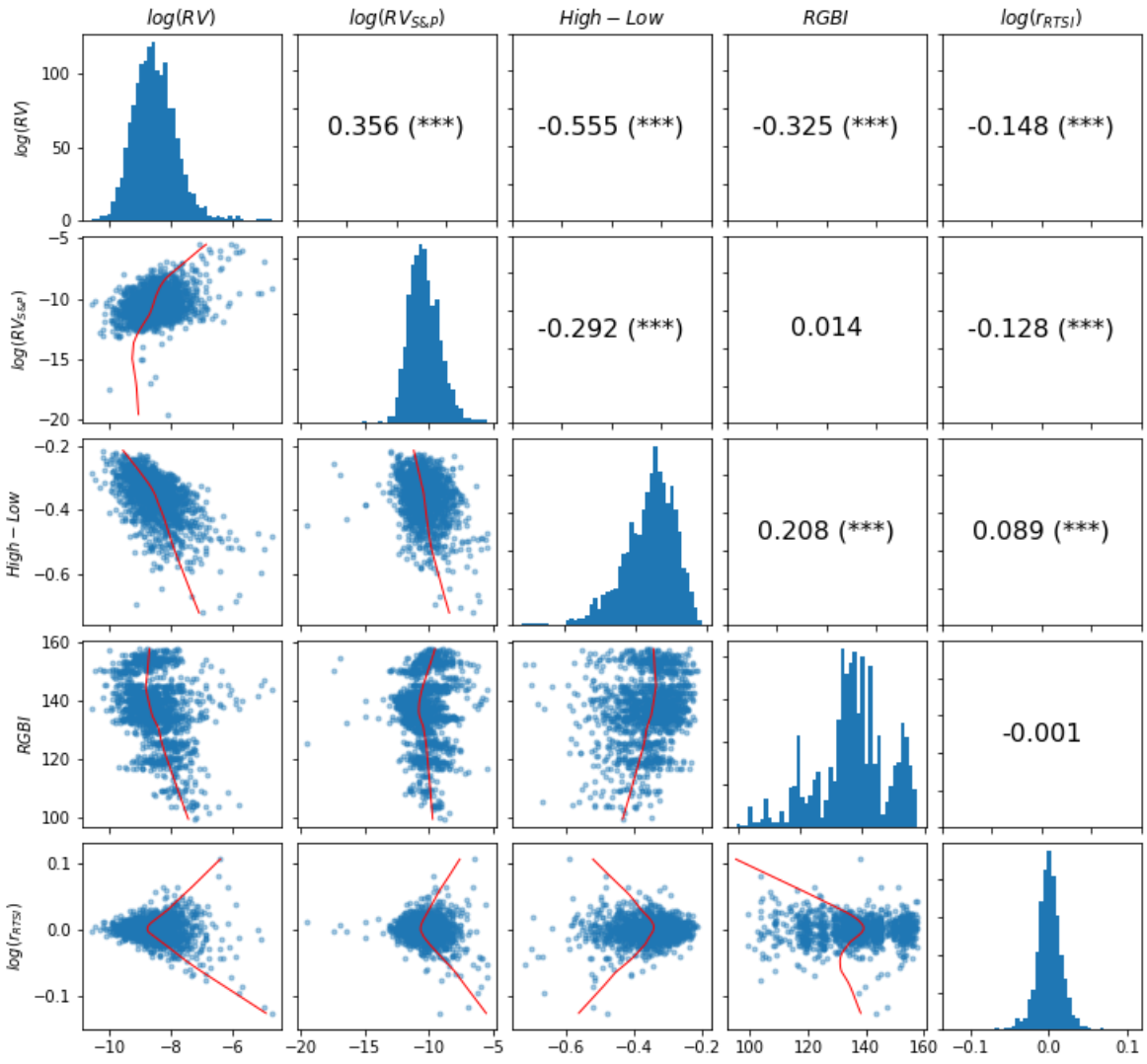


Figure 32: Correlations and dependencies between selected variables, SBERBANK

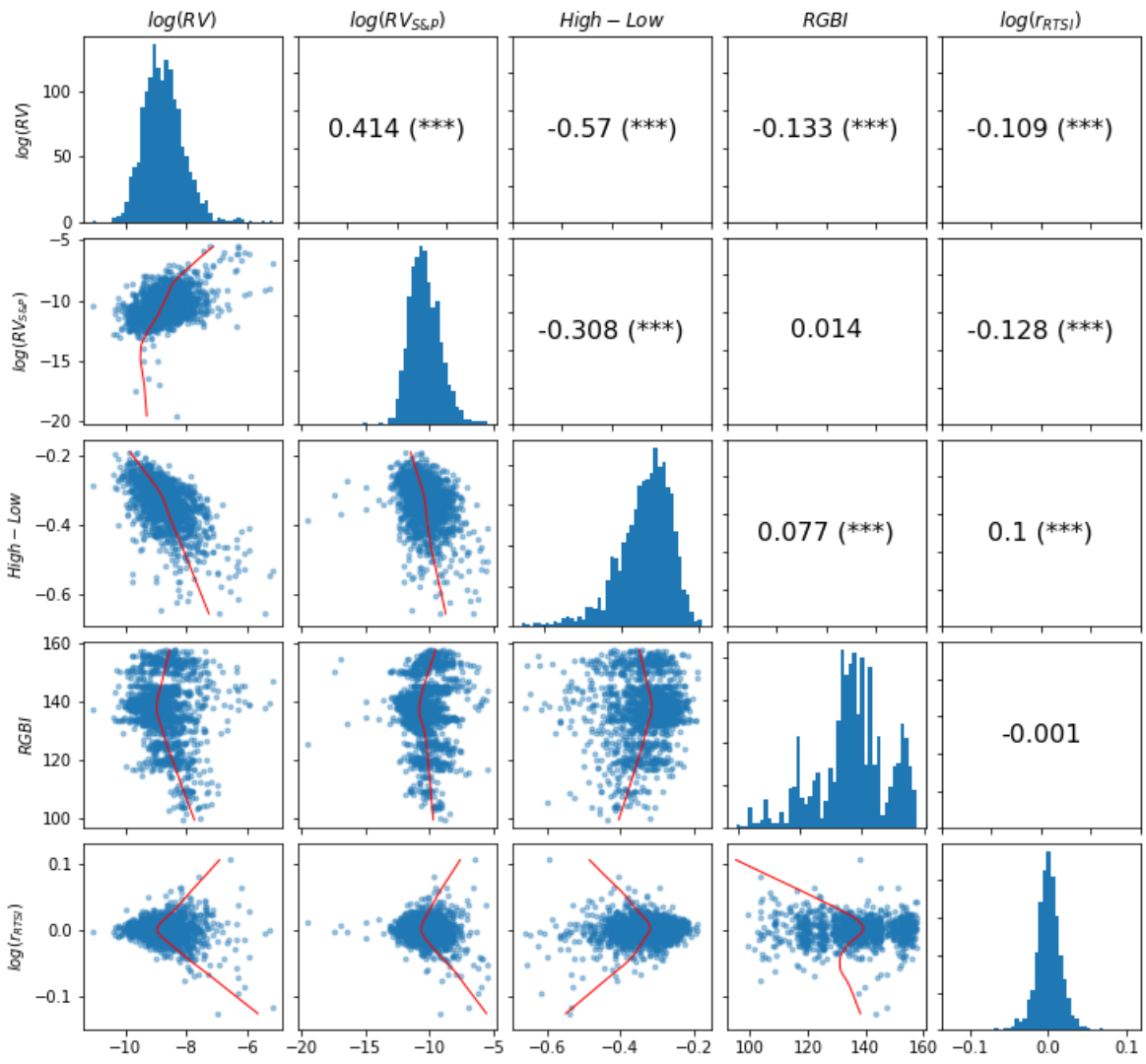


Figure 33: Correlations and dependencies between selected variables, GAZPROM

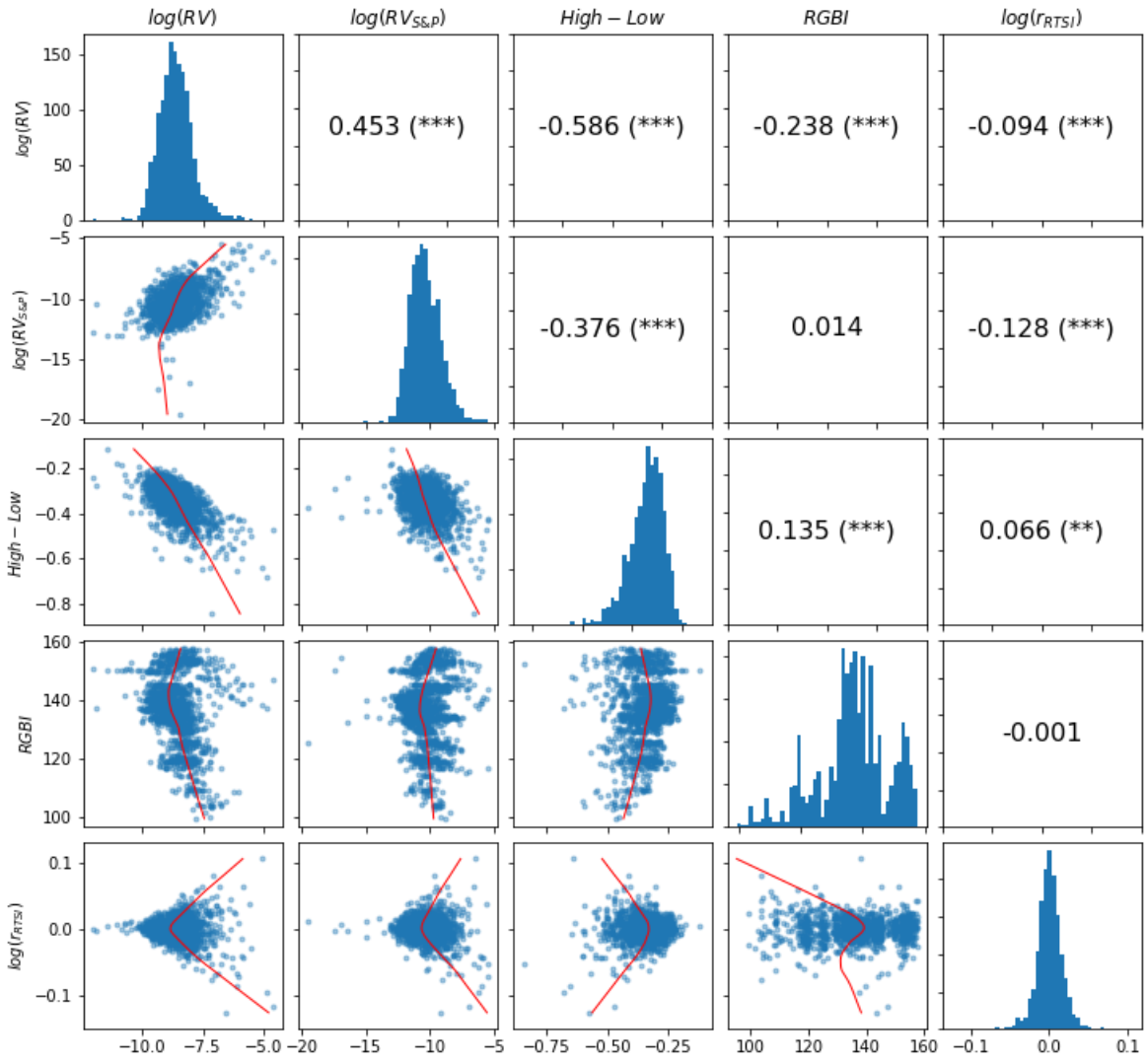


Figure 34: Correlations and dependencies between selected variables, LUKOIL

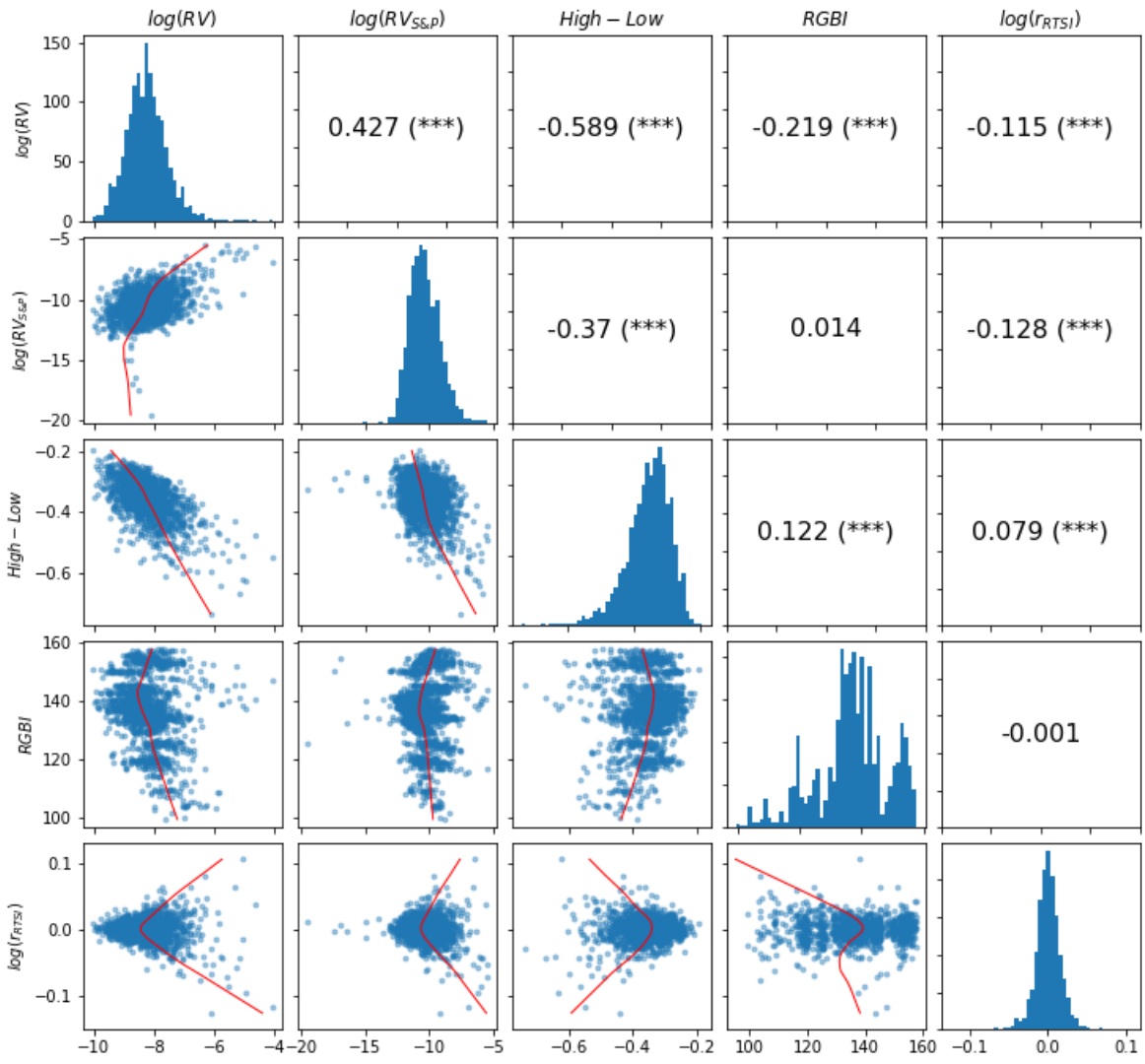


Figure 35: Correlations and dependencies between selected variables, NOVATEK

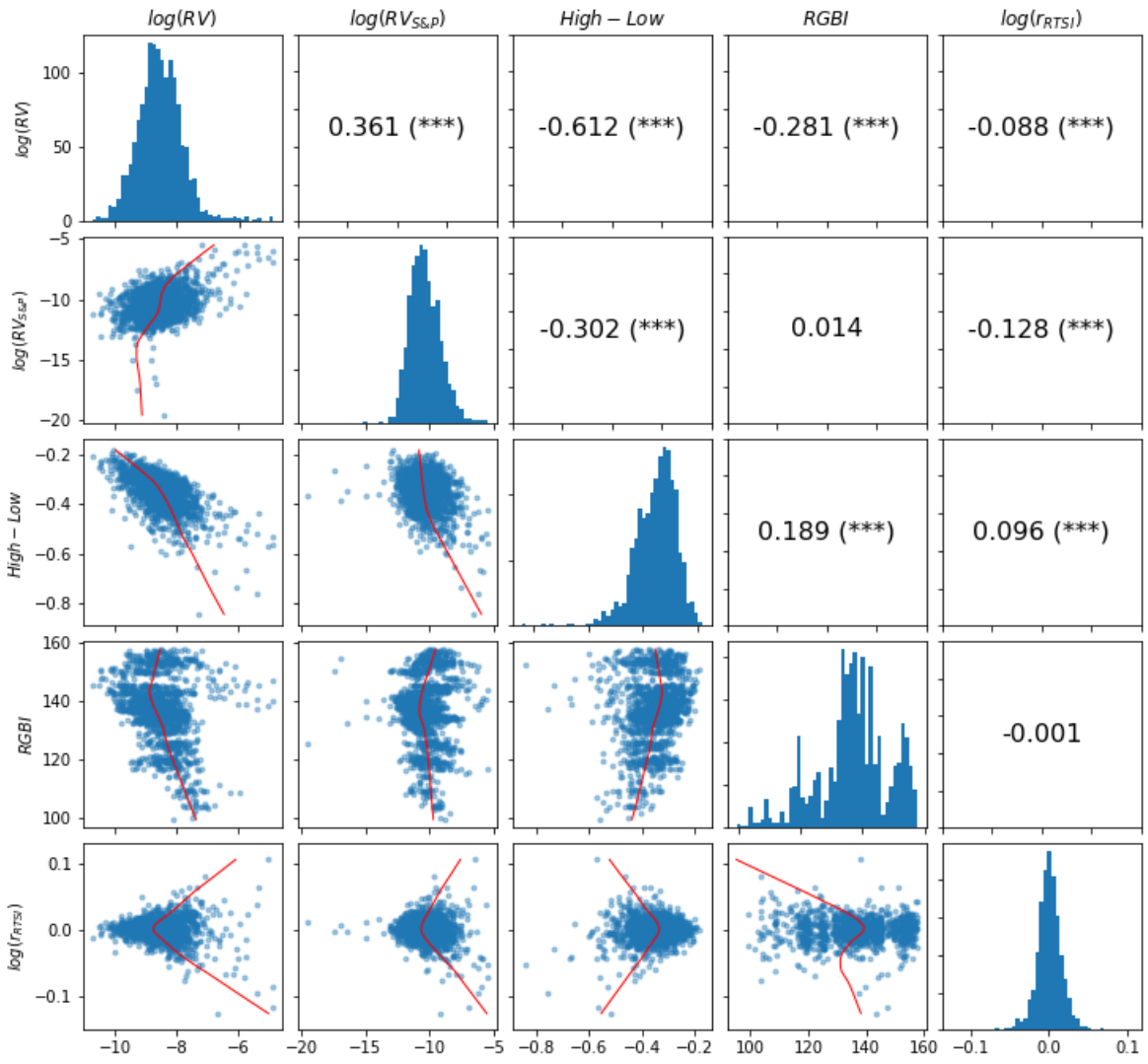


Figure 36: Correlations and dependencies between selected variables, ROSNEFT

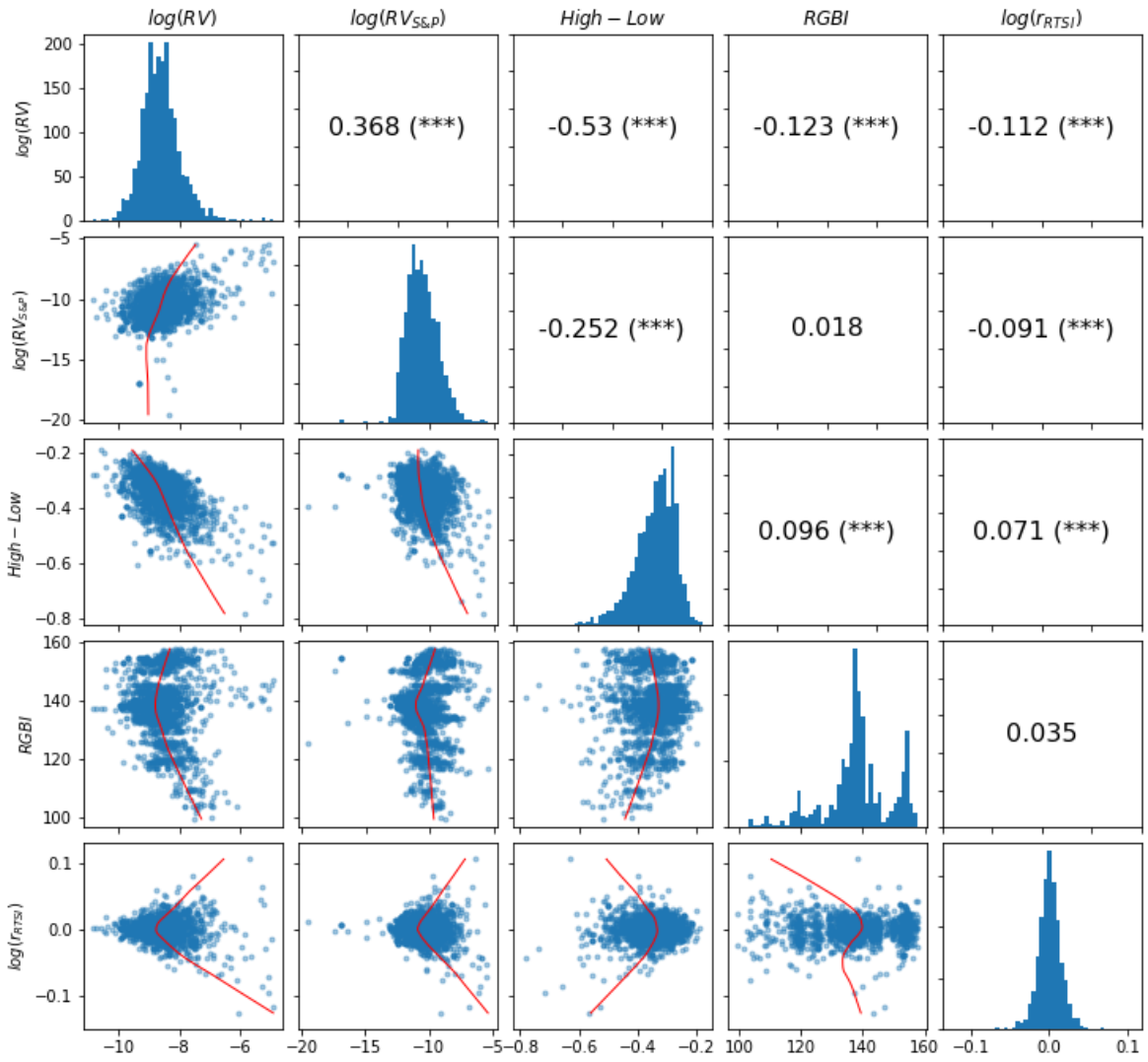


Figure 37: Correlations and dependencies between selected variables, NORNICKEL

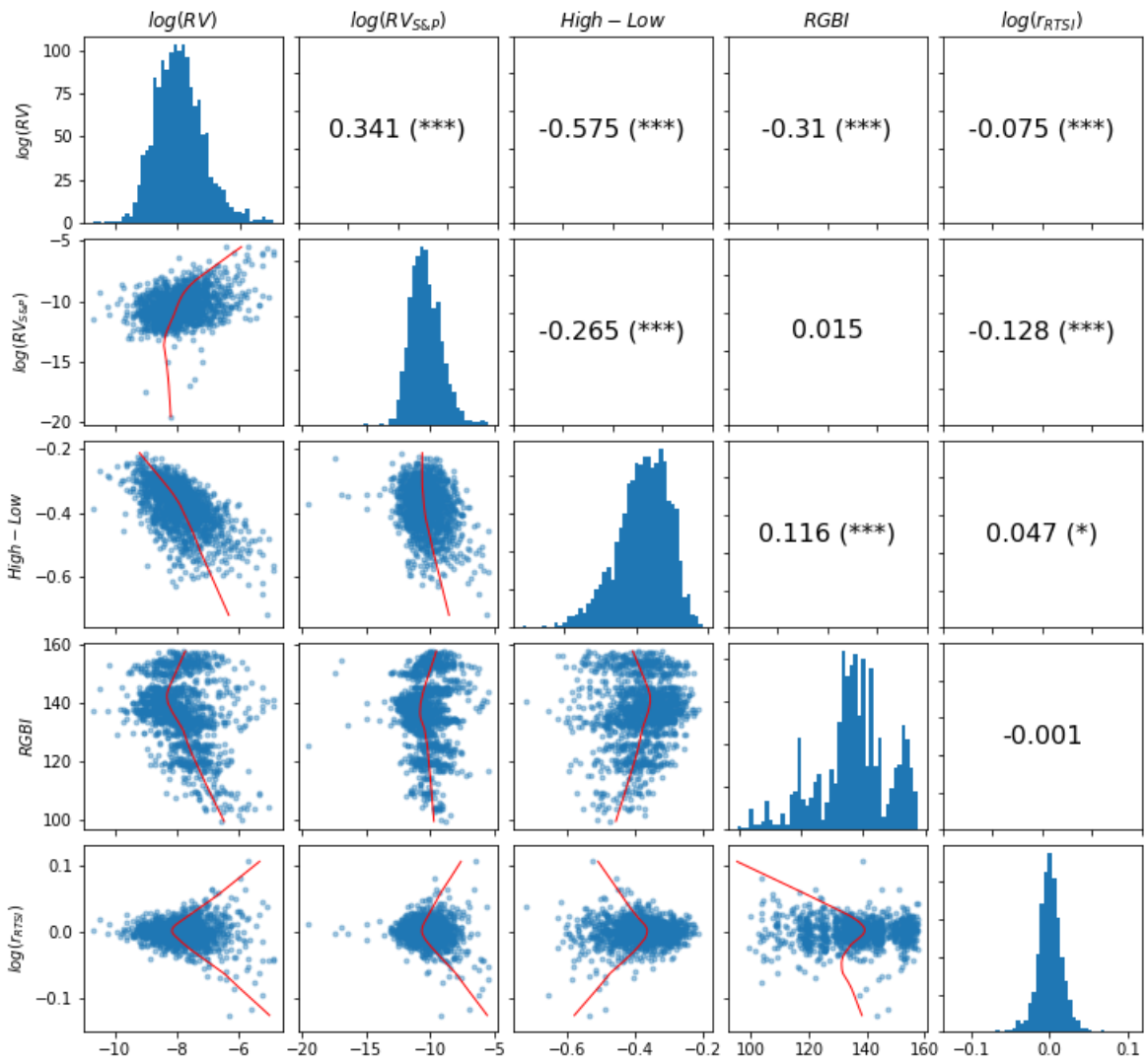


Figure 38: Correlations and dependencies between selected variables, POLYMETAL

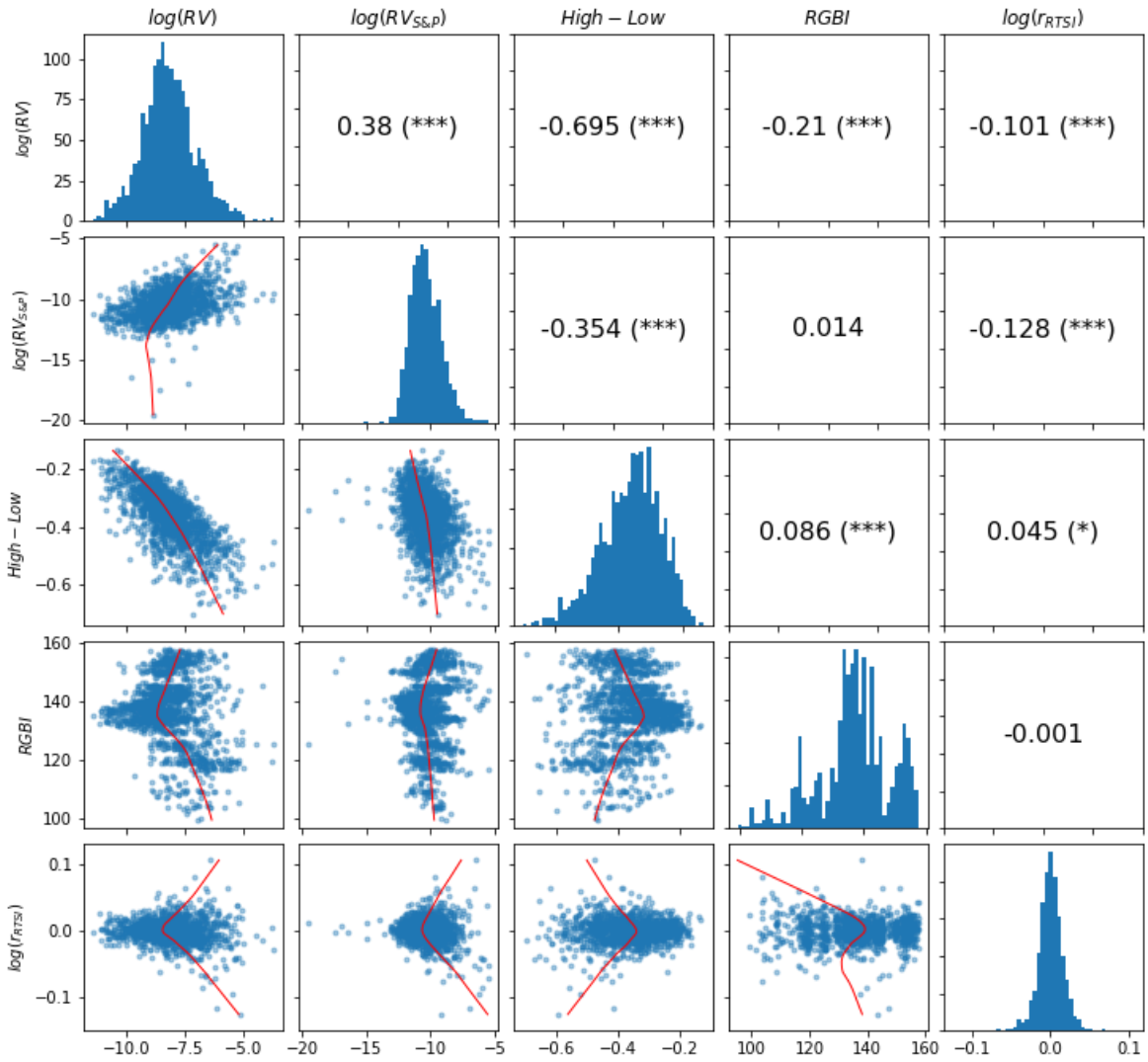


Figure 39: Correlations and dependencies between selected variables, POLYUS

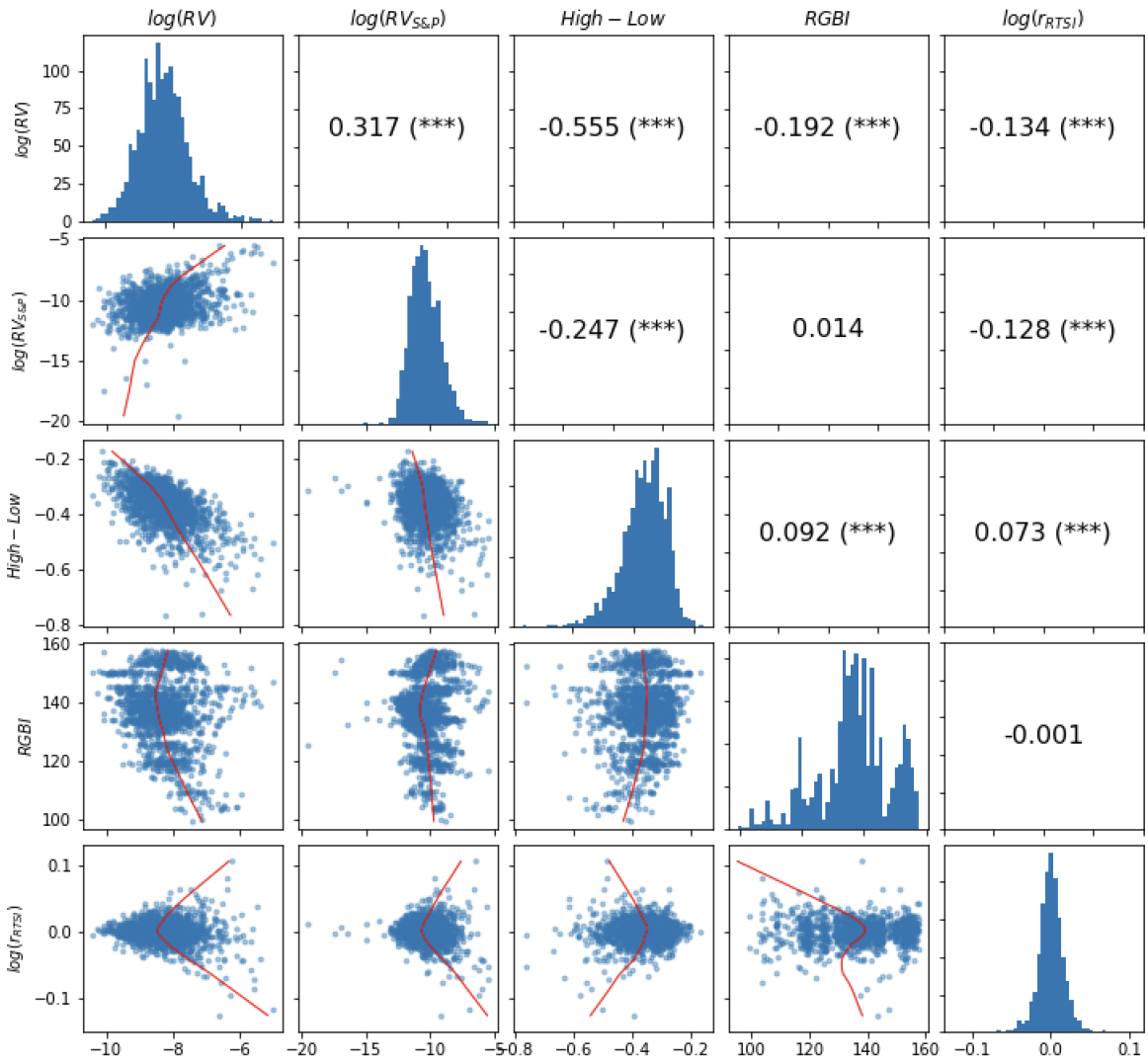


Figure 40: Correlations and dependencies between selected variables, MAGNIT

A Quality of approximation of correlation parameter

Attractiveness of the methodology relies heavily on the quality of the approximation (1.6). It suggests using a correlation of pseudo-observations $\mathcal{U}$ from either Gaussian or t copula as an approximation for the copula correlation matrix parameter P. We demonstrate the scope of this approximation for these two copulas in the bivariate case, i.e. for the copula parameter

$$P = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{A.1}$$

The approximation (1.6) suggests that

$$\mathrm{cor}(u_1, u_2) \approx \rho, \tag{A.2}$$

where $(u_1, u_2)' \sim C_P$. We run a simulation to evaluate $\mathrm{cor}(u_1, u_2)$ from $B = 2^{26}$ simulated values of $(u_1, u_2)'$ from the Gaussian and t copulas; in the latter case, the degrees of freedom parameter ν varies in $\{2, 4, 8, 10, 16\}$). We evaluate the error of this approximation for different values of ρ. The results are summarized in Figure 41. In a nutshell, the approximation error is negligible in all cases of practical interest.

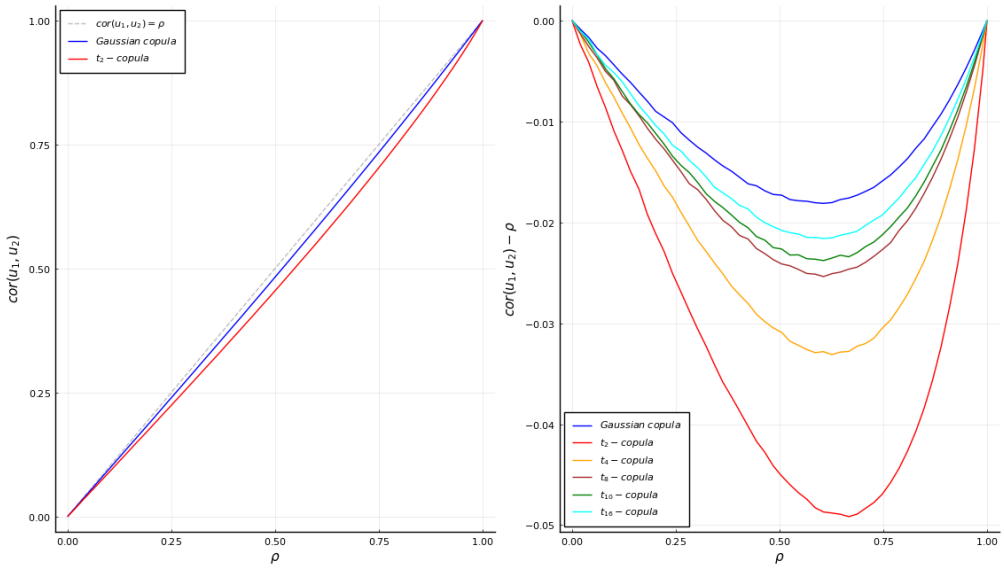


Figure 41: Approximation of copula parameter by correlation of pseudo-observations

B Portfolio selection and evaluation technique

Assume we have historical data on stock prices (daily, close) for a set of p stocks over the period of T days, $\{S_t^i\}_{i=1,...,p,t=1,...,T}$. We call *a portfolio* a p-dimensional vector of shares, $\alpha = (\alpha_1, ..., \alpha_p)$, such that $\forall i = 1, ..., p\ \alpha_i \geq 0$, and $\sum_{i=1}^p \alpha_i = 1$. The *value of portfolio* α is then the corresponding linear combination of the stock prices:

$$\pi_t(\alpha) = \sum_{i=1}^{p} \alpha_i S_t^i. \tag{B.1}$$

We use the portion of the historical data for the periods $t = 1, ..., n < T$ to fit a particular model for the stock price dynamics, based on which a particular portfolio is selected according to some criteria introduced below. The portfolio is then held for the rest $T - n$ time periods, $t = n + 1, ..., T$. The ratio of the current value of the portfolio to its initial value is then what we call *cumulative return* of the portfolio up to that time period:

$$X_t(\alpha) = \frac{\pi_t(\alpha)}{\pi_n(\alpha)},\ t > n. \tag{B.2}$$

We use the following modeling technique.

1. Since all price series are non-stationary, to model price dynamics we switch to daily log-returns,
$$r_t^i = \log\left(S_t^i\right) - \log\left(S_{t-1}^i\right).$$

