Working Paper Series 699
(ISSN 1211-3298)

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

Stanislav Anatolyev
Vladimir Pyrlik

CERGE-EI
Prague, August 2021

ISBN 978-80-7343-506-6 (Univerzita Karlova, Centrum pro ekonomický výzkum a doktorské studium)

ISBN 978-80-7344-601-7 (Národohospodářský ústav AV ČR, v. v. i.)

Shrinkage for Gaussian and t Copulas in Ultra-High Dimensions

Stanislav Anatolyev* Vladimir Pyrlik†

CERGE-EI and NES CERGE-EI

July 2021

Abstract

Copulas are a convenient framework to synthesize joint distributions, particularly in higher dimensions. Currently, copula-based high dimensional settings are used for as many as a few hundred variables and require large data samples for estimation to be precise. In this paper, we employ shrinkage techniques for large covariance matrices in the problem of estimation of Gaussian and t copulas whose dimensionality goes well beyond that typical in the literature. Specifically, we use the covariance matrix shrinkage of Ledoit and Wolf to estimate large matrix parameters of Gaussian and t copulas for up to thousands of variables, using up to 20 times lower sample sizes. The simulation study shows that the shrinkage

*Address: Stanislav Anatolyev, CERGE-EI, a joint workplace of Center for Economic Research and Graduate Education, Charles University and the Economics Institute of the Czech Academy of Sciences, Politickych vězňů 7, P.O. Box 882, 111 21 Prague 1, Czech Republic; e-mail stanislav.anatolyev@cerge-ei.cz. This research was supported by the grant 20-28055S from the Czech Science Foundation.

†Address: Vladimir Pyrlik, CERGE-EI, a joint workplace of Center for Economic Research and Graduate Education, Charles University and the Economics Institute of the Czech Academy of Sciences, Politických vězňů 7, P.O. Box 882, 111 21 Prague 1, Czech Republic; e-mail vladimir.pyrlik@cerge-ei.cz.

estimation significantly outperforms traditional estimators, both in low and especially high dimensions. We also apply this approach to the problem of allocation of large portfolios.

KEYWORDS: Gaussian copula, t copula, high dimensionality, large covariance matrices, shrinkage, portfolio allocation

JEL CODES: C31, C46, C55, C58

HIGHLIGHTS:

- Methods of covariance matrix shrinkage are applied to estimate parameters of Gaussian and t copulas in ultra-high dimensions.
- Simulations illustrate dominance of shrinkage estimators over traditional copula estimators.
- The approach is applied for a large portfolio allocation problem with up to 3600 assets.

1 Introduction

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).

Another 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).

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). 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). 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), or vine copulas. 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.

In this paper, we focus on elliptical copulas in high dimensions. We focus on the two most commonly used in modeling and practical applications: Gaussian and t copulas. These copulas are used in a vast variety of applications and 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 geographical 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).

Nevertheless, 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.

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. In low dimensions, 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.

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. An important property of these copulas is that their matrix parameter is very close the correlation matrix of pseudo-observations (Demarta and McNeil, 2005; Kojadinovic and Yan, 2010). This allows one 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.[1] In particular, 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.[2] 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

[1] 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 the method of maximum pseudo-likelihood.

[2] 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.

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 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 over 3600 U.S. stocks, we construct portfolios of up to 3600 assets and simulate buy-and-hold portfolio strategies. The joint distributional models of asset returns are estimated over the period of six months (120 observations), hence the problem is ultra-high dimensional, with the ratio of data dimensionality to sample size being 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 paper is organized as follows. Section 2 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 3, 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 4.