2. For each of the log-return series, we use the historical data over the period $t = 1, ..., n$ to estimate a series of *ARMA-EGARCH* models of order up to (6,6)-(1,1,1). We run a simple in-sample diagnostics of each specification dropping those that do not pass the Ljung-Box test for standardized residual autocorrelation or the LM test for autoregressive conditional heteroskedasticity, and from the remaining specifications we pick the one with the minimal BIC value. For each asset we then record the estimates of the conditional mean rquation and conditional variance specifications and extract the corresponding standardized residual series, $\{e_t^i\}_{i=1,...,p,t=1,...,n}$.

3. Two different types of models are then used to model the *joint distribution* of residuals across all stocks, $\mathbf{e}_t = (e_t^1, ..., e_t^p)'$:

   - The multivariate normal (MVN) model:

$$\mathbf{e}_t \sim \text{i.i.d. } \mathcal{N}(\mathbb{O}_p, \Omega), \tag{B.3}$$

     where Ω is the correlation matrix, which is estimated by either linear or non-linear shrinkage, $\widehat{\Omega}$, of the standardized residuals over the period $t = 1, ..., n$.

   - The t copula model with EDF marginals:

$$\mathbf{e}_t \sim \text{i.i.d. } C_{P,\nu}^t \left(\widehat{F}^1(e^1), ..., \widehat{F}^p(e^p) \right), \tag{B.4}$$

     where $\widehat{F}^i(e)$ is the EDF of the i^{th} standardized residual series estimated over the period $t = 1, ..., n$. The matrix parameter P can be estimated by any of the method-of-moments-like estimators described earlier in the paper (Sections 1.3.2 & 1.3.3), and the degrees-of-freedom parameter ν is estimated via MPLE. We use only the two shrinkage estimators of the matrix parameter.

4. From each model, we generate $B = 2^{10}$ trajectories of future error terms for the period $t = n+1, ..., T$, $\{e_t^i(b)\}_{i=1,...,p, t=n+1,...,T, b=1,...,B}$, and use the fitted ARMA-EGARCH specifications to calculate the corresponding trajectories of future stock prices, $\{\widehat{S}_t^i(b)\}_{...}$. We use these simulated data to calculate the simulation analogs of the portfolio value (B.1) and return (B.2) as

$$\widehat{\pi}_t(\alpha, b) = \sum_{i=1}^{p} \alpha_i \widehat{S}_t^i(b), \text{ and} \tag{B.5}$$

$$\widehat{X}_t(\alpha, b) = \frac{\widehat{\pi}_t(\alpha, b)}{\pi_n(\alpha)}. \tag{B.6}$$

5. For each portfolio α, we use as the main performance criterion the simulated sample Sharpe ratio based on the cumulative returns in the final period T estimated over the

simulations $b = 1, ..., B$ (and assuming zero risk-free return):

$$\xi(\alpha) = \frac{B^{-1} \sum_{b=1}^{B} \widehat{X}_T(\alpha, b)}{\sqrt{B^{-1} \sum_{b=1}^{B} \left(\widehat{X}_t(\alpha, b) - B^{-1} \sum_{b=1}^{B} \widehat{X}_T(\alpha, b) \right)^2}}. \tag{B.7}$$

6. We choose the portfolio with the best Sharpe ratio:

$$\alpha^* = \arg\max_{\alpha} \xi(\alpha). \tag{B.8}$$

   This results in 4 different model-based portfolio choices: $\alpha^*_{\text{MVN-LSh}}$, $\alpha^*_{\text{MVN-NLSh}}$, $\alpha^*_{\text{tc-LSh}}$, $\alpha^*_{\text{tc-NLSh}}$, depending on which model is used to simulate the stock price trajectories and calculate the simulated portfolio returns (B.6).

7. As a benchmark for a given set of p stocks we use the equally weighted portfolio, $\alpha_{1/p} = (p^{-1}, ..., p^{-1})$. To evaluate the actual performance of the portfolios over the period $t = n+1, ..., T$, we calculate, for each set of assets and the corresponding choices of α^*, the ratios of the actual return in time period T of the different model-based portfolios to each other:

$$\tilde{R}(M_1, M_2) = \frac{X_T(\alpha^*_{M_1})}{X_T(\alpha^*_{M_2})}, \tag{B.9}$$

   where $M_1 \in \{\text{tc-LSh},\ \text{tc-NLSh}\}$, $M_2 \in \{1/p, \text{MVN-LSh}, \text{MVN-NLSh}, \text{tc-NLSh}\}/M_1$.

The interpretation of the measures (B.9) is the following. The purpose of this empirical exercise is to show the potential gains of the combination of copula-based models and shrinkage-based estimators over traditional techniques. The higher the relative cumulative return of a model-based portfolio $R(M_1, M_2)$ is, the better is the model's choice M_1 over M_2, with the preferred range of the criterion being above 1.

Still, the resulting portfolio performance measures (B.9) are single numbers, and the result is random for a particular set of assets, choice of sample sizes, and dates. We therefore run another simulation to compare different model-based portfolio choices.

First, we set as a modeling period approximately the last 9 months of the year 2017. We use $n = 120$ daily observations to fit and estimate the models. The remaining $T - n =$

60 observations are used to run the simulations, select the portfolios, and evaluate their performance. The sample sizes are intentionally very low. One reason to keep them such is that, clearly, the quality of simulations of stock prices crucially depends on the quality of univariate conditional mean models of the log-return series. In our example, these models are very simplistic, and one should not expect that their performance can remain relevant for a long period of time. However, normally, the shorter the samples are, the lower should be the number of assets in potential portfolios, exactly due to the curse of dimensionality. In our case, this is another reason to keep the samples short so that we can make the point that the high-dimensionality adjustment in estimation techniques can be beneficial even when the sample is very short.

Second, in the interest of not over-complicating asset selection for potential portfolios, from all securities for which we managed to access the data, we drop the series whose log-returns fail stationarity tests or for which we could not select an ARMA-EGARCH specification (for example, if none of the specifications deliver residuals that pass the Ljung-Box or LM tests). This leaves us with approximately 4980 securities from over 5000 initially.

From the remaining securities we randomly choose $K > 2^7$ subsets[16] of size $p = 3600$, and for each of them perform steps 1–7 above. Thus, we obtain a distribution of the overall performance of different strategies of portfolio construction (B.9) over 135 randomly chosen sets of $p = 3600$ assets.

Finally, under $p = 3600$ the optimization problem (B.7) is very high-dimensional. To make its solution computationally practical (in each simulation it needs to be solved up 4 times), we substitute the actual optimization (B.7) with a choice over a number greater than 10^6 of portfolios α randomly and uniformly generated from p-dimensional simplex. The set of alternative αs is pre-generated and remains fixed across all simulations as long as the dimensionality p remains the same. The resulting choices of the portfolios α^* are not guaranteed to be optimal, however, given the dimensionality of the optimization problem, and its simulation nature, the search on a randomly pre-generated set of alternatives is believed to be the best computationally feasible choice.

[16] we ran the simulations for over $K = 150$ subsets to obtain result for 135 of them, the remaining 15 were dropped due to poor convergence of optimization or numerical errros during paralleled computations

C Technical remarks

C.1 Computational software

All the calculations for the simulation study were performed using R language (R Core Team, 2013). The packages *foreach* (Weston, 2019b) and *doParallel* (Weston, 2019a) were used to perform parallel computations. The package *copula* (Hofert et al., 2018; Jun Yan, 2007; Ivan Kojadinovic and Jun Yan, 2010; Marius Hofert and Martin Mächler, 2011) was used to simulate the random variables from the copula-based multivariate distributions and calculate copula density functions. To perform linear and nonlinear shrinkage covariance matrix estimators, the package *nlshrink* (Ramprasad, 2016) was used. Other packages used in particular calculations include *Matrix, matrixcalc, pcaPP, corrplot* (Bates and Maechler, 2019; Novomestky, 2012; Filzmoser et al., 2018; Wei and Simko, 2017), and others.

The empirical example was evaluated in the Julia programming language (Bezanson et al., 2017). Particularly, the package *ARCHModels* (Broda and Paolella, 2020) was applied to estimate and select ARMA-EGARCH models.

C.2 Evaluation time of estimators

We assess the time required for evaluation of the four estimators of matrix parameters of the t copula for different true matrix parameter structures (identity or arbitrary) and under different dimensionality $p/n \in \{1/2,\ 1,\ 2\}$. The results are reported in Table 4. The assessment of evaluation time was performed on an Intel(R) Core(TM) i7-7700K CPU @4.20GHz machine with 16GB of RAM running on Windows 10 Home edition. For assessing evaluation time, no parallel computing was used. The R package *microbenchmark* (Mersmann, 2019) was used to record the running time of the four estimators evaluation.

D Dynamic portfolio allocation technique

D.1 Definitions

This section presents the overview of the portfolio selection and evaluation algorithm we use in Chapter 2 to produce the results discussed in Section 2.4. We apply the following definitions.

1. *Stock prices*, $\{S_{ij}\}_{i=0,1,...N,j=1,...p}$, are daily close prices of stocks, indexed by $j = 1, ...p$, over the time period from $i = 1$ to N days.

2. *Daily stock returns*, $r_{ij} = \log S_{ij} - \log S_{(i-1)j}$, $i = 1, ...N$, $j = 1, ...p$.

3. *A portfolio*, $\alpha = (\alpha_1, ...\alpha_p)'$, is a vector of shares that correspond to the input of each stock to the portfolio, i.e. $\alpha_j \geq 0$, $\forall j = 1, ...p$, and $\sum_{j=1}^{p} \alpha_j = 1$.

4. *Equally-weighted portfolio (EWP)* or *naïve portfolio* is $\alpha^{1/p} = (1/p, ...1/p)'$.

5. *A portfolio's value* at time $i = 1, ...N$, $X_i(\alpha) = \sum_{j=1}^{p} S_{ij}\alpha_j$.

6. *A portfolio's accumulated return* over a period from i_1 to i_2, $x_{i_1,i_2}(\alpha) = \frac{X_{i_2}(\alpha)}{X_{i_1}(\alpha)}$.

7. *A portfolio's Sharpe's ratio* (assuming zero risk-free return) for the period from i_1 to i_2, $\xi_{i_1,i_2}(\alpha) = \frac{\mathbb{E}x_{i_1,i_2}(\alpha)}{\sqrt{\mathrm{Var}x_{i_1,i_2}(\alpha)}}$.

Investment strategy

In this study, we employ dynamic portfolio allocation, mimicking an investment strategy under conditions of full reinvestment and no transaction costs. The whole timeframe $i = 1., , , N$ is divided into equal periods $t = 1, ...T$, with starting points i_1^t and end points i_2^t, such that

$$i_1^1 < i_2^1 = i_1^2 < i_2^2 = i_1^3 < ...i_2^{T-1} = i_1^T < i_2^T = N,$$

and

$$i_1^1 = i_2^t - i_1^t = n, \ \forall t = 1, ...T.$$

An investment strategy is a vector of period-specific portfolios, $A = (\alpha^1, ...\alpha^T)$. Each of the portfolios, α^t, is selected at the time point i_1^t, and over the period t it generates accumulated return $x_{i_1^t, i_2^t}(\alpha^t)$. At the beginning of the next period $t+1$, the full final value of the previous portfolio $X_{i_2^t}(\alpha^t)$ is re-invested into the next portfolio α^{t+1}. Thus, at each time-point i between i_1^1 and i_2^T, the *accumulated return of an investment strategy* is

$$Z_i(A) = \begin{cases} x_{i_1,i}(\alpha^1), \text{ if } i_1^1 \leq i < i_2^1, \\ x_{i_1,i}(\alpha^{\bar{t}}) \prod_{t=1}^{\bar{t}-1} x_{i_1^t, i_2^t}(\alpha^t), \text{ if } i_1^{\bar{t}} \leq i \leq i_2^{\bar{t}}, \ \bar{t} = 2, ...T. \end{cases}$$

D.2 Portfolio selection technique

According to an investment strategy, the portfolio is selected (updated) T times. We perform each portfolio selection $t = 1, ...T$following the same algorithm:

1. Each of the stocks $j = 1, ...p$ returns series r_{ij} over the historical period of time $i = i_1^t - n + 1, ...i_1^t$ is filtered using *ARMA-(E)GARCH* model of sufficient small order to obtain serially-uncorrelated residual returns e_{ij}.

2. The sample $\{e_{ij}\}_{i=i_1^t-n+1,...i_1^t, j=1,...p}$ is transformed into pseudo-observations (2.14) using ECDF of each particular series $j = 1, ...p$. Skew-t copula is fitted to the pseudo-observations using the algorithm described in Section 2.3.2.

3. The estimate of the copula is used as the model of the joint distribution of the residual returns, and the fitted *ARMA-(E)GARCH* models are used as conditional mean models for the assets' returns. From these models, we simulate trajectories of stock prices S_{ij}, $j = 1, ...p$ for the period $i = i_1^t + 1, ...i_2^t$ to solve the optimal Sharpe's ratio portfolio selection problem:

$$\alpha^{*t} = \arg\max_a \xi_{i_1^t, i_2^t}(a).$$

We perform the *ARMA-(E)GARCH* filtration step in the same manner as by Anatolyev and Pyrlik (2022), with a lower maximum order of the models, as in this study the sample size $n = 20$. Our simulation of random values from skew-t copula distribution is performed using the stochastic representation of MSTD (2.2) - (2.6) and the transformation (2.14).

We measure the realized performance of the portfolio in terms of the whole investment strategy accumulated return $Z_N(A^*)$ compared to the naïve portfolio's return $x_{i_1^1,N}(\alpha^{1/p})$ and the accumulated return of the market index $\frac{S_{i_1^1,\text{EUROSTOXX50}}}{S_{N,\text{EUROSTOXX50}}}$.

Technical remarks

We performed all the calculations for this study in the Julia programming language (Bezanson et al., 2017). The package *Distributed* (Jeff Bezanson and other contributors, 2022) was used to perform parallel computations. We handled the optimization problems in SGMM application using package *Optim* (Mogensen and Riseth, 2018). We apply package *ARCH-Models* (Broda and Paolella, 2020) to estimate and select ARMA-EGARCH models. We produce visualizations using *PlotlyJS* (Lyon, 2022). Other packages that we use in particular calculations include *LinearAlgebra, Distributions, Roots*, and others.

E Machine Learning Algorithms

Lasso

Lasso (*least absolute shrinkage and selection operator*) is a type of regularization that performs variable selection. This method aims to enhance predictive power of linear models and to highlight the most valuable predictors. As OLS models tend to exhibit low bias and high variance of forecasts, they may be overfitted. Lasso regularization helps to decrease the risk of overfitting.

In comparison with the Ordinary Least Squares method, there is an added penalty in Lasso:

$$||Xw - y||_2^2 - \lambda||w||_1 \to min \tag{E.1}$$

where λ is a hyperparameter of Lasso regularization that can be interpreted as the penalty rate.

Gradient Boosting

Gradient Boosting was firstly described by Friedman (2001), and its main idea is to create an ensemble of simple models by sequentially fitting and adding parameterized functions. The main goal of this algorithm is to find a function $F^*(x)$ that maps x to y and specific (differentiable) loss function $\Psi(y, F(x))$ is minimized:

$$F^*(x) = \arg\min_{F(x)} \Psi(y, F(x)) \tag{E.2}$$

$F(X)$ is represented as a linear combination of simple models (base learners):

$$F(x) = \sum_{m=o}^{M} \beta_m h(x; a_m) \implies F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m) \tag{E.3}$$

where $h(a; x)$ (base learner) is a simple function of x with parameters $a = \{a_1, a_2, ...\}$, β_m is the coefficient behind each base learner h_m to make a linear combination.

Suppose, we have a training sample $\{y, x\}$ of size N. So, function $F^*(x)$ could be fitted by repeating two steps m times. Firstly, function $h(x; a)$ is fitted via least squares

$$a_m =_a \sum_{i=1}^{N} [y^*_{im} - h(a; x_i)]^2 \tag{E.4}$$

to the current 'pseudo'-residuals

$$y^*_{im} = -\left[\frac{\partial \Psi(y_{im}, F(x_{im}))}{\partial F(x_{im})}\right]_{F(x)=F_{m-1}(x)} \tag{E.5}$$

Then the optimal value of β_m is determined

$$\beta_m = \arg\min_{\beta} \sum_{i=0}^{N} (\Psi(y_i, F_{m-1}(x_i) + \beta h(a_m; x_i))) \tag{E.6}$$

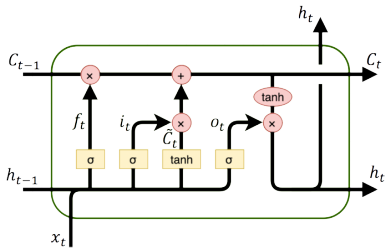


Figure 42: LSTM scheme

This method replaces a potentially difficult function optimization problem with one based on least squares followed by a single parameter optimization based on general loss criterion Ψ.

The most commonly used base learners are binary decision trees (we use them in our study). Further, we use *L1* regularization on this algorithm to deal with overfitting. This type of regularization works on Gradient Boosting by constraining the leaf weights, rather than the feature weights.

Random Forest

Random forest was firstly described by Breiman (2001). The algorithm consists of tree-structured algorithms $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$ where $\{\Theta_k\}$ are i.i.d. random vectors. Each tree casts a unit vote at input $\mathbf{x}$. For the regression problem, RF extracts the mean vote of all the trees. The mechanism of tree-structured algorithms is the same as for GB, described above.

LSTM

Long Short-Term Memory (LSTM) was first described by Hochreiter and Schmidhuber (1997). LSTM is a kind of recurring neural network tahn can be represented as a sequence of blocks.

The principal work of a block is described in Figure 42. This block can also be rewritten as:

$$
\begin{aligned}
f_t &= \sigma(x_t U^f + h_{t-1} W^f + b^f), \\
i_t &= \sigma(x_t U^i + h_{t-1} W^i + b^i), \\
o_t &= \sigma(x_t U^o + h_{t-1} W^o + b^o), \\
\tilde{C}_t &= \tanh(x_t U^g + h_{t-1} W^g + b^g), \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t, \\
h_t &= \tanh(C_t) * o_t.
\end{aligned}
$$

where W, U and b are the parameters of the block, x_t is an input vector, h_t is an output vector and C_t is a vector of conditions. The two functions σ and $tanh$ are $\sigma(x) = \frac{1}{1-e^{-x}}$, and $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. LSTM is typically trained with the use of *stochastic gradient descent* or one of its modifications. In our study, we use firstly *ADAM* optimizer, but we switch to alternatives in case it fails.