Section 5 concludes. Appendices contain some additional technical material, including tables with detailed results of the simulations in the Supplementary Appendix

2 Methodology

2.1 Sklar's theorem and copula classes

A convenient way to introduce the copula approach is through the Sklar's theorem (Sklar, 1959), the key result in the copula theory. Given $X \equiv (X_1, ..., X_p)' \in \mathbb{R}^p$ a p-dimensional random vector from a distribution with the joint cumulative distribution function (CDF) $F_X(x)$ and marginal CDFs $\{F_i(x_i)\}_{i=1,...,p}$, there exists a copula function $C(u)$,

$$C : [0,1]^p \to [0,1], \tag{2.1}$$

such that for all $x = (x_1, ..., x_p)' \in \mathbb{R}^p$,

$$F_X(x) = C\big(F_1(x_1), ..., F_p(x_p)\big). \tag{2.2}$$

This theorem is particularly useful as its converse also holds: given a set of univariate distributions with CDFs $\{F_i(x_i)\}_{i=1,...,p}$ and a copula function $C(u)$, the corresponding function $F_X(x)$ defined for these functions from (2.2) is a legit joint CDF with marginals $\{F_i(x_i)\}_{i=1,...,p}$.

Thus, marginal distributions of the quantities of interest can be modeled separately from the interdependence embedded by the copula function $C(u)$. This brings on a variety of classes of copulas developed in the literature over the years. We only briefly recall some of the main existing classes of copulas focusing on the Gaussian and t copulas.

One of essential classes of copulas is the Archimedean copulas, whose members are often used in modeling bivariate distributions. A major advantage of this class of cop-

ulas is that most of them have a closed-form representation. Further, by construction, any Archimedean copula is extendable to an arbitrary dimensionality p. However, the parameter space (uni-dimensional in most cases) is disconnected from the data dimensionality resulting in insufficient flexibility of dependence structures as data dimensionality grows (Hofert et al., 2012). The tightness of parameterization of Archimedean copulas is the main reason for this class to be rarely chosen to model dependence beyond bivariate settings.

Another class of copulas is pair copula constructions (PCCs), also known as vine copulas, based on sequential construction of the multivariate distribution (2.2) from the marginals and a series of corresponding bivariate conditional copulas. In general, the bivariate copulas for all pairs are chosen independently of the marginal models and of each other. Thus, this class allows one to attain maximal flexibility in copula construction. However, the cost of this flexibility is an ultimately growing number of alternative specifications with higher data dimensionality. To make the PCCs operational in practice, simplifying assumptions are made to restrict the structure, and heuristic algorithms are applied to identify and distinguish between alternative restricted structures (Aas et al., 2009; Brechmann et al., 2012; Brechmann and Czado, 2013; Czado et al., 2013; Dissmann et al., 2013). Further, the PCCs are well extendable to high data dimensionality. Yet, they become computationally demanding because of both the heuristic algorithms used to pre-identify the vine structure, and actual estimation of parameters of a high-dimensional vine. For the heuristic algorithms to work and deliver sustainable results, the sample sizes need to remain comparable with the number of variables. So far, the highest dimensionality has been reached by Müller and Czado (2019), with a vine applied to 2131 variables (stock returns from industrial sectors) with 999 observations (daily data). The dimensionality ratio is thus slightly above 2, and the structure of the vine is sparse.

We stick to the class of elliptical copulas that are defined directly from elliptical

distributions by inversion of (2.2). The copulas are parameterized in a way close to the corresponding distributions from which they are defined. As all elliptical distributions are transformations of the multivariate normal, a key parameter is the matrix corresponding to the covariance matrix of the underlying normal random variable. This makes the dimensionality of the parameter space naturally connected to the dimensionality of the data and correspondingly interpretable.

The most popular elliptical copula is Gaussian, which is the copula of the multivariate normal distribution. Another important elliptical copula is a natural extension to the normal one, the t copula of the multivariate Student's t distribution. These two elliptical copulas inherit their main limitations from the underlying elliptical distributions. Thus, both Gaussian and t copulas are symmetric, and only the t copula exhibits (also symmetric) non-zero tail dependence (Demarta and McNeil, 2005). Nevertheless, these copulas are often used as building blocks in more complicated settings seeking to capture desired properties of the data (Zimmer, 2012; Patton, 2012; De Leon and Chough, 2013; Patton, 2013; Oh and Patton, 2017).