F Machine Learning Results

F.1 Graphs with Top-1 Models

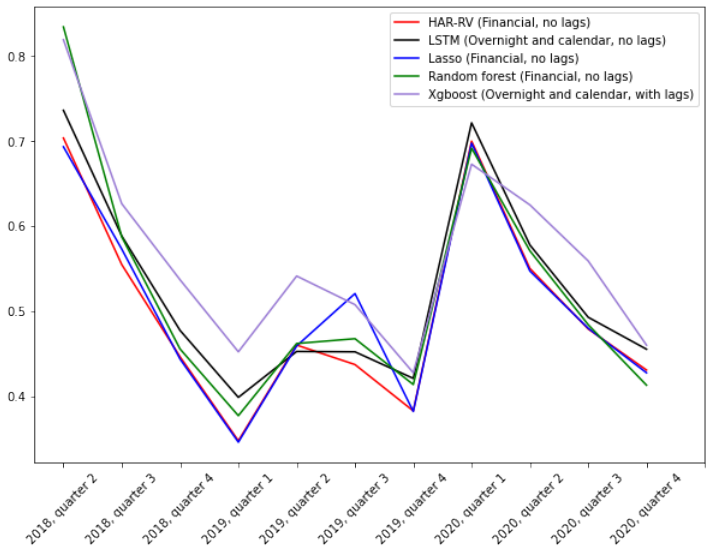


Figure 43: Results for top-1 ML models, SBERBANK

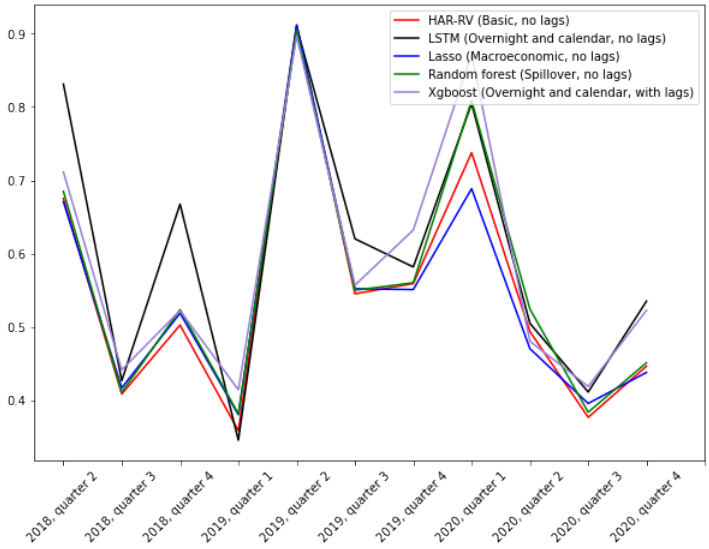


Figure 44: Results for top-1 ML models, GAZPROM

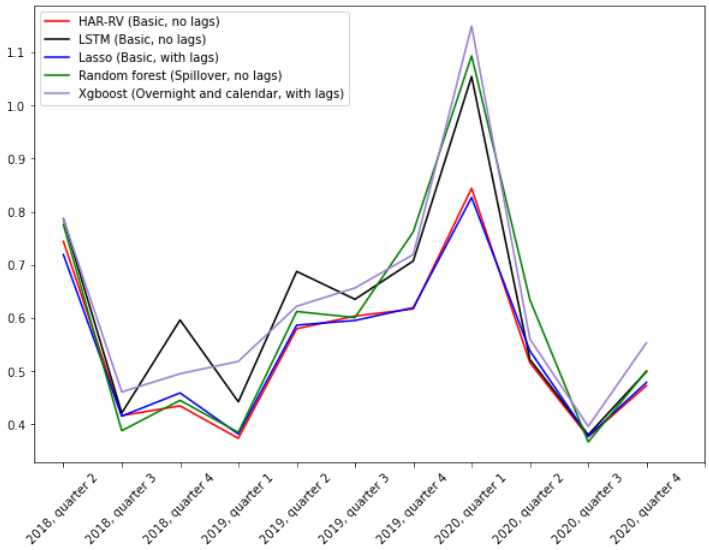


Figure 45: Results for top-1 ML models, LUKOIL

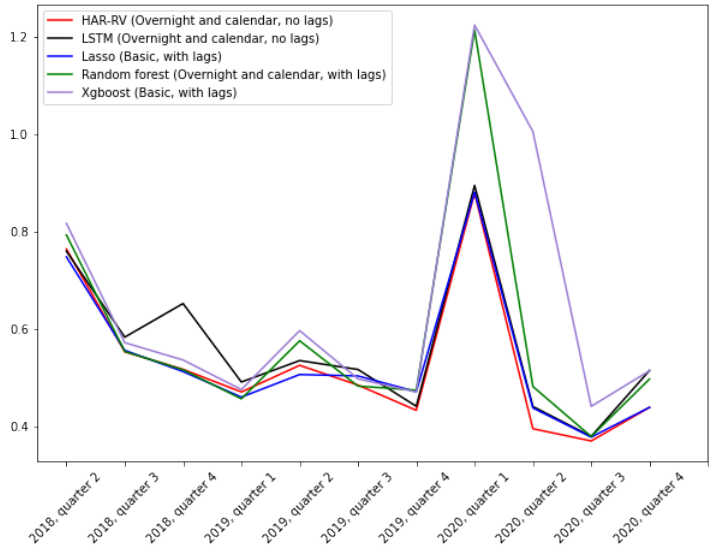


Figure 46: Results for top-1 ML models, NOVATEK

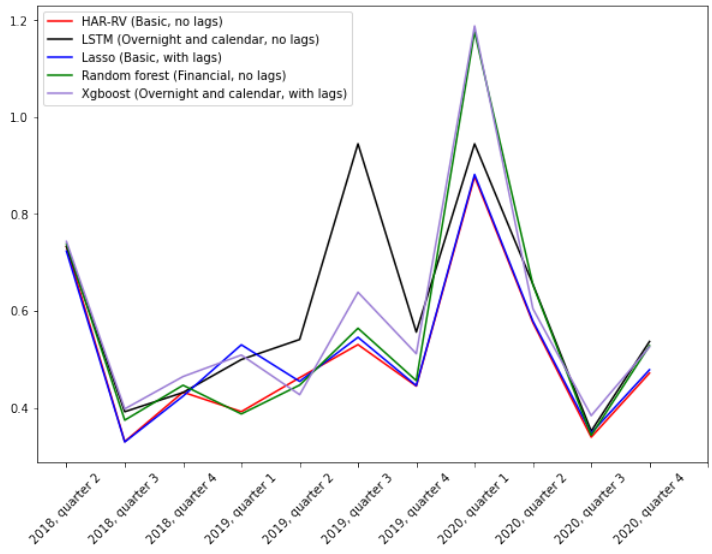


Figure 47: Results for top-1 ML models, ROSNEFT

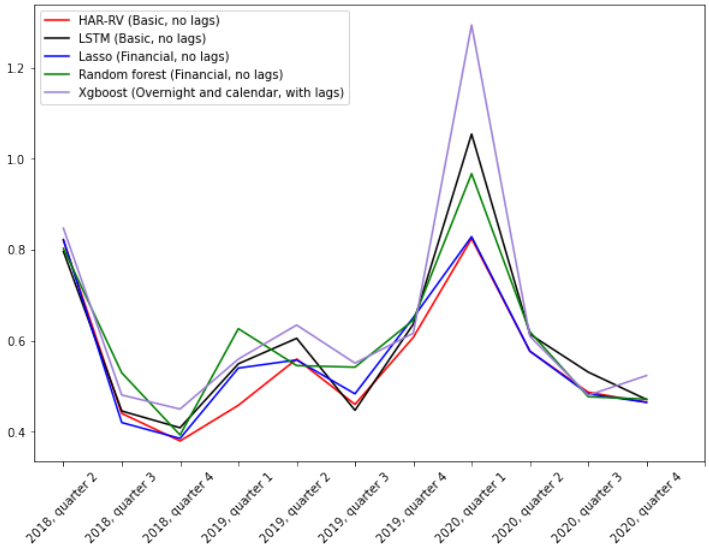


Figure 48: Results for top-1 ML models, NORNICKEL

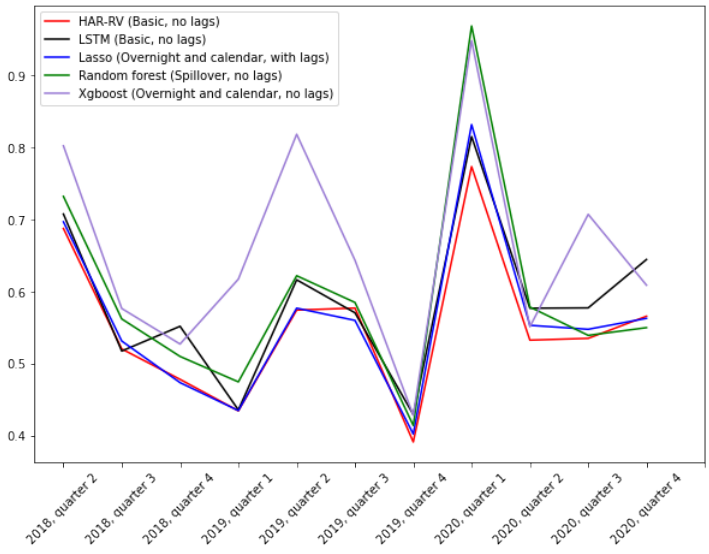


Figure 49: Results for top-1 ML models, POLYMETAL

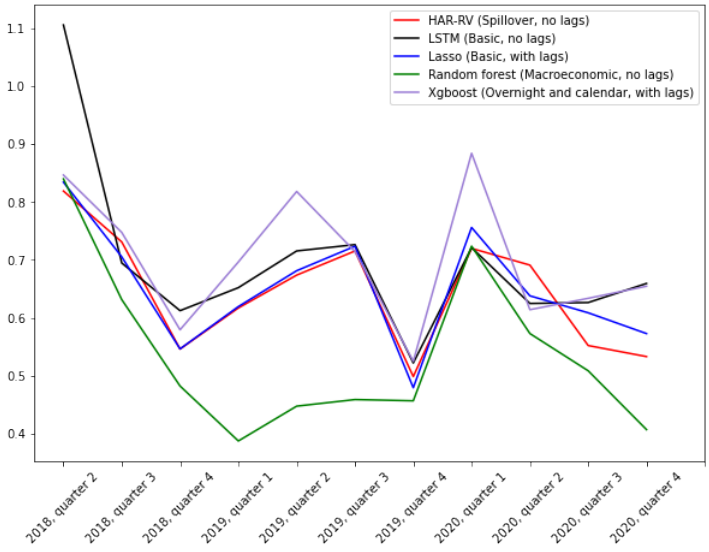


Figure 50: Results for top-1 ML models, POLYUS

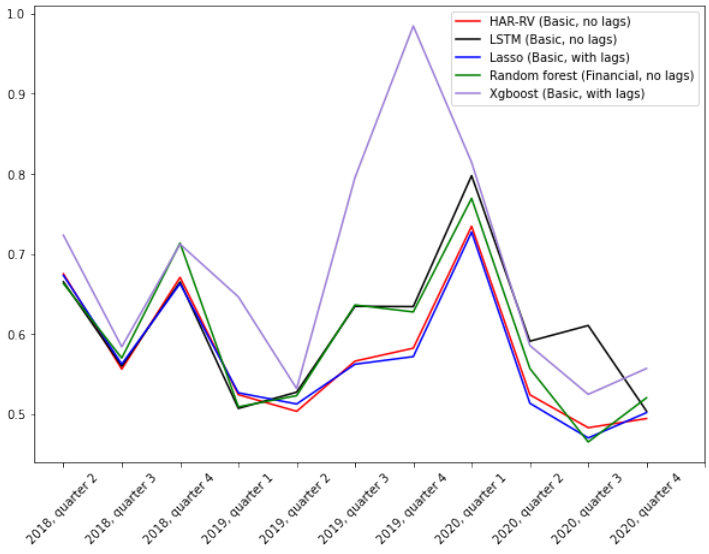


Figure 51: Results for top-1 ML models, MAGNIT

F.2 Tables with Top-3 Models

Table 23: Results of models for SBERBANK with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Financial	Financial	Spillover	Basic	Overnight and calendar	Basic	Basic	Financial	Spillover	Basic	Overnight and calendar	Basic	Spillover
	no lags	no lags	no lags	with lags	with lags	with lags	no lags	no lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.7	0.69	0.68	0.7	0.82	0.82	0.81	0.83	0.83	0.8	0.74	0.79	0.76
2018 Q3	0.55	0.57	0.58	0.58	0.63	0.62	0.62	0.59	0.6	0.67	0.59	0.63	0.6
2018 Q4	0.45	0.44	0.45	0.45	0.54	0.54	0.51	0.46	0.5	0.51	0.48	0.49	0.59
2019 Q1	0.35	0.35	0.36	0.35	0.45	0.47	0.43	0.38	0.39	0.35	0.4	0.43	0.41
2019 Q2	0.46	0.46	0.48	0.48	0.54	0.53	0.5	0.46	0.46	0.48	0.45	0.46	9.19
2019 Q3	0.44	0.52	0.52	0.51	0.51	0.56	0.5	0.47	0.46	0.48	0.45	0.61	0.6
2019 Q4	0.38	0.38	0.38	0.39	0.43	0.41	0.5	0.41	0.42	0.39	0.42	0.39	0.83
2020 Q1	0.7	0.7	0.7	0.69	0.67	0.68	0.84	0.69	0.72	0.71	0.72	0.69	1.35
2020 Q2	0.55	0.55	0.54	0.54	0.62	0.61	0.56	0.57	0.55	0.55	0.58	0.55	0.58
2020 Q3	0.48	0.48	0.48	0.48	0.56	0.56	0.55	0.48	0.49	0.54	0.49	0.48	0.57
2020 Q4	0.43	0.43	0.43	0.41	0.46	0.43	0.44	0.41	0.42	0.42	0.46	0.47	0.59
Mean RMSE	0.5	0.51	0.51	0.51	0.57	0.57	0.57	0.52	0.53	0.54	0.52	0.54	1.46

Table 24: Results of models for GAZPROM with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Macroeconomic	Financial	Financial	Overnight and calendar	Basic	Spillover	Spillover	Financial	Spillover	Overnight and calendar	Basic	Financial
	no lags	no lags	no lags	with lags	with lags	with lags	no lags	no lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.68	0.67	0.67	0.67	0.71	0.74	0.72	0.68	0.69	0.7	0.83	0.72	0.75
2018 Q3	0.41	0.42	0.4	0.42	0.44	0.44	0.43	0.41	0.42	0.39	0.43	0.41	0.43
2018 Q4	0.5	0.52	0.49	0.51	0.52	0.53	0.56	0.52	0.52	0.51	0.67	0.67	0.67
2019 Q1	0.36	0.38	0.37	0.35	0.41	0.43	0.4	0.38	0.38	0.37	0.35	0.43	0.43
2019 Q2	0.91	0.91	0.92	0.91	0.9	0.88	0.89	0.91	0.92	0.93	0.91	0.91	0.96
2019 Q3	0.55	0.55	0.55	0.55	0.56	0.57	0.64	0.55	0.53	0.55	0.62	0.55	0.62
2019 Q4	0.56	0.55	0.55	0.55	0.63	0.65	0.63	0.56	0.55	0.59	0.58	0.76	0.58
2020 Q1	0.74	0.69	0.72	0.74	0.87	0.9	0.98	0.81	0.81	0.81	0.8	0.86	0.76
2020 Q2	0.49	0.47	0.52	0.49	0.48	0.49	0.56	0.52	0.54	0.51	0.51	0.54	0.86
2020 Q3	0.38	0.4	0.38	0.4	0.42	0.41	0.46	0.38	0.4	0.38	0.41	0.46	0.74
2020 Q4	0.45	0.44	0.45	0.45	0.52	0.51	0.51	0.45	0.45	0.47	0.54	0.46	0.74
Mean RMSE	0.55	0.55	0.55	0.55	0.59	0.6	0.62	0.56	0.56	0.56	0.6	0.61	0.68

Table 25: Results of models for LUKOIL with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Basic	Overnight and calendar	Financial	Overnight and calendar	Basic	Spillover	Spillover	Overnight and calendar	Overnight and calendar	Basic	Overnight and calendar	Spillover
	no lags	with lags	with lags	no lags	with lags	with lags	with lags	no lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.74	0.72	0.72	0.74	0.79	0.83	0.75	0.78	0.78	0.77	0.79	0.76	0.77
2018 Q3	0.42	0.42	0.42	0.4	0.46	0.47	0.44	0.39	0.41	0.42	0.42	0.5	0.4
2018 Q4	0.43	0.46	0.46	0.44	0.5	0.49	0.48	0.45	0.46	0.44	0.6	0.44	0.45
2019 Q1	0.37	0.38	0.38	0.38	0.52	0.5	0.66	0.38	0.39	0.38	0.44	0.46	0.66
2019 Q2	0.58	0.59	0.59	0.56	0.62	0.63	0.63	0.61	0.57	0.57	0.69	0.61	0.64
2019 Q3	0.6	0.6	0.6	0.59	0.66	0.65	0.72	0.6	0.61	0.62	0.64	0.63	0.65
2019 Q4	0.62	0.62	0.62	0.63	0.72	0.7	0.85	0.76	0.71	0.7	0.71	0.82	0.74
2020 Q1	0.84	0.83	0.84	0.9	1.15	1.16	1.11	1.09	1.04	1.16	1.05	1.02	1.35
2020 Q2	0.52	0.54	0.54	0.59	0.56	0.56	0.7	0.63	0.78	0.7	0.52	0.54	0.57
2020 Q3	0.38	0.38	0.38	0.37	0.4	0.4	0.37	0.37	0.39	0.38	0.38	0.61	0.38
2020 Q4	0.47	0.48	0.48	0.46	0.55	0.57	0.57	0.5	0.54	0.55	0.5	0.54	0.61
Mean RMSE	0.54	0.55	0.55	0.55	0.63	0.63	0.66	0.6	0.61	0.61	0.61	0.63	0.66

Table 26: Results of models for NOVATEK with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Overnight and calendar	Basic	Overnight and calendar	Overnight and calendar	Basic	Overnight and calendar	Overnight and calendar	Overnight and calendar	Basic	Spillover	Overnight and calendar	Basic	Macroeconomic
	no lags	with lags	with lags	no lags	with lags	with lags	no lags	with lags	with lags	no lags	no lags	no lags	no lags
2018 Q2	0.76	0.75	0.75	0.77	0.82	0.83	0.85	0.79	0.79	0.76	0.76	0.78	0.8
2018 Q3	0.55	0.56	0.56	0.58	0.57	0.57	0.53	0.55	0.56	0.55	0.58	0.56	0.59
2018 Q4	0.52	0.51	0.51	0.53	0.54	0.53	0.58	0.52	0.51	0.53	0.65	0.64	0.57
2019 Q1	0.47	0.46	0.46	0.47	0.47	0.45	0.6	0.46	0.46	0.51	0.49	0.48	1.07
2019 Q2	0.52	0.51	0.51	0.53	0.6	0.6	0.64	0.58	0.57	0.55	0.53	0.53	0.54
2019 Q3	0.48	0.5	0.5	0.5	0.5	0.56	0.54	0.48	0.48	0.51	0.52	0.5	0.63
2019 Q4	0.43	0.47	0.47	0.47	0.47	0.5	0.6	0.47	0.47	0.48	0.44	0.98	0.82
2020 Q1	0.88	0.88	0.88	0.9	1.22	1.14	1.37	1.21	1.23	1.22	0.89	0.9	1.92
2020 Q2	0.39	0.44	0.44	0.44	1.0	1.01	0.92	0.48	0.48	0.49	0.44	0.5	0.52
2020 Q3	0.37	0.38	0.38	0.37	0.44	0.44	0.49	0.38	0.38	0.4	0.38	0.43	0.45
2020 Q4	0.44	0.44	0.44	0.44	0.51	0.54	0.48	0.5	0.5	0.46	0.51	0.47	0.51
Mean RMSE	0.53	0.54	0.54	0.54	0.65	0.65	0.69	0.58	0.59	0.59	0.56	0.62	0.76

Table 27: Results of models for ROSNEFT with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Basic	Overnight and calendar	Financial	Overnight and calendar	Basic	Financial	Financial	Financial	Spillover	Overnight and calendar	Basic	Macroeconomic
	no lags	with lags	with lags	no lags	with lags	with lags	with lags	no lags	with lags	with lags	no lags	no lags	no lags
2018 Q2	0.73	0.72	0.72	0.72	0.74	0.75	0.74	0.74	0.74	0.73	0.73	0.74	0.82
2018 Q3	0.33	0.33	0.33	0.34	0.4	0.39	0.37	0.37	0.34	0.34	0.39	0.34	0.36
2018 Q4	0.43	0.42	0.42	0.42	0.46	0.49	0.48	0.45	0.45	0.45	0.43	0.44	0.42
2019 Q1	0.39	0.53	0.53	0.56	0.51	0.52	0.43	0.39	0.41	0.45	0.5	0.75	0.81
2019 Q2	0.46	0.45	0.45	0.44	0.43	0.44	0.46	0.45	0.44	0.45	0.54	0.64	8.43
2019 Q3	0.53	0.55	0.55	0.53	0.64	0.63	0.74	0.56	0.6	0.59	0.94	0.76	0.72
2019 Q4	0.44	0.45	0.45	0.45	0.51	0.52	0.53	0.46	0.45	0.45	0.56	0.73	0.64
2020 Q1	0.88	0.88	0.88	0.91	1.19	1.12	0.96	1.18	1.22	1.24	0.94	1.02	1.34
2020 Q2	0.58	0.58	0.58	0.64	0.6	0.64	0.92	0.65	0.62	0.59	0.66	0.66	1.15
2020 Q3	0.34	0.35	0.35	0.34	0.38	0.4	0.4	0.34	0.34	0.34	0.35	0.39	0.53
2020 Q4	0.47	0.48	0.48	0.48	0.53	0.51	0.53	0.53	0.51	0.51	0.54	0.48	0.59
Mean RMSE	0.51	0.52	0.52	0.53	0.58	0.58	0.6	0.56	0.56	0.56	0.6	0.63	1.44

Table 28: Results of models for NORNICKEL with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Financial	Basic	Overnight and calendar	Overnight and calendar	Basic	Spillover	Financial	Spillover	Basic	Basic	Overnight and calendar	Financial
	no lags	no lags	with lags	with lags	with lags	with lags	no lags	no lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.82	0.82	0.85	0.86	0.85	0.84	0.77	0.8	0.79	0.84	0.8	0.81	0.95
2018 Q3	0.44	0.42	0.45	0.45	0.48	0.48	0.54	0.53	0.51	0.48	0.45	0.43	0.44
2018 Q4	0.38	0.39	0.38	0.39	0.45	0.44	0.48	0.39	0.43	0.41	0.41	0.45	0.4
2019 Q1	0.46	0.54	0.54	0.54	0.56	0.57	0.82	0.63	0.64	0.53	0.55	0.56	0.58
2019 Q2	0.56	0.56	0.57	0.59	0.63	0.64	0.57	0.55	0.55	0.59	0.61	0.61	0.63
2019 Q3	0.46	0.48	0.46	0.46	0.55	0.54	0.55	0.54	0.55	0.47	0.45	0.61	0.46
2019 Q4	0.61	0.65	0.6	0.61	0.62	0.64	0.71	0.65	0.65	0.62	0.64	0.62	0.77
2020 Q1	0.82	0.83	0.82	0.84	1.29	1.28	0.95	0.97	0.98	1.11	1.05	0.9	1.4
2020 Q2	0.58	0.58	0.58	0.59	0.61	0.64	0.91	0.62	0.63	0.62	0.61	0.61	0.81
2020 Q3	0.49	0.48	0.49	0.48	0.48	0.48	0.57	0.48	0.48	0.53	0.53	0.78	0.45
2020 Q4	0.47	0.47	0.47	0.47	0.52	0.53	0.54	0.47	0.47	0.49	0.47	0.53	0.54
Mean RMSE	0.55	0.57	0.57	0.57	0.64	0.64	0.68	0.6	0.61	0.61	0.6	0.63	0.67

Table 29: Results of models for POLYMETAL with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Overnight and calendar	Basic	Basic	Overnight and calendar	Basic	Overnight and calendar	Spillover	Macroeconomic	Financial	Basic	Overnight and calendar	Financial
	no lags	with lags	no lags	with lags	no lags	with lags	with lags	no lags	no lags	no lags	no lags	no lags	no lags
2018 Q2	0.69	0.7	0.69	0.69	0.8	0.74	0.74	0.73	0.72	0.73	0.71	0.71	0.87
2018 Q3	0.52	0.53	0.52	0.52	0.58	0.63	0.62	0.56	0.52	0.57	0.52	0.54	0.56
2018 Q4	0.48	0.47	0.61	0.48	0.53	0.53	0.52	0.51	0.5	0.5	0.55	0.48	0.57
2019 Q1	0.43	0.44	0.39	0.46	0.62	0.68	0.68	0.47	0.48	0.49	0.44	0.48	0.64
2019 Q2	0.57	0.58	0.57	0.58	0.82	0.84	0.86	0.62	0.66	0.62	0.62	0.97	7.81
2019 Q3	0.58	0.56	0.58	0.57	0.64	0.6	0.6	0.59	0.59	0.58	0.57	0.56	0.57
2019 Q4	0.39	0.4	0.4	0.4	0.43	0.44	0.44	0.41	0.42	0.43	0.43	0.44	0.46
2020 Q1	0.77	0.83	0.78	0.85	0.95	0.94	1.02	0.97	0.99	1.01	0.82	0.9	0.94
2020 Q2	0.53	0.55	0.55	0.55	0.55	0.61	0.61	0.58	0.62	0.62	0.58	0.55	0.69
2020 Q3	0.54	0.55	0.53	0.55	0.71	0.62	0.63	0.54	0.56	0.52	0.58	0.54	0.54
2020 Q4	0.57	0.56	0.57	0.56	0.61	0.63	0.58	0.55	0.54	0.56	0.64	0.65	0.62
Mean RMSE	0.55	0.56	0.56	0.56	0.66	0.66	0.66	0.59	0.6	0.6	0.59	0.62	1.3

Table 30: Results of models for POLYUS with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Spillover	Basic	Basic	Financial	Overnight and calendar	Basic	Spillover	Macroeconomic	Overnight and calendar	Basic	Basic	Overnight and calendar	Financial
	no lags	with lags	no lags	with lags	with lags	with lags	no lags	no lags	no lags	no lags	no lags	no lags	no lags
2018 Q2	0.82	0.83	0.82	0.84	0.85	0.85	0.9	0.84	0.82	0.83	1.11	1.08	0.96
2018 Q3	0.73	0.71	0.71	0.7	0.75	0.76	0.72	0.63	0.68	0.68	0.69	0.77	0.71
2018 Q4	0.55	0.55	0.54	0.61	0.58	0.59	0.56	0.48	0.48	0.47	0.61	0.6	0.56
2019 Q1	0.62	0.62	0.63	0.61	0.7	0.74	0.71	0.39	0.38	0.37	0.65	0.74	0.68
2019 Q2	0.67	0.68	0.67	0.68	0.82	0.78	0.78	0.45	0.48	0.48	0.72	0.84	0.98
2019 Q3	0.72	0.72	0.72	0.73	0.71	0.75	0.83	0.46	0.47	0.47	0.73	0.71	0.73
2019 Q4	0.5	0.48	0.48	0.5	0.52	0.54	0.62	0.46	0.41	0.41	0.52	0.62	0.54
2020 Q1	0.72	0.76	0.77	0.77	0.88	0.89	0.71	0.72	0.68	0.72	0.72	0.84	1.36
2020 Q2	0.69	0.64	0.63	0.62	0.61	0.62	0.78	0.57	0.58	0.58	0.62	0.67	0.67
2020 Q3	0.55	0.61	0.61	0.57	0.63	0.6	0.66	0.51	0.52	0.53	0.63	0.8	0.6
2020 Q4	0.53	0.57	0.58	0.55	0.65	0.68	0.65	0.41	0.42	0.42	0.66	0.64	0.62
Mean RMSE	0.65	0.65	0.65	0.65	0.7	0.71	0.72	0.54	0.54	0.54	0.7	0.76	0.76

Table 31: Results of models for MAGNIT with 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Basic	Overnight and calendar	Financial	Basic	Overnight and calendar	Financial	Financial	Spillover	Financial	Basic	Overnight and calendar	Macroeconomic
	no lags	with lags	with lags	no lags	with lags	with lags	with lags	no lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.68	0.67	0.68	0.67	0.72	0.75	0.76	0.66	0.68	0.66	0.67	0.68	0.68
2018 Q3	0.56	0.56	0.56	0.55	0.58	0.61	0.58	0.57	0.57	0.55	0.56	0.57	0.7
2018 Q4	0.67	0.66	0.68	0.69	0.71	0.73	0.82	0.71	0.75	0.72	0.67	0.76	0.77
2019 Q1	0.52	0.53	0.52	0.52	0.65	0.64	0.56	0.51	0.51	0.51	0.51	0.63	0.54
2019 Q2	0.5	0.51	0.51	0.53	0.53	0.61	0.64	0.52	0.52	0.57	0.53	0.49	0.49
2019 Q3	0.57	0.56	0.56	0.64	0.8	0.87	0.8	0.64	0.65	0.62	0.63	0.65	0.62
2019 Q4	0.58	0.57	0.57	0.58	0.98	0.92	0.74	0.63	0.62	0.62	0.63	0.63	9.07
2020 Q1	0.73	0.73	0.73	0.74	0.81	0.8	0.91	0.77	0.8	0.83	0.8	0.85	1.78
2020 Q2	0.52	0.51	0.52	0.53	0.59	0.57	0.68	0.56	0.54	0.55	0.59	0.62	0.54
2020 Q3	0.48	0.47	0.47	0.46	0.53	0.47	0.51	0.47	0.47	0.48	0.61	0.48	0.48
2020 Q4	0.49	0.5	0.5	0.51	0.56	0.52	0.52	0.52	0.5	0.5	0.5	0.57	0.63
Mean RMSE	0.57	0.57	0.57	0.58	0.68	0.68	0.68	0.6	0.6	0.6	0.61	0.63	1.48