In the next subsection we formally introduce the Gaussian and t copulas and some of their properties that are important for our analysis.

2.2 Gaussian and t copulas

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

$$C_P^{\mathcal{N}}(u) = F_P\big(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_p)\big), \tag{2.3}$$

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

$$C^t_{P,\nu}(u) = t_{P,\nu}\big(t^{-1}_\nu(u_1), ..., t^{-1}_\nu(u_p)\big), \tag{2.4}$$

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 $t^{-1}_\nu(u_i)$ is the quantile function of the standard univariate t-distribution with ν degrees of freedom.

As any other copula function, the copulas (2.3) and (2.4) 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[3]. For a pair of random variables $\{U_i, U_j\}$, Kendall's rank correlation, or Kendall's i-τ, is defined as

$$\tau_{ij} \equiv \mathbb{E}\left[\text{sign}\left((U_i - \tilde{U}_i)(U_j - \tilde{U}_j)\right)\right], \tag{2.5}$$

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 $C(u) = C^{\mathcal{N}}_P(u)$ or $C(u) = C^t_{P,\nu}(u)$ it holds that:

$$\tau_{ij} = \frac{2}{\pi}\arcsin\left(P_{ij}\right). \tag{2.6}$$

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[4]. Firstly, in the case of multivariate normal distribution and its copula,

[3] 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.

[4] 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*

i.e. $\mathcal{U} \sim C_P^{\mathcal{N}}(u)$, the relation has the following analytical form:

$$\text{Corr}(\mathcal{U}) = \frac{6}{\pi}\text{asin}\left(\frac{P}{2}\right). \tag{2.7}$$

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

$$\text{Corr}(\mathcal{U}) \approx P. \tag{2.8}$$

Further, in the case of t copula, $\mathcal{U} \sim C_{P,\nu}^{t}(u)$, there is no closed form expression for $\text{Corr}(\mathcal{U})$. However, the relations (2.7) and (2.8) 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, yet it approaches the level of the one for the Gaussian copula rather quickly as the degrees of freedom value 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, (2.8) and (2.6) 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 Section 2.3.

Another construct related to the copula function is the copula density function, the probability density function (PDF) associated with the copula function $C(u)$ as a CDF:

$$c(u) = \frac{\partial^p C(u)}{\partial u_1 ... \partial u_p}. \tag{2.9}$$

In the case of Gaussian and t copulas defined by (2.3) and (2.4) it is easy to show using

(2.9) that the corresponding copula log-densities are

$$\log c_P^{\mathcal{N}}(u) = -\frac{1}{2}\log|P| - \frac{1}{2}\phi'(u)\cdot(P^{-1} - I_p)\cdot\phi(u), \tag{2.10}$$

and

$$\begin{aligned}\log c_{P,\nu}^{t}(u) = {} & \log\Gamma\left(\frac{\nu+p}{2}\right) + (p-1)\log\Gamma\left(\frac{\nu}{2}\right) - p\log\Gamma\left(\frac{\nu+1}{2}\right) - \frac{1}{2}\log|P| \\ & - \frac{\nu+p}{2}\log\left(1 + \frac{\psi_\nu'(u)P^{-1}\psi_\nu(u)}{\nu}\right) + \sum_{i=1}^{p}\log\left(1 + \frac{t_\nu^{-1}(u_i)^2}{\nu}\right),\end{aligned} \tag{2.11}$$

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 (2.10) and (2.11) are used in evaluation of estimation precision via the Kullback–Leibler information criterion (KLIC) presented later in Section 3.1.2.

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 $\tilde{P} = \{P_{i_{s_1} i_{s_2}}\}_{s_1, s_2 \in \{1,...,\tilde{p}\}}$.

2.3 Copula estimation

2.3.1 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 (2.2) 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*

$$U_{it} = \hat{F}_i(X_{it}), \tag{2.12}$$

and the copula function (2.1) is treated as the joint distribution function of the pseudo-observations (2.12), 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 (2.12) 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 (2.8) and (2.6) 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

$$\hat{P}^{\text{smpl}} = \{\hat{P}^{\text{smpl}}_{ij}\}_{i,j=1,\dots,p} = \left\{\widehat{\text{corr}}(U_i, U_j)\right\}_{i,j=1,\dots,p}, \tag{2.13}$$

and

$$\hat{P}^{\text{i-}\tau} = \{\hat{P}^{\text{i-}\tau}_{ij}\}_{i,j=1,\dots,p} = \left\{\sin\left(\frac{\pi}{2}\hat{\tau}_{ij}\right)\right\}_{i,j=1,\dots,p}, \tag{2.14}$$

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 (2.14) 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 (2.13) 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 (2.7) 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 (2.8) are relatively small, both for Gaussian and t copulas, and the estimator (2.13) 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.

2.3.2 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; **?**).

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

$$S_n = \frac{X_n' X_n}{n}. \tag{2.15}$$

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 (2.15):

$$\Sigma^* = \rho_1 I_p + \rho_2 S_n. \tag{2.16}$$

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

$$S^* = \hat{\vartheta}\hat{\mu}I_p + (1 - \hat{\vartheta})S_n, \tag{2.17}$$

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 (2.17) 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):

$$\mu = \mathbb{E}\left[\frac{1}{p}\sum_{i=1}^{p} l_i\right] = \frac{1}{p}\sum_{i=1}^{p}\lambda_i. \tag{2.18}$$

Also, Ledoit and Wolf (2004b) show that

$$\frac{1}{p}\,\mathbb{E}\left[\sum_{i=1}^{p}(l_i - \mu)^2\right] = \frac{1}{p}\sum_{i=1}^{p}(\lambda_i - \mu)^2 + \mathbb{E}||S_n - \Sigma||^2. \tag{2.19}$$

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 (2.17) reduces the bias of the sample covariance matrix eigenvalues by shifting them towards their grand mean (2.18) shrinking the distribution of the sample eigenvalues. The shrunk eigenvalues corresponding to the optimal linear shrinkage estimator (2.17) are

$$\lambda_i^* = \vartheta\mu + (1 - \vartheta)l_i, \tag{2.20}$$

where the coefficients ϑ and μ are probability limits, under dimension asymptotics, of $\hat{\vartheta}$ and $\hat{\mu}$ in (2.17), and so the shrunk eigenvalues can then be estimated from the data similarly to how the estimator (2.17) estimates (2.16):

$$l_i^* = \hat{\vartheta}\hat{\mu} + (1 - \hat{\vartheta})l_i, \tag{2.21}$$

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

$$S^* = \Gamma_n \text{diag}\{l_i^*\}_{i=1,...,p}\Gamma_n', \tag{2.22}$$

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 (2.22) 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 (2.17) estimates the non-feasible estimator (2.16). The non-linear shrinkage estimator preserves the form of the rotation equivariant estimator (2.22), with the linearly shrunk eigenvalues l^*s (2.21)

replaced by the non-linearly shrunk versions:

$$l_i^{**} = \frac{l_i}{|1 - \frac{p}{n} - \frac{p}{n} l_i \breve{m}_F^*(l_i)|^2}. \tag{2.23}$$

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 (2.13). However, certain adaptations are in order.

First, since the shrinkage estimators estimate the population covariance matrix, they need to be transformed to estimates of the correlation matrix. Alternatively, the shrinkage estimators can be applied to standardized pseudo-observations. Given that the univariate means and variances of the pseudo-observations are known constants (respectively, $1/2$ and $1/12$, coming from $u_i \sim U[0,1]\ \forall i$), preliminary standardization

of pseudo-observations is preferred to avoid extra noise and computational time of converting covariance matrices into correlations.

Second, the shrinkage estimators and their properties rely on i.i.d. data samples, while in copula estimation the pseudo-observations (2.12) 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-iidness" of pseudo-observations and applying the shrinkage estimators will perform sufficiently better than the traditional estimators (2.13) and (2.14).