Table 32: Results of models for SBERBANK with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Financial	Financial	Spillover	Basic	Basic	Overnight and calendar	Overnight and calendar	Financial	Financial	Spillover	Overnight and calendar	Basic	Spillover
	no lags	no lags	no lags	with lags	no lags	no lags	with lags	with lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.7	0.69	0.68	0.7	0.81	0.81	0.82	0.82	0.83	0.83	0.74	0.79	0.76
2018 Q3	0.55	0.57	0.58	0.58	0.62	0.62	0.63	0.6	0.59	0.6	0.59	0.63	0.6
2018 Q4	0.45	0.44	0.45	0.45	0.51	0.51	0.54	0.48	0.46	0.49	0.48	0.49	0.59
2019 Q1	0.35	0.35	0.36	0.35	0.43	0.43	0.45	0.35	0.38	0.36	0.4	0.43	0.41
2019 Q2	0.46	0.46	0.48	0.48	0.5	0.51	0.54	0.46	0.46	0.46	0.45	0.46	9.19
2019 Q3	0.44	0.52	0.52	0.51	0.5	0.52	0.51	0.47	0.47	0.47	0.45	0.61	0.6
2019 Q4	0.38	0.38	0.38	0.39	0.5	0.5	0.43	0.41	0.41	0.41	0.42	0.39	0.83
2020 Q3	0.48	0.48	0.48	0.48	0.55	0.56	0.56	0.5	0.48	0.49	0.49	0.48	0.57
2020 Q4	0.43	0.43	0.43	0.41	0.44	0.43	0.46	0.41	0.41	0.41	0.46	0.47	0.59
Mean RMSE	0.47	0.48	0.48	0.49	0.54	0.54	0.55	0.5	0.5	0.5	0.5	0.53	1.57

Table 33: Results of models for GAZPROM with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Financial	Financial	Basic	Overnight and calendar	Basic	Spillover	Financial	Spillover	Spillover	Overnight and calendar	Basic	Financial
	no lags	no lags	with lags	no lags	with lags	with lags	no lags	no lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.68	0.67	0.67	0.68	0.71	0.74	0.72	0.69	0.68	0.7	0.83	0.72	0.75
2018 Q3	0.41	0.4	0.42	0.41	0.44	0.44	0.43	0.42	0.41	0.39	0.43	0.41	0.43
2018 Q4	0.5	0.49	0.51	0.53	0.52	0.53	0.56	0.52	0.52	0.51	0.67	0.67	0.67
2019 Q1	0.36	0.37	0.35	0.37	0.41	0.43	0.4	0.38	0.38	0.37	0.35	0.43	0.43
2019 Q2	0.91	0.92	0.91	0.91	0.9	0.88	0.89	0.92	0.91	0.93	0.91	0.91	0.96
2019 Q3	0.55	0.55	0.55	0.55	0.56	0.57	0.64	0.53	0.55	0.55	0.62	0.55	0.62
2019 Q4	0.56	0.55	0.55	0.56	0.63	0.65	0.63	0.55	0.56	0.59	0.58	0.76	0.58
2020 Q3	0.38	0.38	0.4	0.39	0.42	0.41	0.46	0.4	0.38	0.38	0.41	0.46	0.74
2020 Q4	0.45	0.45	0.45	0.44	0.52	0.51	0.51	0.45	0.45	0.47	0.54	0.46	0.74
Mean RMSE	0.53	0.53	0.53	0.54	0.57	0.57	0.58	0.54	0.54	0.54	0.59	0.6	0.66

Table 34: Results of models for LUKOIL with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Financial	Financial	Basic	Financial	Overnight and calendar	Basic	Overnight and calendar	Overnight and calendar	Financial	Basic	Basic	Spillover	Overnight and calendar
	no lags	no lags	with lags	with lags	with lags	with lags	no lags	with lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.74	0.74	0.72	0.74	0.79	0.83	0.76	0.77	0.77	0.77	0.79	0.77	0.76
2018 Q3	0.39	0.4	0.42	0.41	0.46	0.47	0.43	0.42	0.39	0.42	0.42	0.4	0.5
2018 Q4	0.44	0.44	0.46	0.45	0.5	0.49	0.51	0.44	0.44	0.44	0.6	0.45	0.44
2019 Q1	0.37	0.38	0.38	0.39	0.52	0.5	0.55	0.38	0.39	0.38	0.44	0.66	0.46
2019 Q2	0.56	0.56	0.59	0.58	0.62	0.63	0.68	0.57	0.59	0.56	0.69	0.64	0.61
2019 Q3	0.59	0.59	0.6	0.59	0.66	0.65	0.64	0.62	0.6	0.62	0.64	0.65	0.63
2019 Q4	0.63	0.63	0.62	0.61	0.72	0.7	0.75	0.7	0.79	0.7	0.71	0.74	0.82
2020 Q3	0.39	0.37	0.38	0.38	0.4	0.4	0.39	0.38	0.36	0.39	0.38	0.38	0.61
2020 Q4	0.47	0.46	0.48	0.47	0.55	0.57	0.54	0.55	0.51	0.54	0.5	0.61	0.54
Mean RMSE	0.51	0.51	0.51	0.51	0.58	0.58	0.58	0.54	0.54	0.54	0.57	0.59	0.6

Table 35: Results of models for NOVATEK with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Financial	Financial	Basic	Overnight and calendar	Basic	Overnight and calendar	Financial	Overnight and calendar	Basic	Spillover	Overnight and calendar	Basic	Financial
	no lags	no lags	with lags	with lags	with lags	with lags	no lags	with lags	with lags	no lags	no lags	no lags	no lags
2018 Q2	0.76	0.77	0.75	0.75	0.82	0.83	0.82	0.79	0.79	0.76	0.76	0.78	0.81
2018 Q3	0.55	0.55	0.56	0.56	0.57	0.57	0.56	0.55	0.56	0.55	0.58	0.56	0.63
2018 Q4	0.51	0.52	0.51	0.51	0.54	0.53	0.54	0.52	0.51	0.53	0.65	0.64	0.58
2019 Q1	0.47	0.47	0.46	0.46	0.47	0.45	0.56	0.46	0.46	0.51	0.49	0.48	1.1
2019 Q2	0.51	0.51	0.51	0.51	0.6	0.6	0.57	0.58	0.57	0.55	0.53	0.53	0.62
2019 Q3	0.5	0.5	0.5	0.5	0.5	0.56	0.71	0.48	0.48	0.51	0.52	0.5	0.51
2019 Q4	0.4	0.45	0.47	0.47	0.47	0.5	0.64	0.47	0.47	0.48	0.44	0.98	0.58
2020 Q3	0.37	0.38	0.38	0.38	0.44	0.44	0.42	0.38	0.38	0.4	0.38	0.43	0.47
2020 Q4	0.43	0.43	0.44	0.44	0.51	0.54	0.45	0.5	0.5	0.46	0.51	0.47	0.49
Mean RMSE	0.5	0.51	0.51	0.51	0.55	0.56	0.59	0.52	0.53	0.53	0.54	0.6	0.64

Table 36: Results of models for ROSNEFT with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Basic	Overnight and calendar	Macroeconomic	Overnight and calendar	Basic	Financial	Financial	Financial	Spillover	Overnight and calendar	Basic	Macroeconomic
	no lags	no lags	no lags	no lags	with lags	with lags	no lags	with lags	no lags	with lags	no lags	no lags	no lags
2018 Q2	0.73	0.75	0.75	0.71	0.74	0.75	0.73	0.74	0.74	0.73	0.73	0.74	0.82
2018 Q3	0.33	0.33	0.34	0.35	0.4	0.39	0.39	0.34	0.37	0.34	0.39	0.34	0.36
2018 Q4	0.43	0.42	0.42	0.43	0.46	0.49	0.47	0.45	0.45	0.45	0.43	0.44	0.42
2019 Q1	0.39	0.44	0.44	0.44	0.51	0.52	0.38	0.41	0.39	0.45	0.5	0.75	0.81
2019 Q2	0.46	0.47	0.46	0.47	0.43	0.44	0.46	0.44	0.45	0.45	0.54	0.64	8.43
2019 Q3	0.53	0.53	0.53	0.56	0.64	0.63	0.74	0.6	0.56	0.59	0.94	0.76	0.72
2019 Q4	0.44	0.44	0.45	0.45	0.51	0.52	0.59	0.45	0.46	0.45	0.56	0.73	0.64
2020 Q3	0.34	0.34	0.34	0.33	0.38	0.4	0.39	0.34	0.34	0.34	0.35	0.39	0.53
2020 Q4	0.47	0.47	0.47	0.47	0.53	0.51	0.51	0.51	0.53	0.51	0.54	0.48	0.59
Mean RMSE	0.46	0.47	0.47	0.47	0.51	0.52	0.52	0.48	0.48	0.48	0.55	0.58	1.48

Table 37: Results of models for NORNICKEL with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Basic	Financial	Basic	Overnight and calendar	Basic	Spillover	Basic	Overnight and calendar	Financial	Basic	Financial	Macroeconomic
	no lags	no lags	no lags	with lags	with lags	with lags	no lags	with lags	with lags	with lags	no lags	no lags	no lags
2018 Q2	0.82	0.91	0.82	0.85	0.85	0.84	0.77	0.84	0.84	0.81	0.8	0.95	0.93
2018 Q3	0.44	0.46	0.42	0.45	0.48	0.48	0.54	0.48	0.47	0.47	0.45	0.44	0.48
2018 Q4	0.38	0.37	0.39	0.38	0.45	0.44	0.48	0.41	0.41	0.39	0.41	0.4	0.46
2019 Q1	0.46	0.48	0.54	0.54	0.56	0.57	0.82	0.53	0.53	0.59	0.55	0.58	0.61
2019 Q2	0.56	0.56	0.56	0.57	0.63	0.64	0.57	0.59	0.6	0.58	0.61	0.63	0.54
2019 Q3	0.46	0.44	0.48	0.46	0.55	0.54	0.55	0.47	0.47	0.52	0.45	0.46	0.48
2019 Q4	0.61	0.62	0.65	0.6	0.62	0.64	0.71	0.62	0.62	0.68	0.64	0.77	0.73
2020 Q3	0.49	0.48	0.48	0.49	0.48	0.48	0.57	0.53	0.54	0.5	0.53	0.45	0.61
2020 Q4	0.47	0.48	0.47	0.47	0.52	0.53	0.54	0.49	0.49	0.51	0.47	0.54	0.5
Mean RMSE	0.52	0.53	0.53	0.54	0.57	0.57	0.62	0.55	0.55	0.56	0.54	0.58	0.59

Table 38: Results of models for POLYMETAL with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Overnight and calendar	Basic	Basic	Overnight and calendar	Basic	Macroeconomic	Macroeconomic	Spillover	Financial	Basic	Overnight and calendar	Financial
	no lags	with lags	with lags	no lags	with lags	with lags	no lags	no lags	no lags	no lags	no lags	no lags	no lags
2018 Q2	0.69	0.7	0.69	0.69	0.74	0.74	0.75	0.72	0.73	0.73	0.71	0.71	0.87
2018 Q3	0.52	0.53	0.52	0.52	0.62	0.63	0.64	0.52	0.56	0.57	0.52	0.54	0.56
2018 Q4	0.48	0.47	0.48	0.61	0.52	0.53	0.58	0.5	0.51	0.5	0.55	0.48	0.57
2019 Q1	0.43	0.44	0.46	0.39	0.68	0.68	0.71	0.48	0.47	0.49	0.44	0.48	0.64
2019 Q2	0.57	0.58	0.58	0.57	0.86	0.84	0.73	0.66	0.62	0.62	0.62	0.97	7.81
2019 Q3	0.58	0.56	0.57	0.58	0.6	0.6	0.65	0.59	0.59	0.58	0.57	0.56	0.57
2019 Q4	0.39	0.4	0.4	0.4	0.44	0.44	0.43	0.42	0.41	0.43	0.43	0.44	0.46
2020 Q3	0.54	0.55	0.55	0.53	0.63	0.62	0.61	0.56	0.54	0.52	0.58	0.54	0.54
2020 Q4	0.57	0.56	0.56	0.57	0.58	0.63	0.6	0.54	0.55	0.56	0.64	0.65	0.62
Mean RMSE	0.53	0.53	0.53	0.54	0.63	0.63	0.63	0.55	0.55	0.56	0.56	0.6	1.41

Table 39: Results of models for POLYUS with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Spillover	Spillover	Financial	Overnight and calendar	Overnight and calendar	Basic	Financial	Macroeconomic	Basic	Overnight and calendar	Basic	Financial	Macroeconomic
	no lags	no lags	no lags	with lags	with lags	with lags	no lags	no lags	no lags	no lags	no lags	no lags	no lags
2018 Q2	0.82	0.81	0.82	0.84	0.85	0.85	0.88	0.84	0.83	0.82	1.11	0.96	1.27
2018 Q3	0.73	0.73	0.72	0.71	0.75	0.76	0.74	0.63	0.68	0.68	0.69	0.71	0.77
2018 Q4	0.55	0.55	0.57	0.55	0.58	0.59	0.56	0.48	0.47	0.48	0.61	0.56	0.56
2019 Q1	0.62	0.62	0.62	0.62	0.7	0.74	0.67	0.39	0.37	0.38	0.65	0.68	0.6
2019 Q2	0.67	0.7	0.67	0.68	0.82	0.78	0.76	0.45	0.48	0.48	0.72	0.98	0.91
2019 Q3	0.72	0.72	0.72	0.72	0.71	0.75	0.84	0.46	0.47	0.47	0.73	0.73	0.71
2019 Q4	0.5	0.48	0.49	0.48	0.52	0.54	0.67	0.46	0.41	0.41	0.52	0.54	0.52
2020 Q3	0.55	0.61	0.6	0.61	0.63	0.6	0.57	0.51	0.53	0.52	0.63	0.6	0.68
2020 Q4	0.53	0.54	0.55	0.57	0.65	0.68	0.65	0.41	0.42	0.42	0.66	0.62	0.66
Mean RMSE	0.63	0.64	0.64	0.64	0.69	0.7	0.71	0.51	0.52	0.52	0.7	0.71	0.74

Table 40: Results of models for MAGNIT with NO Q1 and Q2 of 2020

Period	HAR-RV	Lasso			Random Forest			Gradient Boosting			LSTM		
	Basic	Basic	Overnight and calendar	Basic	Financial	Basic	Financial	Financial	Financial	Spillover	Basic	Overnight and calendar	Macroeconomic
	no lags	no lags	no lags	with lags	with lags	with lags	no lags	with lags	no lags	no lags	no lags	no lags	no lags
2018 Q2	0.68	0.68	0.68	0.67	0.76	0.72	0.78	0.66	0.66	0.68	0.67	0.68	0.68
2018 Q3	0.56	0.56	0.56	0.56	0.58	0.58	0.64	0.55	0.57	0.57	0.56	0.57	0.7
2018 Q4	0.67	0.67	0.67	0.66	0.82	0.71	0.79	0.72	0.71	0.75	0.67	0.76	0.77
2019 Q1	0.52	0.52	0.52	0.53	0.56	0.65	0.56	0.51	0.51	0.51	0.51	0.63	0.54
2019 Q2	0.5	0.5	0.5	0.51	0.64	0.53	0.6	0.57	0.52	0.52	0.53	0.49	0.49
2019 Q3	0.57	0.55	0.55	0.56	0.8	0.8	0.86	0.62	0.64	0.65	0.63	0.65	0.62
2019 Q4	0.58	0.56	0.56	0.57	0.74	0.98	0.75	0.62	0.63	0.62	0.63	0.63	9.07
2020 Q3	0.48	0.48	0.48	0.47	0.51	0.53	0.48	0.48	0.47	0.47	0.61	0.48	0.48
2020 Q4	0.49	0.49	0.49	0.5	0.52	0.56	0.61	0.5	0.52	0.5	0.5	0.57	0.63
Mean RMSE	0.56	0.56	0.56	0.56	0.66	0.67	0.67	0.58	0.58	0.59	0.59	0.61	1.55

F.3 Prediction-Based Importance of Variables

Table 41: Best variables, chosen by Lasso, SBERBANK

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV, is Friday, is after weekend
Frequently chosen	log RV S&P, log RV Brent, growth rate of export, growth rate of housing starts, RGBI

Table 42: Best variables, chosen by Lasso, GAZPROM

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV, is Friday, is after weekend
Frequently chosen	log RV S&P, log RV Brent, growth rate of import, growth rate of CPI, growth rate of housing starts, overnight returns

Table 43: Best variables, chosen by Lasso, LUKOIL

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV, is Friday, is after weekend
Frequently chosen	log RV S&P, log RV Brent, growth rate of import, growth rate of export, growth rate of housing starts, RGBI

Table 44: Best variables, chosen by Lasso, NOVATEK

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV, is Friday, is after weekend
Frequently chosen	Log RV S&P, log RV Brent, growth rate of import, growth rate of export, growth rate of CPI, overnight returns, RGBI

Table 45: Best variables, chosen by Lasso, ROSNEFT

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV, is after weekend
Frequently chosen	Log RV S&P, log RV Brent, growth rate of import, growth rate of CPI, is Friday, overnight returns, RGBI, earning price ratio

Table 46: Best variables, chosen by Lasso, NORNICKEL

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV, is after weekend
Frequently chosen	Log RV S&P, log RV Brent, growth rate of export, growth rate of CPI, growth rate of GDP, is Friday, overnight returns, RGBI, growth rate of earning price ratio, growth rate of housing starts

Table 47: Best variables, chosen by Lasso, POLYMETAL

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV
Frequently chosen	Log RV S&P, log RV Brent, growth rate of export, growth rate of CPI, is after weekend, is after holiday, is Friday, overnight returns, RGBI, growth rate of earning price ratio, growth rate of housing starts

Table 48: Best variables, chosen by Lasso, POLYUS

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV
Frequently chosen	Log RV S&P, log RV Brent, growth rate of export, is after holiday, is after weekend, is Friday, overnight returns, RGBI, High-Low, growth rate of housing starts

Table 49: Best variables, chosen by Lasso, MAGNIT

	Group of variables
Sustainably chosen	Log RV, log weekly RV, log monthly RV, is after weekend, is Friday
Frequently chosen	Log RV S&P, log RV Brent, growth rate of import, growth rate of GDP, overnight returns, RGBI, growth rate of earning price ratio, growth rate of dividend price ratio, growth rate of housing starts

Supplementary material (not for publication)

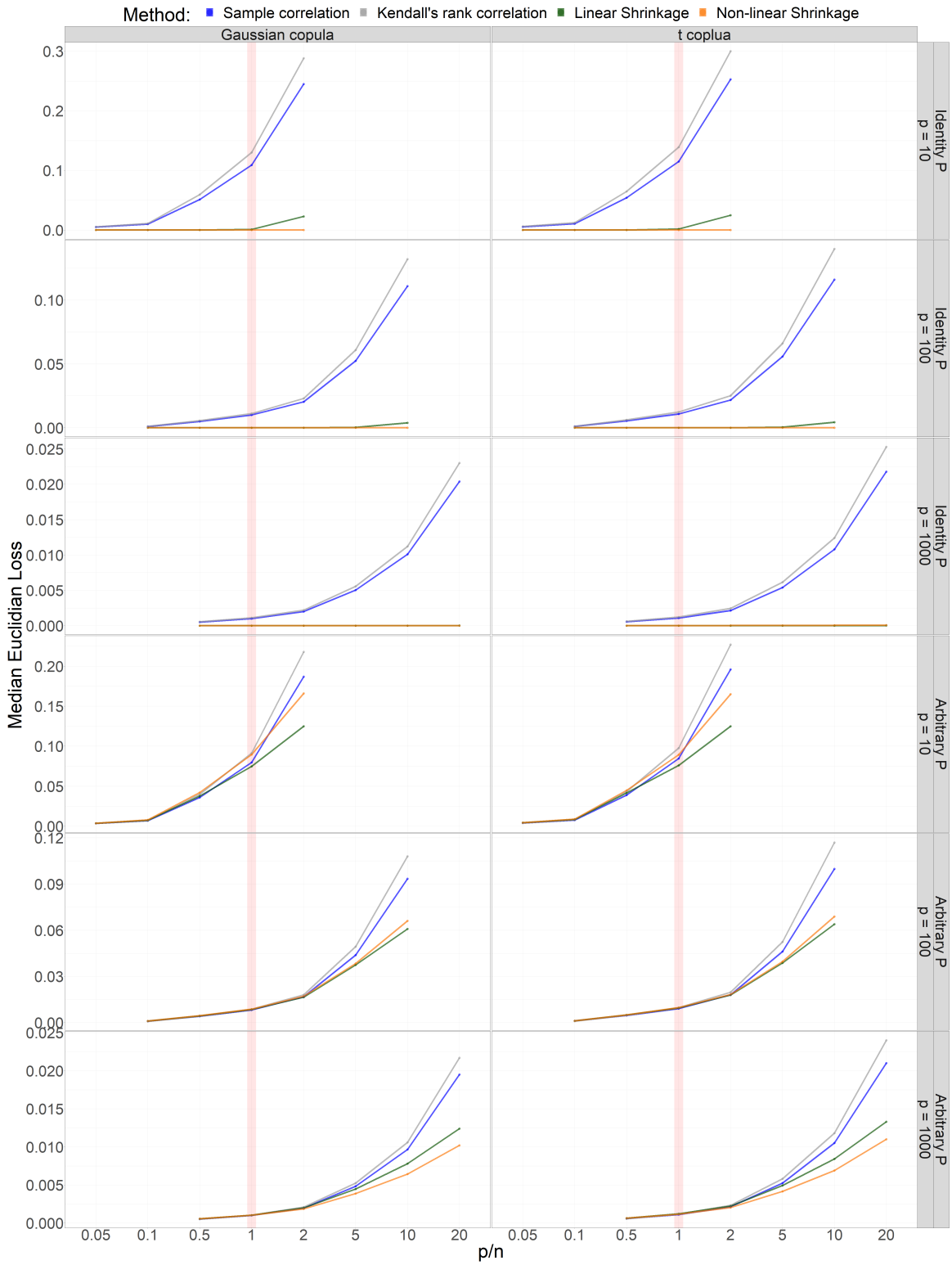


Figure SA1: Median values of Euclidean Loss of different estimators across simulations

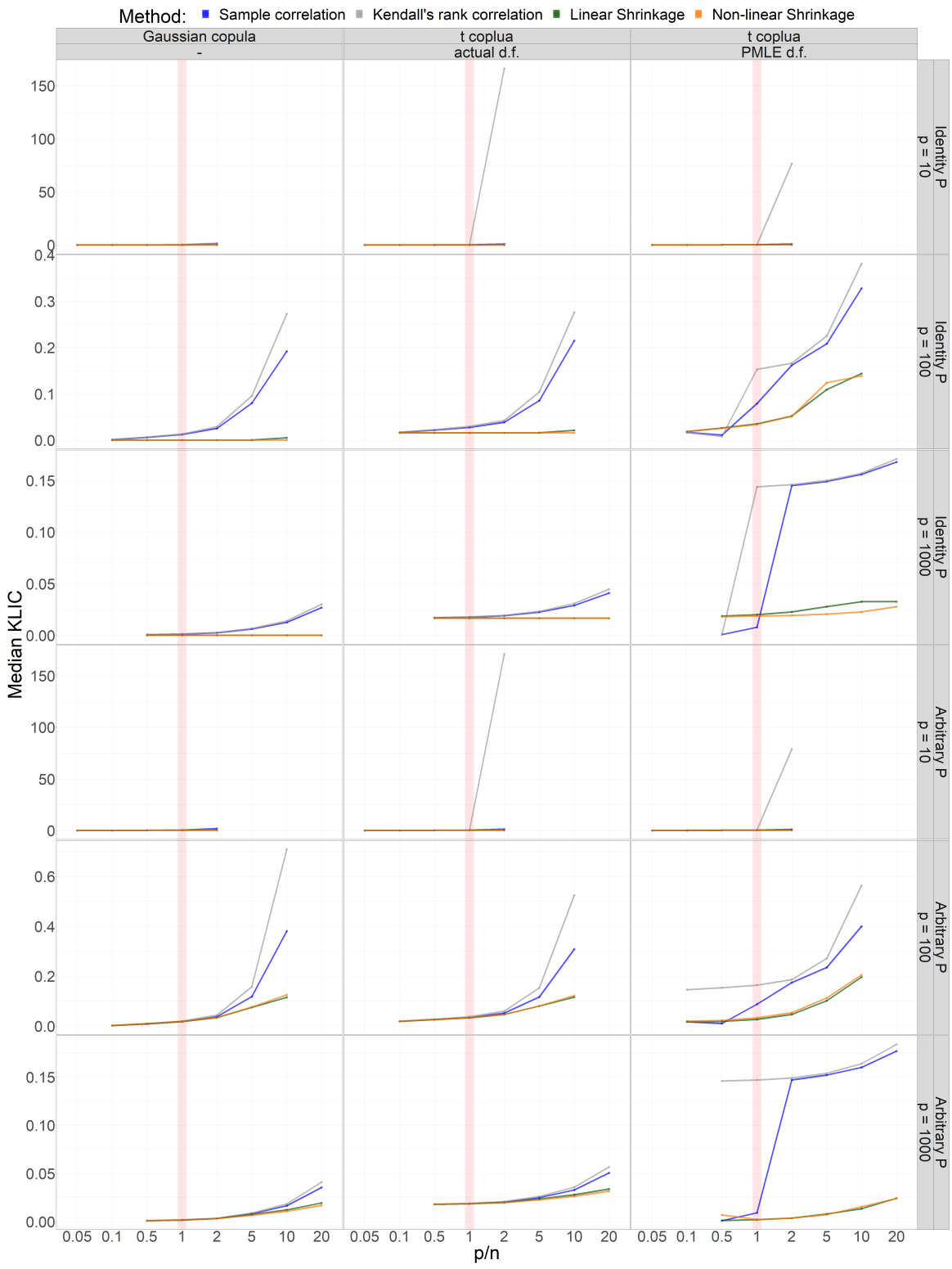


Figure SA2: Median values of KLIC of different estimators across simulations

Table SA1: Precision of estimates for $p = 10$, identity P, Gaussian copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	4.90×10^{-3}	4.99×10^{-3}	1.04×10^{-3}	5.40×10^{-3}	5.51×10^{-3}	1.14×10^{-3}	$\mathbf{0}^{*}$	4.50×10^{-5}	1.31×10^{-4}	2.03×10^{-21}	7.64×10^{-5}	2.00×10^{-4}
0.1	9.84×10^{-3}	9.97×10^{-3}	2.13×10^{-3}	1.10×10^{-2}	1.11×10^{-2}	2.36×10^{-3}	$\mathbf{0}^{*}$	9.95×10^{-5}	3.54×10^{-4}	2.55×10^{-21}	1.47×10^{-4}	4.60×10^{-4}
0.5	5.13×10^{-2}	5.20×10^{-2}	1.04×10^{-2}	5.93×10^{-2}	6.03×10^{-2}	1.18×10^{-2}	$\mathbf{0}^{*}$	9.70×10^{-4}	2.31×10^{-3}	1.18×10^{-20}	7.48×10^{-4}	2.20×10^{-3}
1	1.09×10^{-1}	1.11×10^{-1}	2.00×10^{-2}	1.30×10^{-1}	1.32×10^{-1}	2.33×10^{-2}	1.12×10^{-3}	4.09×10^{-3}	6.88×10^{-3}	$\mathbf{2.21 \times 10^{-8}}$	1.42×10^{-3}	4.79×10^{-3}
2	2.45×10^{-1}	2.51×10^{-1}	3.89×10^{-2}	2.88×10^{-1}	2.95×10^{-1}	4.26×10^{-2}	2.27×10^{-2}	3.00×10^{-2}	2.66×10^{-2}	$\mathbf{3.41 \times 10^{-10}}$	2.70×10^{-3}	1.09×10^{-2}

(b) KLIC

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	5.88×10^{-3}	7.71×10^{-3}	6.65×10^{-3}	6.48×10^{-3}	8.53×10^{-3}	7.38×10^{-3}	$\mathbf{8.88 \times 10^{-21}}$	7.03×10^{-5}	2.42×10^{-4}	6.25×10^{-14}	1.19×10^{-4}	3.88×10^{-4}
0.1	1.32×10^{-2}	1.64×10^{-2}	1.39×10^{-2}	1.49×10^{-2}	1.85×10^{-2}	1.59×10^{-2}	$\mathbf{8.88 \times 10^{-21}}$	1.53×10^{-4}	5.84×10^{-4}	5.68×10^{-14}	2.36×10^{-4}	9.94×10^{-4}
0.5	7.69×10^{-2}	1.10×10^{-1}	1.15×10^{-1}	9.48×10^{-2}	1.37×10^{-1}	1.49×10^{-1}	$\mathbf{8.88 \times 10^{-21}}$	1.46×10^{-3}	4.16×10^{-3}	1.35×10^{-13}	1.15×10^{-3}	4.46×10^{-3}
1	2.04×10^{-1}	3.90×10^{-1}	5.46×10^{-1}	2.83×10^{-1}	*NaN***		1.08×10^{-3}	6.17×10^{-3}	1.30×10^{-2}	$\mathbf{1.47 \times 10^{-7}}$	2.13×10^{-3}	9.14×10^{-3}
2	1.30×10^{0}	*NaN***		1.61×10^{0}	*NaN***		2.98×10^{-2}	4.92×10^{-2}	5.80×10^{-2}	$\mathbf{1.36 \times 10^{-8}}$	3.85×10^{-3}	1.76×10^{-2}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA2: Precision of estimates for $p = 10$, identity P, t copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	5.37×10^{-3}	5.41×10^{-3}	1.11×10^{-3}	6.12×10^{-3}	6.18×10^{-3}	1.27×10^{-3}	$\mathbf{0}^{*}$	5.28×10^{-5}	1.55×10^{-4}	1.92×10^{-21}	1.45×10^{-4}	3.14×10^{-4}
0.1	1.07×10^{-2}	1.08×10^{-2}	2.22×10^{-3}	1.23×10^{-2}	1.25×10^{-2}	2.56×10^{-3}	$\mathbf{0}^{*}$	1.17×10^{-4}	3.24×10^{-4}	3.19×10^{-21}	3.01×10^{-4}	6.29×10^{-4}
0.5	5.43×10^{-2}	5.55×10^{-2}	1.11×10^{-2}	6.46×10^{-2}	6.57×10^{-2}	1.29×10^{-2}	$\mathbf{0}^{*}$	1.11×10^{-3}	2.40×10^{-3}	1.45×10^{-20}	1.30×10^{-3}	2.98×10^{-3}
1	1.15×10^{-1}	1.17×10^{-1}	2.17×10^{-2}	1.39×10^{-1}	1.41×10^{-1}	2.53×10^{-2}	1.52×10^{-3}	4.69×10^{-3}	7.85×10^{-3}	$\mathbf{2.57 \times 10^{-8}}$	2.18×10^{-3}	6.44×10^{-3}
2	2.53×10^{-1}	2.60×10^{-1}	4.33×10^{-2}	3.00×10^{-1}	3.06×10^{-1}	4.84×10^{-2}	2.48×10^{-2}	3.21×10^{-2}	2.84×10^{-2}	$\mathbf{4.72 \times 10^{-10}}$	4.24×10^{-3}	1.33×10^{-2}

(b) KLIC (known true d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	2.34×10^{-2}	2.51×10^{-2}	6.18×10^{-3}	2.43×10^{-2}	2.62×10^{-2}	7.07×10^{-3}	$\mathbf{1.77 \times 10^{-2}}$	1.78×10^{-2}	2.64×10^{-4}	$\mathbf{1.77 \times 10^{-2}}$	1.79×10^{-2}	4.90×10^{-4}
0.1	2.97×10^{-2}	3.33×10^{-2}	1.28×10^{-2}	3.20×10^{-2}	3.58×10^{-2}	1.51×10^{-2}	$\mathbf{1.77 \times 10^{-2}}$	1.78×10^{-2}	5.28×10^{-4}	$\mathbf{1.77 \times 10^{-2}}$	1.81×10^{-2}	1.11×10^{-3}
0.5	8.88×10^{-2}	1.15×10^{-1}	9.63×10^{-2}	1.06×10^{-1}	1.38×10^{-1}	1.20×10^{-1}	$\mathbf{1.77 \times 10^{-2}}$	1.92×10^{-2}	3.92×10^{-3}	$\mathbf{1.77 \times 10^{-2}}$	1.96×10^{-2}	8.19×10^{-3}
1	1.97×10^{-1}	3.04×10^{-1}	3.31×10^{-1}	2.55×10^{-1}	1.08×10^{0}	1.04×10^{1}	1.90×10^{-2}	2.43×10^{-2}	1.32×10^{-2}	$\mathbf{1.77 \times 10^{-2}}$	2.09×10^{-2}	1.19×10^{-2}
2	9.02×10^{-1}	*NaN***		1.66×10^{2}	*NaN***		4.70×10^{-2}	6.35×10^{-2}	5.14×10^{-2}	$\mathbf{1.77 \times 10^{-2}}$	2.33×10^{-2}	2.14×10^{-2}

(c) KLIC (MPLE d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	2.90×10^{-2}	3.01×10^{-2}	9.31×10^{-3}	2.98×10^{-2}	3.09×10^{-2}	9.86×10^{-3}	2.49×10^{-2}	2.55×10^{-2}	7.65×10^{-3}	$\mathbf{2.48 \times 10^{-2}}$	2.56×10^{-2}	7.66×10^{-3}
0.1	3.92×10^{-2}	4.07×10^{-2}	1.69×10^{-2}	4.00×10^{-2}	4.24×10^{-2}	1.86×10^{-2}	$\mathbf{3.11 \times 10^{-2}}$	3.19×10^{-2}	1.30×10^{-2}	3.12×10^{-2}	3.20×10^{-2}	1.30×10^{-2}
0.5	1.04×10^{-1}	1.31×10^{-1}	1.08×10^{-1}	1.48×10^{-1}	1.76×10^{-1}	1.48×10^{-1}	$\mathbf{9.93 \times 10^{-2}}$	1.00×10^{-1}	3.84×10^{-2}	$\mathbf{9.93 \times 10^{-2}}$	9.97×10^{-2}	3.98×10^{-2}
1	3.18×10^{-1}	3.98×10^{-1}	2.57×10^{-1}	3.69×10^{-1}	7.96×10^{-1}	4.79×10^{0}	1.48×10^{-1}	1.53×10^{-1}	1.54×10^{-2}	$\mathbf{1.47 \times 10^{-1}}$	1.48×10^{-1}	1.56×10^{-2}
2	8.78×10^{-1}	*NaN***		7.65×10^{1}	*NaN***		1.77×10^{-1}	1.93×10^{-1}	4.88×10^{-2}	$\mathbf{1.47 \times 10^{-1}}$	1.53×10^{-1}	2.08×10^{-2}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA3: Precision of estimates for $p = 10$, arbitrary P, Gaussian copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	$\mathbf{3.38 \times 10^{-3}}$	3.75×10^{-3}	1.65×10^{-3}	$\mathbf{3.38 \times 10^{-3}}$	3.69×10^{-3}	1.59×10^{-3}	3.68×10^{-3}	4.08×10^{-3}	1.81×10^{-3}	3.79×10^{-3}	4.17×10^{-3}	1.84×10^{-3}
0.1	$\mathbf{6.85 \times 10^{-3}}$	7.48×10^{-3}	3.27×10^{-3}	7.01×10^{-3}	7.61×10^{-3}	3.28×10^{-3}	7.27×10^{-3}	8.17×10^{-3}	3.68×10^{-3}	7.80×10^{-3}	8.53×10^{-3}	3.77×10^{-3}
0.5	$\mathbf{3.58 \times 10^{-2}}$	3.95×10^{-2}	1.76×10^{-2}	3.98×10^{-2}	4.33×10^{-2}	1.86×10^{-2}	3.77×10^{-2}	4.18×10^{-2}	1.93×10^{-2}	4.20×10^{-2}	4.62×10^{-2}	2.13×10^{-2}
1	7.98×10^{-2}	8.56×10^{-2}	3.59×10^{-2}	9.13×10^{-2}	9.75×10^{-2}	3.89×10^{-2}	$\mathbf{7.47 \times 10^{-2}}$	7.92×10^{-2}	3.21×10^{-2}	8.96×10^{-2}	9.41×10^{-2}	3.79×10^{-2}
2	1.87×10^{-1}	2.05×10^{-1}	8.48×10^{-2}	2.18×10^{-1}	2.35×10^{-1}	9.07×10^{-2}	$\mathbf{1.25 \times 10^{-1}}$	1.28×10^{-1}	4.65×10^{-2}	1.66×10^{-1}	1.58×10^{-1}	4.83×10^{-2}

(b) KLIC

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	7.80×10^{-3}	8.75×10^{-3}	5.22×10^{-3}	8.40×10^{-3}	9.58×10^{-3}	5.82×10^{-3}	$\mathbf{7.74 \times 10^{-3}}$	8.63×10^{-3}	5.07×10^{-3}	8.22×10^{-3}	9.23×10^{-3}	5.44×10^{-3}
0.1	1.57×10^{-2}	1.81×10^{-2}	1.09×10^{-2}	1.78×10^{-2}	2.05×10^{-2}	1.26×10^{-2}	$\mathbf{1.49 \times 10^{-2}}$	1.67×10^{-2}	9.64×10^{-3}	1.65×10^{-2}	1.87×10^{-2}	1.08×10^{-2}
0.5	9.66×10^{-2}	1.22×10^{-1}	9.78×10^{-2}	1.24×10^{-1}	1.64×10^{-1}	1.46×10^{-1}	$\mathbf{5.82 \times 10^{-2}}$	6.68×10^{-2}	4.03×10^{-2}	7.07×10^{-2}	8.07×10^{-2}	4.98×10^{-2}
1	2.70×10^{-1}	4.27×10^{-1}	5.27×10^{-1}	4.33×10^{-1}	*NaN***		$\mathbf{9.34 \times 10^{-2}}$	1.08×10^{-1}	6.77×10^{-2}	1.10×10^{-1}	1.28×10^{-1}	8.19×10^{-2}
2	1.92×10^{0}	*NaN***		*NaN***			$\mathbf{1.36 \times 10^{-1}}$	1.66×10^{-1}	1.20×10^{-1}	1.56×10^{-1}	1.64×10^{-1}	1.61×10^{-1}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA4: Precision of estimates for $p = 10$, arbitrary P, t copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	3.96×10^{-3}	4.33×10^{-3}	1.94×10^{-3}	$\mathbf{3.81 \times 10^{-3}}$	4.21×10^{-3}	1.90×10^{-3}	4.40×10^{-3}	4.79×10^{-3}	2.12×10^{-3}	4.47×10^{-3}	4.84×10^{-3}	2.15×10^{-3}
0.1	$\mathbf{7.66 \times 10^{-3}}$	8.45×10^{-3}	3.80×10^{-3}	7.98×10^{-3}	8.58×10^{-3}	3.67×10^{-3}	8.42×10^{-3}	9.41×10^{-3}	4.38×10^{-3}	8.85×10^{-3}	9.69×10^{-3}	4.43×10^{-3}
0.5	$\mathbf{3.88 \times 10^{-2}}$	4.34×10^{-2}	2.13×10^{-2}	4.33×10^{-2}	4.77×10^{-2}	2.22×10^{-2}	4.16×10^{-2}	4.62×10^{-2}	2.27×10^{-2}	4.48×10^{-2}	5.00×10^{-2}	2.42×10^{-2}
1	8.46×10^{-2}	9.27×10^{-2}	4.07×10^{-2}	9.77×10^{-2}	1.06×10^{-1}	4.43×10^{-2}	$\mathbf{7.61 \times 10^{-2}}$	8.29×10^{-2}	3.46×10^{-2}	8.92×10^{-2}	9.53×10^{-2}	3.97×10^{-2}
2	1.96×10^{-1}	2.09×10^{-1}	8.31×10^{-2}	2.27×10^{-1}	2.40×10^{-1}	9.06×10^{-2}	$\mathbf{1.25 \times 10^{-1}}$	1.27×10^{-1}	4.66×10^{-2}	1.65×10^{-1}	1.57×10^{-1}	4.83×10^{-2}

(b) KLIC (known true d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	2.45×10^{-2}	2.56×10^{-2}	5.18×10^{-3}	2.62×10^{-2}	2.75×10^{-2}	6.41×10^{-3}	$\mathbf{2.41 \times 10^{-2}}$	2.52×10^{-2}	4.84×10^{-3}	2.47×10^{-2}	2.58×10^{-2}	5.21×10^{-3}
0.1	3.21×10^{-2}	3.40×10^{-2}	1.01×10^{-2}	3.51×10^{-2}	3.76×10^{-2}	1.25×10^{-2}	$\mathbf{3.07 \times 10^{-2}}$	3.22×10^{-2}	8.88×10^{-3}	3.20×10^{-2}	3.40×10^{-2}	9.94×10^{-3}
0.5	1.05×10^{-1}	1.22×10^{-1}	7.39×10^{-2}	1.27×10^{-1}	1.58×10^{-1}	1.15×10^{-1}	$\mathbf{6.81 \times 10^{-2}}$	7.40×10^{-2}	3.22×10^{-2}	7.77×10^{-2}	8.55×10^{-2}	4.05×10^{-2}
1	2.49×10^{-1}	3.20×10^{-1}	2.71×10^{-1}	3.65×10^{-1}	*NaN***		$\mathbf{1.02 \times 10^{-1}}$	1.10×10^{-1}	5.27×10^{-2}	1.15×10^{-1}	1.28×10^{-1}	6.68×10^{-2}
2	1.26×10^{0}	*NaN***		1.71×10^{2}	*NaN***		$\mathbf{1.40 \times 10^{-1}}$	1.60×10^{-1}	9.04×10^{-2}	1.51×10^{-1}	1.55×10^{-1}	9.75×10^{-2}

(c) KLIC (MPLE d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.05	3.33×10^{-2}	3.45×10^{-2}	1.05×10^{-2}	$\mathbf{2.74 \times 10^{-2}}$	2.88×10^{-2}	1.11×10^{-2}	3.95×10^{-2}	4.06×10^{-2}	1.08×10^{-2}	3.63×10^{-2}	3.75×10^{-2}	1.08×10^{-2}
0.1	4.19×10^{-2}	4.42×10^{-2}	1.74×10^{-2}	$\mathbf{3.73 \times 10^{-2}}$	4.48×10^{-2}	3.05×10^{-2}	5.48×10^{-2}	5.71×10^{-2}	1.80×10^{-2}	4.89×10^{-2}	5.12×10^{-2}	1.81×10^{-2}
0.5	$\mathbf{1.11 \times 10^{-1}}$	1.34×10^{-1}	9.59×10^{-2}	2.50×10^{-1}	2.74×10^{-1}	1.03×10^{-1}	1.62×10^{-1}	1.61×10^{-1}	3.81×10^{-2}	1.56×10^{-1}	1.57×10^{-1}	4.80×10^{-2}
1	3.52×10^{-1}	4.04×10^{-1}	2.06×10^{-1}	4.52×10^{-1}	*NaN***		$\mathbf{2.00 \times 10^{-1}}$	2.10×10^{-1}	4.72×10^{-2}	2.09×10^{-1}	2.22×10^{-1}	6.00×10^{-2}
2	1.10×10^{0}	*NaN***		7.87×10^{1}	*NaN***		$\mathbf{2.31 \times 10^{-1}}$	2.53×10^{-1}	8.00×10^{-2}	2.26×10^{-1}	2.38×10^{-1}	8.40×10^{-2}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA5: Precision of estimates for $p = 100$, identity P, Gaussian copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	1.00×10^{-3}	1.00×10^{-3}	1.97×10^{-5}	1.10×10^{-3}	1.10×10^{-3}	2.16×10^{-5}	$\mathbf{0}^{*}$	1.17×10^{-7}	3.03×10^{-7}	1.30×10^{-21}	1.96×10^{-7}	6.30×10^{-7}
0.5	5.03×10^{-3}	5.03×10^{-3}	9.54×10^{-5}	5.55×10^{-3}	5.55×10^{-3}	1.05×10^{-4}	2.93×10^{-8}	1.05×10^{-6}	2.17×10^{-6}	$\mathbf{1.93 \times 10^{-20}}$	9.96×10^{-7}	2.74×10^{-6}
1	1.01×10^{-2}	1.01×10^{-2}	2.09×10^{-4}	1.12×10^{-2}	1.12×10^{-2}	2.31×10^{-4}	1.27×10^{-6}	4.86×10^{-6}	8.04×10^{-6}	$\mathbf{5.12 \times 10^{-8}}$	2.25×10^{-6}	6.44×10^{-6}
2	2.04×10^{-2}	2.04×10^{-2}	4.06×10^{-4}	2.30×10^{-2}	2.30×10^{-2}	4.52×10^{-4}	2.07×10^{-5}	2.91×10^{-5}	3.02×10^{-5}	$\mathbf{1.61 \times 10^{-10}}$	4.02×10^{-6}	1.16×10^{-5}
5	5.25×10^{-2}	5.26×10^{-2}	9.74×10^{-4}	6.09×10^{-2}	6.10×10^{-2}	1.11×10^{-3}	4.35×10^{-4}	4.55×10^{-4}	1.81×10^{-4}	$\mathbf{1.23 \times 10^{-11}}$	7.91×10^{-6}	2.73×10^{-5}
10	1.11×10^{-1}	1.11×10^{-1}	1.92×10^{-3}	1.32×10^{-1}	1.32×10^{-1}	2.19×10^{-3}	3.99×10^{-3}	4.06×10^{-3}	7.27×10^{-4}	$\mathbf{1.31 \times 10^{-11}}$	1.40×10^{-5}	5.76×10^{-5}

(b) KLIC

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	1.16×10^{-3}	1.50×10^{-3}	1.27×10^{-3}	1.28×10^{-3}	1.65×10^{-3}	1.39×10^{-3}	$\mathbf{1.78 \times 10^{-20}}$	1.53×10^{-7}	1.93×10^{-6}	1.79×10^{-14}	3.03×10^{-7}	3.73×10^{-6}
0.5	5.87×10^{-3}	7.44×10^{-3}	6.31×10^{-3}	6.52×10^{-3}	8.26×10^{-3}	7.00×10^{-3}	$\mathbf{1.78 \times 10^{-20}}$	1.26×10^{-6}	7.09×10^{-6}	6.04×10^{-14}	1.22×10^{-6}	8.48×10^{-6}
1	1.21×10^{-2}	1.61×10^{-2}	1.40×10^{-2}	1.33×10^{-2}	1.81×10^{-2}	1.58×10^{-2}	$\mathbf{1.78 \times 10^{-20}}$	7.29×10^{-6}	2.07×10^{-5}	6.99×10^{-8}	2.79×10^{-6}	1.43×10^{-5}
2	2.54×10^{-2}	3.48×10^{-2}	3.07×10^{-2}	2.91×10^{-2}	4.00×10^{-2}	3.56×10^{-2}	1.83×10^{-5}	4.36×10^{-5}	7.32×10^{-5}	$\mathbf{1.08 \times 10^{-8}}$	6.30×10^{-6}	3.48×10^{-5}
5	7.99×10^{-2}	1.12×10^{-1}	1.13×10^{-1}	9.56×10^{-2}	1.40×10^{-1}	1.56×10^{-1}	5.17×10^{-4}	6.79×10^{-4}	6.18×10^{-4}	$\mathbf{4.85 \times 10^{-10}}$	1.21×10^{-5}	6.87×10^{-5}
10	1.92×10^{-1}	3.59×10^{-1}	5.23×10^{-1}	2.73×10^{-1}	*NaN***		4.85×10^{-3}	5.71×10^{-3}	4.19×10^{-3}	$\mathbf{1.07 \times 10^{-9}}$	2.33×10^{-5}	2.02×10^{-4}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA6: Precision of estimates for $p = 100$, identity P, t copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	1.08×10^{-3}	1.08×10^{-3}	2.28×10^{-5}	1.22×10^{-3}	1.22×10^{-3}	2.65×10^{-5}	$\mathbf{0}^{*}$	1.42×10^{-7}	3.85×10^{-7}	1.67×10^{-21}	3.11×10^{-6}	4.30×10^{-6}
0.5	5.40×10^{-3}	5.40×10^{-3}	1.14×10^{-4}	6.17×10^{-3}	6.17×10^{-3}	1.39×10^{-4}	$\mathbf{5.10 \times 10^{-8}}$	1.37×10^{-6}	2.72×10^{-6}	1.82×10^{-5}	1.82×10^{-5}	1.92×10^{-5}
1	1.08×10^{-2}	1.08×10^{-2}	2.37×10^{-4}	1.24×10^{-2}	1.24×10^{-2}	3.03×10^{-4}	$\mathbf{2.17 \times 10^{-6}}$	5.08×10^{-6}	7.62×10^{-6}	3.85×10^{-5}	4.11×10^{-5}	3.62×10^{-5}
2	2.18×10^{-2}	2.18×10^{-2}	5.43×10^{-4}	2.52×10^{-2}	2.53×10^{-2}	7.31×10^{-4}	$\mathbf{2.27 \times 10^{-5}}$	3.31×10^{-5}	3.52×10^{-5}	6.49×10^{-5}	8.27×10^{-5}	7.28×10^{-5}
5	5.58×10^{-2}	5.59×10^{-2}	1.59×10^{-3}	6.62×10^{-2}	6.63×10^{-2}	2.25×10^{-3}	4.99×10^{-4}	5.29×10^{-4}	2.12×10^{-4}	$\mathbf{1.30 \times 10^{-4}}$	1.90×10^{-4}	2.08×10^{-4}
10	1.16×10^{-1}	1.17×10^{-1}	3.55×10^{-3}	1.40×10^{-1}	1.41×10^{-1}	5.05×10^{-3}	4.44×10^{-3}	4.54×10^{-3}	7.87×10^{-4}	$\mathbf{1.38 \times 10^{-4}}$	3.12×10^{-4}	4.65×10^{-4}

(b) KLIC (known true d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	1.68×10^{-2}	1.71×10^{-2}	1.27×10^{-3}	1.69×10^{-2}	1.73×10^{-2}	1.44×10^{-3}	$\mathbf{1.56 \times 10^{-2}}$	1.57×10^{-2}	2.82×10^{-6}	$\mathbf{1.56 \times 10^{-2}}$	1.57×10^{-2}	1.64×10^{-5}
0.5	2.17×10^{-2}	2.35×10^{-2}	6.43×10^{-3}	2.26×10^{-2}	2.47×10^{-2}	7.45×10^{-3}	$\mathbf{1.56 \times 10^{-2}}$	1.57×10^{-2}	9.75×10^{-6}	$\mathbf{1.56 \times 10^{-2}}$	1.57×10^{-2}	5.10×10^{-5}
1	2.76×10^{-2}	3.13×10^{-2}	1.40×10^{-2}	2.95×10^{-2}	3.39×10^{-2}	1.65×10^{-2}	$\mathbf{1.56 \times 10^{-2}}$	1.57×10^{-2}	2.42×10^{-5}	1.57×10^{-2}	1.57×10^{-2}	9.67×10^{-5}
2	3.87×10^{-2}	4.80×10^{-2}	2.99×10^{-2}	4.24×10^{-2}	5.39×10^{-2}	3.55×10^{-2}	$\mathbf{1.57 \times 10^{-2}}$	1.57×10^{-2}	9.75×10^{-5}	$\mathbf{1.57 \times 10^{-2}}$	1.58×10^{-2}	2.06×10^{-4}
5	8.55×10^{-2}	1.16×10^{-1}	1.03×10^{-1}	1.04×10^{-1}	1.41×10^{-1}	1.32×10^{-1}	1.62×10^{-2}	1.64×10^{-2}	6.47×10^{-4}	$\mathbf{1.57 \times 10^{-2}}$	1.59×10^{-2}	5.10×10^{-4}
10	2.15×10^{-1}	3.21×10^{-1}	3.40×10^{-1}	2.76×10^{-1}	*NaN***		2.12×10^{-2}	2.20×10^{-2}	4.44×10^{-3}	$\mathbf{1.58 \times 10^{-2}}$	1.61×10^{-2}	1.02×10^{-3}

(c) KLIC (MPLE d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	1.68×10^{-2}	1.70×10^{-2}	2.03×10^{-3}	$\mathbf{1.67 \times 10^{-2}}$	1.69×10^{-2}	2.13×10^{-3}	1.86×10^{-2}	1.86×10^{-2}	1.71×10^{-3}	1.86×10^{-2}	1.86×10^{-2}	1.70×10^{-3}
0.5	1.08×10^{-2}	1.22×10^{-2}	7.52×10^{-3}	$\mathbf{7.85 \times 10^{-3}}$	1.20×10^{-2}	1.80×10^{-2}	2.64×10^{-2}	2.66×10^{-2}	4.23×10^{-3}	2.60×10^{-2}	2.62×10^{-2}	4.17×10^{-3}
1	7.92×10^{-2}	8.26×10^{-2}	2.62×10^{-2}	1.53×10^{-1}	1.57×10^{-1}	1.65×10^{-2}	3.51×10^{-2}	3.56×10^{-2}	6.48×10^{-3}	$\mathbf{3.39 \times 10^{-2}}$	3.46×10^{-2}	6.18×10^{-3}
2	1.62×10^{-1}	1.71×10^{-1}	2.92×10^{-2}	1.66×10^{-1}	1.77×10^{-1}	3.44×10^{-2}	5.24×10^{-2}	5.35×10^{-2}	9.99×10^{-3}	$\mathbf{5.13 \times 10^{-2}}$	5.23×10^{-2}	8.99×10^{-3}
5	2.08×10^{-1}	2.34×10^{-1}	9.15×10^{-2}	2.25×10^{-1}	2.57×10^{-1}	1.14×10^{-1}	$\mathbf{1.09 \times 10^{-1}}$	1.09×10^{-1}	1.48×10^{-2}	1.24×10^{-1}	1.22×10^{-1}	1.18×10^{-2}
10	3.28×10^{-1}	4.06×10^{-1}	2.64×10^{-1}	3.81×10^{-1}	*NaN***		1.44×10^{-1}	1.45×10^{-1}	4.53×10^{-3}	$\mathbf{1.39 \times 10^{-1}}$	1.39×10^{-1}	1.04×10^{-3}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA7: Precision of estimates for $p = 100$, arbitrary P, Gaussian copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	9.33×10^{-4}	9.62×10^{-4}	1.97×10^{-4}	$\mathbf{8.46 \times 10^{-4}}$	8.70×10^{-4}	1.63×10^{-4}	9.95×10^{-4}	1.02×10^{-3}	2.14×10^{-4}	1.00×10^{-3}	1.03×10^{-3}	2.17×10^{-4}
0.5	$\mathbf{4.13 \times 10^{-3}}$	4.27×10^{-3}	8.68×10^{-4}	4.20×10^{-3}	4.36×10^{-3}	8.44×10^{-4}	4.36×10^{-3}	4.53×10^{-3}	9.86×10^{-4}	4.47×10^{-3}	4.64×10^{-3}	1.07×10^{-3}
1	$\mathbf{8.18 \times 10^{-3}}$	8.49×10^{-3}	1.68×10^{-3}	8.70×10^{-3}	8.92×10^{-3}	1.69×10^{-3}	8.48×10^{-3}	8.77×10^{-3}	1.92×10^{-3}	8.76×10^{-3}	9.09×10^{-3}	2.01×10^{-3}
2	1.67×10^{-2}	1.72×10^{-2}	3.30×10^{-3}	1.80×10^{-2}	1.85×10^{-2}	3.35×10^{-3}	$\mathbf{1.65 \times 10^{-2}}$	1.71×10^{-2}	3.83×10^{-3}	1.72×10^{-2}	1.76×10^{-2}	3.75×10^{-3}
5	4.39×10^{-2}	4.52×10^{-2}	8.63×10^{-3}	4.93×10^{-2}	5.05×10^{-2}	8.99×10^{-3}	$\mathbf{3.74 \times 10^{-2}}$	3.84×10^{-2}	8.73×10^{-3}	3.84×10^{-2}	3.98×10^{-2}	9.26×10^{-3}
10	9.34×10^{-2}	9.60×10^{-2}	1.77×10^{-2}	1.08×10^{-1}	1.11×10^{-1}	1.89×10^{-2}	$\mathbf{6.08 \times 10^{-2}}$	6.21×10^{-2}	1.27×10^{-2}	6.61×10^{-2}	6.73×10^{-2}	1.41×10^{-2}

(b) KLIC

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	2.14×10^{-3}	2.20×10^{-3}	6.82×10^{-4}	$\mathbf{1.80 \times 10^{-3}}$	1.87×10^{-3}	5.64×10^{-4}	2.38×10^{-3}	2.45×10^{-3}	7.57×10^{-4}	2.41×10^{-3}	2.48×10^{-3}	7.61×10^{-4}
0.5	$\mathbf{8.81 \times 10^{-3}}$	9.16×10^{-3}	2.70×10^{-3}	9.29×10^{-3}	9.65×10^{-3}	2.91×10^{-3}	9.72×10^{-3}	1.01×10^{-2}	2.99×10^{-3}	9.87×10^{-3}	1.02×10^{-2}	2.90×10^{-3}
1	$\mathbf{1.80 \times 10^{-2}}$	1.86×10^{-2}	5.59×10^{-3}	1.96×10^{-2}	2.07×10^{-2}	6.47×10^{-3}	1.84×10^{-2}	1.92×10^{-2}	5.71×10^{-3}	1.86×10^{-2}	1.94×10^{-2}	5.48×10^{-3}
2	3.74×10^{-2}	3.91×10^{-2}	1.25×10^{-2}	4.38×10^{-2}	4.60×10^{-2}	1.55×10^{-2}	$\mathbf{3.42 \times 10^{-2}}$	3.56×10^{-2}	1.03×10^{-2}	3.44×10^{-2}	3.56×10^{-2}	9.79×10^{-3}
5	1.18×10^{-1}	1.25×10^{-1}	4.75×10^{-2}	1.58×10^{-1}	1.73×10^{-1}	8.04×10^{-2}	$\mathbf{7.46 \times 10^{-2}}$	7.66×10^{-2}	2.05×10^{-2}	7.61×10^{-2}	7.86×10^{-2}	2.08×10^{-2}
10	3.81×10^{-1}	*NaN***		7.09×10^{-1}	*NaN***		$\mathbf{1.15 \times 10^{-1}}$	1.17×10^{-1}	2.71×10^{-2}	1.25×10^{-1}	1.26×10^{-1}	2.55×10^{-2}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA8: Precision of estimates for $p = 100$, arbitrary P, t copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	1.12×10^{-3}	1.17×10^{-3}	2.30×10^{-4}	$\mathbf{9.42 \times 10^{-4}}$	9.71×10^{-4}	1.76×10^{-4}	1.22×10^{-3}	1.26×10^{-3}	2.50×10^{-4}	1.23×10^{-3}	1.27×10^{-3}	2.50×10^{-4}
0.5	$\mathbf{4.66 \times 10^{-3}}$	4.80×10^{-3}	9.17×10^{-4}	4.79×10^{-3}	4.92×10^{-3}	8.93×10^{-4}	5.04×10^{-3}	5.17×10^{-3}	1.05×10^{-3}	5.06×10^{-3}	5.24×10^{-3}	1.12×10^{-3}
1	$\mathbf{9.12 \times 10^{-3}}$	9.38×10^{-3}	1.81×10^{-3}	9.67×10^{-3}	9.98×10^{-3}	1.81×10^{-3}	9.49×10^{-3}	9.84×10^{-3}	2.13×10^{-3}	9.68×10^{-3}	1.00×10^{-2}	2.14×10^{-3}
2	1.81×10^{-2}	1.86×10^{-2}	3.52×10^{-3}	1.97×10^{-2}	2.03×10^{-2}	3.65×10^{-3}	$\mathbf{1.79 \times 10^{-2}}$	1.85×10^{-2}	4.05×10^{-3}	1.83×10^{-2}	1.87×10^{-2}	3.82×10^{-3}
5	4.62×10^{-2}	4.77×10^{-2}	8.65×10^{-3}	5.24×10^{-2}	5.41×10^{-2}	9.18×10^{-3}	$\mathbf{3.87 \times 10^{-2}}$	4.00×10^{-2}	8.64×10^{-3}	3.96×10^{-2}	4.11×10^{-2}	9.12×10^{-3}
10	9.98×10^{-2}	1.03×10^{-1}	1.97×10^{-2}	1.17×10^{-1}	1.20×10^{-1}	2.13×10^{-2}	$\mathbf{6.39 \times 10^{-2}}$	6.50×10^{-2}	1.32×10^{-2}	6.89×10^{-2}	6.96×10^{-2}	1.41×10^{-2}

(b) KLIC (known true d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	$\mathbf{1.89 \times 10^{-2}}$	1.90×10^{-2}	5.01×10^{-4}	1.95×10^{-2}	1.96×10^{-2}	7.25×10^{-4}	1.90×10^{-2}	1.90×10^{-2}	5.30×10^{-4}	1.90×10^{-2}	1.91×10^{-2}	5.32×10^{-4}
0.5	$\mathbf{2.56 \times 10^{-2}}$	2.58×10^{-2}	2.52×10^{-3}	2.73×10^{-2}	2.76×10^{-2}	3.25×10^{-3}	2.57×10^{-2}	2.59×10^{-2}	2.54×10^{-3}	2.57×10^{-2}	2.59×10^{-2}	2.47×10^{-3}
1	3.39×10^{-2}	3.47×10^{-2}	5.39×10^{-3}	3.74×10^{-2}	3.84×10^{-2}	6.79×10^{-3}	$\mathbf{3.28 \times 10^{-2}}$	3.36×10^{-2}	5.00×10^{-3}	$\mathbf{3.28 \times 10^{-2}}$	3.35×10^{-2}	4.74×10^{-3}
2	5.20×10^{-2}	5.32×10^{-2}	1.12×10^{-2}	6.02×10^{-2}	6.16×10^{-2}	1.41×10^{-2}	4.68×10^{-2}	4.76×10^{-2}	8.77×10^{-3}	$\mathbf{4.63 \times 10^{-2}}$	4.71×10^{-2}	8.37×10^{-3}
5	1.17×10^{-1}	1.23×10^{-1}	3.70×10^{-2}	1.53×10^{-1}	1.64×10^{-1}	6.28×10^{-2}	$\mathbf{8.00 \times 10^{-2}}$	8.13×10^{-2}	1.68×10^{-2}	8.07×10^{-2}	8.21×10^{-2}	1.76×10^{-2}
10	3.08×10^{-1}	*NaN***		5.24×10^{-1}	*NaN***		$\mathbf{1.16 \times 10^{-1}}$	1.17×10^{-1}	2.31×10^{-2}	1.22×10^{-1}	1.23×10^{-1}	2.21×10^{-2}

(c) KLIC (MPLE d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.1	$\mathbf{1.74 \times 10^{-2}}$	1.75×10^{-2}	2.11×10^{-3}	1.46×10^{-1}	1.44×10^{-1}	1.89×10^{-2}	1.84×10^{-2}	1.86×10^{-2}	2.12×10^{-3}	1.92×10^{-2}	1.95×10^{-2}	2.15×10^{-3}
0.5	$\mathbf{1.03 \times 10^{-2}}$	1.05×10^{-2}	3.10×10^{-3}	1.54×10^{-1}	1.54×10^{-1}	5.45×10^{-3}	1.91×10^{-2}	1.93×10^{-2}	3.64×10^{-3}	2.30×10^{-2}	2.32×10^{-2}	3.87×10^{-3}
1	8.76×10^{-2}	9.31×10^{-2}	2.78×10^{-2}	1.64×10^{-1}	1.65×10^{-1}	9.15×10^{-3}	$\mathbf{2.67 \times 10^{-2}}$	2.75×10^{-2}	5.98×10^{-3}	3.27×10^{-2}	3.32×10^{-2}	5.97×10^{-3}
2	1.74×10^{-1}	1.75×10^{-1}	1.25×10^{-2}	1.86×10^{-1}	1.88×10^{-1}	1.60×10^{-2}	$\mathbf{4.67 \times 10^{-2}}$	4.74×10^{-2}	9.57×10^{-3}	5.38×10^{-2}	5.42×10^{-2}	9.49×10^{-3}
5	2.35×10^{-1}	2.40×10^{-1}	3.46×10^{-2}	2.71×10^{-1}	2.79×10^{-1}	5.33×10^{-2}	$\mathbf{1.02 \times 10^{-1}}$	1.04×10^{-1}	1.82×10^{-2}	1.13×10^{-1}	1.15×10^{-1}	1.90×10^{-2}
10	4.00×10^{-1}	*NaN***		5.64×10^{-1}	*NaN***		$\mathbf{1.96 \times 10^{-1}}$	1.96×10^{-1}	2.14×10^{-2}	2.05×10^{-1}	2.06×10^{-1}	1.69×10^{-2}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA9: Precision of estimates for $p = 1000$, identity P, Gaussian copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	5.00×10^{-4}	5.00×10^{-4}	1.01×10^{-6}	5.49×10^{-4}	5.49×10^{-4}	1.10×10^{-6}	1.18×10^{-11}	1.18×10^{-9}	2.50×10^{-9}	$\mathbf{5.78 \times 10^{-18}}$	6.07×10^{-10}	1.84×10^{-9}
1	1.00×10^{-3}	1.00×10^{-3}	2.04×10^{-6}	1.10×10^{-3}	1.10×10^{-3}	2.24×10^{-6}	1.17×10^{-9}	4.57×10^{-9}	7.68×10^{-9}	$\mathbf{2.52 \times 10^{-8}}$	2.15×10^{-8}	1.55×10^{-8}
2	2.00×10^{-3}	2.00×10^{-3}	4.00×10^{-6}	2.20×10^{-3}	2.20×10^{-3}	4.38×10^{-6}	2.03×10^{-8}	2.88×10^{-8}	2.88×10^{-8}	$\mathbf{2.92 \times 10^{-11}}$	2.25×10^{-9}	6.99×10^{-9}
5	5.03×10^{-3}	5.03×10^{-3}	1.01×10^{-5}	5.55×10^{-3}	5.55×10^{-3}	1.12×10^{-5}	4.24×10^{-7}	4.44×10^{-7}	1.88×10^{-7}	$\mathbf{2.72 \times 10^{-10}}$	5.30×10^{-9}	1.51×10^{-8}
10	1.01×10^{-2}	1.01×10^{-2}	1.93×10^{-5}	1.12×10^{-2}	1.12×10^{-2}	2.13×10^{-5}	3.73×10^{-6}	3.78×10^{-6}	7.45×10^{-7}	$\mathbf{2.34 \times 10^{-9}}$	1.12×10^{-8}	2.51×10^{-8}
20	2.04×10^{-2}	2.04×10^{-2}	4.08×10^{-5}	2.30×10^{-2}	2.30×10^{-2}	4.61×10^{-5}	3.13×10^{-5}	3.15×10^{-5}	3.21×10^{-6}	$\mathbf{1.36 \times 10^{-8}}$	3.67×10^{-8}	7.47×10^{-8}

(b) KLIC

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	5.95×10^{-4}	7.65×10^{-4}	6.22×10^{-4}	6.57×10^{-4}	8.40×10^{-4}	6.83×10^{-4}	$\mathbf{0}^*$	1.52×10^{-9}	1.53×10^{-7}	$\mathbf{0}^*$	$\mathbf{0}^*$	$\mathbf{0}^*$
1	1.18×10^{-3}	1.49×10^{-3}	1.24×10^{-3}	1.29×10^{-3}	1.64×10^{-3}	1.36×10^{-3}	$\mathbf{0}^*$	$\mathbf{0}^*$	$\mathbf{0}^*$	$\mathbf{0}^*$	2.60×10^{-8}	6.14×10^{-7}
2	2.44×10^{-3}	3.09×10^{-3}	2.47×10^{-3}	2.72×10^{-3}	3.41×10^{-3}	2.72×10^{-3}	$\mathbf{0}^*$	3.86×10^{-8}	7.43×10^{-7}	$\mathbf{0}^*$	$\mathbf{0}^*$	$\mathbf{0}^*$
5	6.05×10^{-3}	7.73×10^{-3}	6.31×10^{-3}	6.68×10^{-3}	8.56×10^{-3}	6.97×10^{-3}	1.39×10^{-7}	4.91×10^{-7}	2.92×10^{-6}	$\mathbf{0}^*$	5.37×10^{-9}	3.32×10^{-7}
10	1.25×10^{-2}	1.64×10^{-2}	1.43×10^{-2}	1.41×10^{-2}	1.83×10^{-2}	1.58×10^{-2}	4.02×10^{-6}	6.22×10^{-6}	1.02×10^{-5}	$\mathbf{7.24 \times 10^{-10}}$	2.38×10^{-8}	4.57×10^{-7}
20	2.69×10^{-2}	3.42×10^{-2}	2.85×10^{-2}	3.01×10^{-2}	3.88×10^{-2}	3.23×10^{-2}	3.79×10^{-5}	4.65×10^{-5}	4.10×10^{-5}	$\mathbf{1.65 \times 10^{-10}}$	7.03×10^{-8}	8.09×10^{-7}

Notes. In each row minimal median value is in **bold**; *****: value numerically indistinguishable from zero; ******: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA10: Precision of estimates for $p = 1000$, identity P, t copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	5.38×10^{-4}	5.38×10^{-4}	1.66×10^{-6}	6.12×10^{-4}	6.12×10^{-4}	2.43×10^{-6}	$\mathbf{8.39 \times 10^{-11}}$	1.47×10^{-9}	2.87×10^{-9}	2.46×10^{-6}	1.57×10^{-6}	1.33×10^{-6}
1	1.08×10^{-3}	1.08×10^{-3}	4.03×10^{-6}	1.22×10^{-3}	1.22×10^{-3}	6.08×10^{-6}	$\mathbf{2.63 \times 10^{-9}}$	6.63×10^{-9}	1.01×10^{-8}	5.11×10^{-6}	3.92×10^{-6}	2.37×10^{-6}
2	2.15×10^{-3}	2.15×10^{-3}	1.05×10^{-5}	2.45×10^{-3}	2.45×10^{-3}	1.69×10^{-5}	$\mathbf{2.66 \times 10^{-8}}$	3.36×10^{-8}	3.10×10^{-8}	9.93×10^{-6}	9.43×10^{-6}	2.75×10^{-6}
5	5.40×10^{-3}	5.40×10^{-3}	3.84×10^{-5}	6.16×10^{-3}	6.16×10^{-3}	6.29×10^{-5}	$\mathbf{4.88 \times 10^{-7}}$	5.14×10^{-7}	2.05×10^{-7}	2.37×10^{-5}	2.42×10^{-5}	5.59×10^{-6}
10	1.08×10^{-2}	1.08×10^{-2}	1.10×10^{-4}	1.24×10^{-2}	1.24×10^{-2}	1.81×10^{-4}	$\mathbf{4.30 \times 10^{-6}}$	4.34×10^{-6}	8.43×10^{-7}	4.54×10^{-5}	4.79×10^{-5}	1.63×10^{-5}
20	2.18×10^{-2}	2.18×10^{-2}	2.90×10^{-4}	2.53×10^{-2}	2.53×10^{-2}	4.80×10^{-4}	$\mathbf{3.60 \times 10^{-5}}$	3.61×10^{-5}	3.52×10^{-6}	8.77×10^{-5}	9.41×10^{-5}	4.34×10^{-5}

(b) KLIC (known true d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	1.71×10^{-2}	1.72×10^{-2}	6.16×10^{-4}	1.72×10^{-2}	1.73×10^{-2}	6.99×10^{-4}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	1.87×10^{-7}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	6.75×10^{-6}
1	1.76×10^{-2}	1.79×10^{-2}	1.19×10^{-3}	1.78×10^{-2}	1.81×10^{-2}	1.35×10^{-3}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	3.54×10^{-7}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	1.30×10^{-5}
2	1.90×10^{-2}	1.96×10^{-2}	2.52×10^{-3}	1.93×10^{-2}	2.01×10^{-2}	2.89×10^{-3}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	9.99×10^{-7}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	2.10×10^{-5}
5	2.24×10^{-2}	2.43×10^{-2}	7.00×10^{-3}	2.32×10^{-2}	2.55×10^{-2}	8.00×10^{-3}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	3.74×10^{-6}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	4.09×10^{-5}
10	2.90×10^{-2}	3.18×10^{-2}	1.23×10^{-2}	3.08×10^{-2}	3.41×10^{-2}	1.43×10^{-2}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	1.20×10^{-5}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	8.70×10^{-5}
20	4.09×10^{-2}	4.95×10^{-2}	2.97×10^{-2}	4.47×10^{-2}	5.56×10^{-2}	3.57×10^{-2}	$\mathbf{1.65 \times 10^{-2}}$	1.65×10^{-2}	5.07×10^{-5}	1.66×10^{-2}	1.66×10^{-2}	1.31×10^{-4}

(c) KLIC (MPLE d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	$\mathbf{6.65 \times 10^{-4}}$	8.27×10^{-4}	6.77×10^{-4}	7.58×10^{-4}	9.41×10^{-4}	7.70×10^{-4}	1.86×10^{-2}	1.86×10^{-2}	1.26×10^{-3}	1.81×10^{-2}	1.81×10^{-2}	1.30×10^{-3}
1	$\mathbf{7.72 \times 10^{-3}}$	7.78×10^{-3}	2.26×10^{-3}	1.44×10^{-1}	1.44×10^{-1}	1.38×10^{-3}	2.00×10^{-2}	2.00×10^{-2}	1.65×10^{-3}	1.86×10^{-2}	1.88×10^{-2}	1.82×10^{-3}
2	1.45×10^{-1}	1.46×10^{-1}	2.58×10^{-3}	1.46×10^{-1}	1.46×10^{-1}	2.95×10^{-3}	2.26×10^{-2}	2.27×10^{-2}	2.61×10^{-3}	$\mathbf{1.93 \times 10^{-2}}$	1.95×10^{-2}	2.36×10^{-3}
5	1.49×10^{-1}	1.51×10^{-1}	7.10×10^{-3}	1.50×10^{-1}	1.52×10^{-1}	8.09×10^{-3}	2.79×10^{-2}	2.81×10^{-2}	3.92×10^{-3}	$\mathbf{2.06 \times 10^{-2}}$	2.09×10^{-2}	2.64×10^{-3}
10	1.56×10^{-1}	1.58×10^{-1}	1.24×10^{-2}	1.57×10^{-1}	1.61×10^{-1}	1.44×10^{-2}	3.25×10^{-2}	3.33×10^{-2}	5.43×10^{-3}	$\mathbf{2.27 \times 10^{-2}}$	2.30×10^{-2}	2.66×10^{-3}
20	1.68×10^{-1}	1.76×10^{-1}	2.90×10^{-2}	1.71×10^{-1}	1.82×10^{-1}	3.45×10^{-2}	3.26×10^{-2}	3.34×10^{-2}	6.07×10^{-3}	$\mathbf{2.78 \times 10^{-2}}$	2.81×10^{-2}	2.08×10^{-3}

Notes. In each row minimal median value is in **bold**; *****: value numerically indistinguishable from zero; ******: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA11: Precision of estimates for $p = 1000$, arbitrary P, Gaussian copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	5.22×10^{-4}	5.29×10^{-4}	3.89×10^{-5}	$\mathbf{5.17 \times 10^{-4}}$	5.20×10^{-4}	2.38×10^{-5}	5.54×10^{-4}	5.63×10^{-4}	5.06×10^{-5}	5.56×10^{-4}	5.63×10^{-4}	4.97×10^{-5}
1	$\mathbf{9.96 \times 10^{-4}}$	1.01×10^{-3}	6.53×10^{-5}	1.03×10^{-3}	1.04×10^{-3}	4.80×10^{-5}	1.04×10^{-3}	1.06×10^{-3}	9.69×10^{-5}	1.02×10^{-3}	1.03×10^{-3}	9.00×10^{-5}
2	1.95×10^{-3}	1.97×10^{-3}	1.09×10^{-4}	2.08×10^{-3}	2.09×10^{-3}	9.47×10^{-5}	1.98×10^{-3}	2.00×10^{-3}	1.82×10^{-4}	$\mathbf{1.85 \times 10^{-3}}$	1.87×10^{-3}	1.54×10^{-4}
5	4.83×10^{-3}	4.86×10^{-3}	2.56×10^{-4}	5.24×10^{-3}	5.26×10^{-3}	2.31×10^{-4}	4.47×10^{-3}	4.55×10^{-3}	4.96×10^{-4}	$\mathbf{3.86 \times 10^{-3}}$	3.92×10^{-3}	3.77×10^{-4}
10	9.65×10^{-3}	9.71×10^{-3}	4.69×10^{-4}	1.06×10^{-2}	1.06×10^{-2}	4.56×10^{-4}	7.80×10^{-3}	7.89×10^{-3}	8.98×10^{-4}	$\mathbf{6.42 \times 10^{-3}}$	6.51×10^{-3}	6.94×10^{-4}
20	1.95×10^{-2}	1.96×10^{-2}	9.45×10^{-4}	2.17×10^{-2}	2.18×10^{-2}	9.48×10^{-4}	1.24×10^{-2}	1.26×10^{-2}	1.45×10^{-3}	$\mathbf{1.02 \times 10^{-2}}$	1.05×10^{-2}	1.33×10^{-3}

(b) KLIC

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	8.54×10^{-4}	8.62×10^{-4}	1.50×10^{-4}	$\mathbf{8.45 \times 10^{-4}}$	8.53×10^{-4}	1.41×10^{-4}	9.13×10^{-4}	9.24×10^{-4}	1.65×10^{-4}	9.63×10^{-4}	9.77×10^{-4}	1.69×10^{-4}
1	$\mathbf{1.62 \times 10^{-3}}$	1.64×10^{-3}	2.69×10^{-4}	1.70×10^{-3}	1.71×10^{-3}	2.74×10^{-4}	1.71×10^{-3}	1.73×10^{-3}	3.00×10^{-4}	1.78×10^{-3}	1.80×10^{-3}	3.00×10^{-4}
2	$\mathbf{3.18 \times 10^{-3}}$	3.21×10^{-3}	5.07×10^{-4}	3.42×10^{-3}	3.46×10^{-3}	5.48×10^{-4}	3.21×10^{-3}	3.25×10^{-3}	5.48×10^{-4}	3.24×10^{-3}	3.26×10^{-3}	5.08×10^{-4}
5	8.02×10^{-3}	8.06×10^{-3}	1.27×10^{-3}	8.82×10^{-3}	8.89×10^{-3}	1.41×10^{-3}	7.19×10^{-3}	7.29×10^{-3}	1.24×10^{-3}	$\mathbf{6.69 \times 10^{-3}}$	6.74×10^{-3}	1.02×10^{-3}
10	1.64×10^{-2}	1.66×10^{-2}	2.72×10^{-3}	1.84×10^{-2}	1.86×10^{-2}	3.09×10^{-3}	1.22×10^{-2}	1.24×10^{-2}	2.22×10^{-3}	$\mathbf{1.08 \times 10^{-2}}$	1.09×10^{-2}	1.80×10^{-3}
20	3.55×10^{-2}	3.58×10^{-2}	6.13×10^{-3}	4.09×10^{-2}	4.14×10^{-2}	7.22×10^{-3}	1.93×10^{-2}	1.96×10^{-2}	3.08×10^{-3}	$\mathbf{1.68 \times 10^{-2}}$	1.71×10^{-2}	2.69×10^{-3}

Notes. In each row minimal median value is in **bold**; *: value numerically indistinguishable from zero; **: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$

Table SA12: Precision of estimates for $p = 1000$, arbitrary P, t copula

(a) Euclidean loss

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	5.98×10^{-4}	6.06×10^{-4}	4.99×10^{-5}	$\mathbf{5.78 \times 10^{-4}}$	5.81×10^{-4}	2.62×10^{-5}	6.48×10^{-4}	6.57×10^{-4}	6.42×10^{-5}	6.47×10^{-4}	6.54×10^{-4}	6.18×10^{-5}
1	$\mathbf{1.12 \times 10^{-3}}$	1.13×10^{-3}	7.85×10^{-5}	1.16×10^{-3}	1.16×10^{-3}	5.09×10^{-5}	1.20×10^{-3}	1.22×10^{-3}	1.14×10^{-4}	1.16×10^{-3}	1.17×10^{-3}	1.01×10^{-4}
2	2.14×10^{-3}	2.16×10^{-3}	1.32×10^{-4}	2.31×10^{-3}	2.33×10^{-3}	1.04×10^{-4}	2.21×10^{-3}	2.26×10^{-3}	2.19×10^{-4}	$\mathbf{2.04 \times 10^{-3}}$	2.07×10^{-3}	1.79×10^{-4}
5	5.24×10^{-3}	5.28×10^{-3}	2.77×10^{-4}	5.82×10^{-3}	5.85×10^{-3}	2.55×10^{-4}	4.93×10^{-3}	5.00×10^{-3}	5.31×10^{-4}	$\mathbf{4.16 \times 10^{-3}}$	4.23×10^{-3}	3.88×10^{-4}
10	1.05×10^{-2}	1.05×10^{-2}	5.46×10^{-4}	1.18×10^{-2}	1.18×10^{-2}	5.47×10^{-4}	8.43×10^{-3}	8.56×10^{-3}	9.78×10^{-4}	$\mathbf{6.89 \times 10^{-3}}$	6.99×10^{-3}	7.42×10^{-4}
20	2.10×10^{-2}	2.11×10^{-2}	1.09×10^{-3}	2.40×10^{-2}	2.41×10^{-2}	1.15×10^{-3}	1.33×10^{-2}	1.34×10^{-2}	1.48×10^{-3}	$\mathbf{1.10 \times 10^{-2}}$	1.12×10^{-2}	1.41×10^{-3}

(b) KLIC (known true d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	$\mathbf{1.79 \times 10^{-2}}$	1.79×10^{-2}	1.21×10^{-4}	1.81×10^{-2}	1.81×10^{-2}	1.67×10^{-4}	$\mathbf{1.79 \times 10^{-2}}$	1.79×10^{-2}	1.23×10^{-4}	$\mathbf{1.79 \times 10^{-2}}$	1.79×10^{-2}	1.21×10^{-4}
1	$\mathbf{1.86 \times 10^{-2}}$	1.87×10^{-2}	2.29×10^{-4}	1.90×10^{-2}	1.90×10^{-2}	2.95×10^{-4}	$\mathbf{1.86 \times 10^{-2}}$	1.86×10^{-2}	2.30×10^{-4}	$\mathbf{1.86 \times 10^{-2}}$	1.86×10^{-2}	2.18×10^{-4}
2	2.02×10^{-2}	2.02×10^{-2}	4.69×10^{-4}	2.07×10^{-2}	2.07×10^{-2}	5.83×10^{-4}	2.00×10^{-2}	2.00×10^{-2}	4.48×10^{-4}	$\mathbf{1.98 \times 10^{-2}}$	1.98×10^{-2}	3.91×10^{-4}
5	2.48×10^{-2}	2.48×10^{-2}	1.22×10^{-3}	2.61×10^{-2}	2.61×10^{-2}	1.45×10^{-3}	2.34×10^{-2}	2.35×10^{-2}	1.10×10^{-3}	$\mathbf{2.27 \times 10^{-2}}$	2.28×10^{-2}	8.98×10^{-4}
10	3.29×10^{-2}	3.31×10^{-2}	2.54×10^{-3}	3.57×10^{-2}	3.60×10^{-2}	3.08×10^{-3}	2.80×10^{-2}	2.81×10^{-2}	1.72×10^{-3}	$\mathbf{2.63 \times 10^{-2}}$	2.64×10^{-2}	1.37×10^{-3}
20	5.06×10^{-2}	5.07×10^{-2}	5.65×10^{-3}	5.68×10^{-2}	5.73×10^{-2}	6.96×10^{-3}	3.39×10^{-2}	3.40×10^{-2}	2.54×10^{-3}	$\mathbf{3.16 \times 10^{-2}}$	3.17×10^{-2}	2.31×10^{-3}

(c) KLIC (MPLE d.f.)

p/n	$\widehat{P}^{\text{smpl}}$			$\widehat{P}^{\text{i-}\tau}$			$\widehat{P}^{\text{LSh}}$			$\widehat{P}^{\text{NLSh}}$		
	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.	median	mean	s.d.
0.5	$\mathbf{9.79 \times 10^{-4}}$	9.99×10^{-4}	1.72×10^{-4}	1.46×10^{-1}	1.46×10^{-1}	4.31×10^{-4}	1.07×10^{-3}	1.10×10^{-3}	1.94×10^{-4}	6.94×10^{-3}	5.79×10^{-3}	2.63×10^{-3}
1	9.03×10^{-3}	3.93×10^{-2}	5.72×10^{-2}	1.47×10^{-1}	1.47×10^{-1}	6.20×10^{-4}	$\mathbf{1.97 \times 10^{-3}}$	2.01×10^{-3}	3.31×10^{-4}	2.08×10^{-3}	1.35×10^{-2}	3.87×10^{-2}
2	1.47×10^{-1}	1.47×10^{-1}	7.44×10^{-4}	1.49×10^{-1}	1.49×10^{-1}	9.73×10^{-4}	3.65×10^{-3}	3.67×10^{-3}	6.01×10^{-4}	$\mathbf{3.60 \times 10^{-3}}$	3.61×10^{-3}	5.29×10^{-4}
5	1.52×10^{-1}	1.52×10^{-1}	1.54×10^{-3}	1.54×10^{-1}	1.54×10^{-1}	1.95×10^{-3}	7.96×10^{-3}	8.07×10^{-3}	1.43×10^{-3}	$\mathbf{7.25 \times 10^{-3}}$	7.37×10^{-3}	1.37×10^{-3}
10	1.60×10^{-1}	1.60×10^{-1}	2.93×10^{-3}	1.64×10^{-1}	1.64×10^{-1}	3.62×10^{-3}	$\mathbf{1.35 \times 10^{-2}}$	1.36×10^{-2}	2.16×10^{-3}	1.55×10^{-2}	1.49×10^{-2}	2.73×10^{-3}
20	1.77×10^{-1}	1.78×10^{-1}	5.79×10^{-3}	1.84×10^{-1}	1.85×10^{-1}	7.15×10^{-3}	2.42×10^{-2}	2.40×10^{-2}	3.67×10^{-3}	$\mathbf{2.41 \times 10^{-2}}$	2.43×10^{-2}	2.64×10^{-3}

Notes. In each row minimal median value is in **bold**; *****: value numerically indistinguishable from zero; ******: $+\infty$ values of KLIC in the samples due to non-PD $\widehat{P}$