3 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 (2.12) 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.

3.1 Simulation design

3.1.1 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

$$p \in \{10, 100, 1000\}. \tag{3.1}$$

- 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:

$$\frac{p}{n} \in \begin{cases} \{1/20,\ 1/10,\ 1/2,\ 1,\ 2\}, & p = 10, \\ \{1/10,\ 1/2,\ 1,\ 2,\ 5,\ 10\}, & p = 100, \\ \{1/2,\ 1,\ 2,\ 5,\ 10,\ 20\}, & p = 1000. \end{cases} \tag{3.2}$$

- 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 1.

- 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 $U[-1, 1]$.

3.1.2 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:

$$L_E(P, \hat{P}) = ||\text{vech}(P - \hat{P})||. \tag{3.3}$$

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 (2.1) itself, we measure the closeness of the estimated one to the true one via the Kullback-Leibler information criterion (KLIC):

$$KLIC_{P|\hat{P}} = \mathbb{E}_{C_P}\left[\log\left(\frac{c_P(u)}{c_{\hat{P}}(u)}\right)\right] = \int \cdots \int_{\mathbb{O}_p}^{\mathbb{1}_p} c_P(u) \log \frac{c_P(u)}{c_{\hat{P}}(u)} \mathrm{d}^p u. \tag{3.4}$$

While the first two criteria are computationally practical even when p is large, calculating KLICa 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 2.2), so for any data dimensionality p we only consider KLICa for 3-dimensional subsets of the data. For $p = 10$, we compute the KLIC for only one triplet; for $p = 100$, we average KLICa over randomly chosen 30 triplets, and for $p = 1000$ the number of triplets we average over is 100. Second, we estimate the expectation in (3.4) 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 (3.4) using the expressions for log-densities of Gaussian and t copulas

(2.10) and (2.11):

$$KLIC_{\tilde{P}|\tilde{\hat{P}}} = M^{-1} \sum_{m=1}^{M} \left(\log c_{\tilde{P}}(\tilde{u}_m) - \log c_{\tilde{\hat{P}}}(\tilde{u}_m) \right). \tag{3.5}$$

3.1.3 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 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 (3.3),
   - KLIC, via (3.5) and averaged over randomized triplets of variables;
   - * for t copulas, KLICa 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.[5]

3.2 Simulation results

The simulation results are presented in Tables SA1 – SA13 in the Supplementary Appendix. For each evaluation criterion, we report the median, mean and standard devia-

[5] 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.

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

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 KLICa 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.

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–SA5). More importantly, the difference in performance is well observed for higher dimensions and smaller samples (see Tables SA10–SA13).

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 2.

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, Figures 2a and 2b). 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 Figure 2c). 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 2d) in which the linear shrinkage based estimator dominates all others, with the nonlinear shrinkage, in this case, only slightly underperforming (see Table SA13c), and the differences between the two can be neglected.

4 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*.[6] 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

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

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 residual terms. 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, 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 4. Figure 3 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 4).

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.

5 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. 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 shrink-

age 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.

Finally, yet another potentially beneficial application that we leave for future research is construction of linear forecast combinations under many alternative predictors. Technically, the forecast combination problem is similar to the portfolio allocation problem. When the number of alternative predictors is large and especially if they belong to one family of predictive models, there is normally a great chance that the forecast errors will be strongly and positively correlated and, importantly, with a certain degree of tail dependence. Accounting for the interdependence of forecast errors from many alternative predictors can crucially improve the quality of combined forecasts. However, the validation samples to train the combined forecasts need to be rather short so that there is no too much noise from outdated information in predictions. This limits the number of alternative predictors to be used when traditional estimators are applied. Thus, using high dimensionality robust estimators will allow one to use more predictors and construct more accurate combined forecasts.

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.

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

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

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

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.

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

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.

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.

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.

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

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.

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.

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.

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.

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

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.

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., 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.

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

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

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.

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*.

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.

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.

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.

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.

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

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.

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.

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.

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

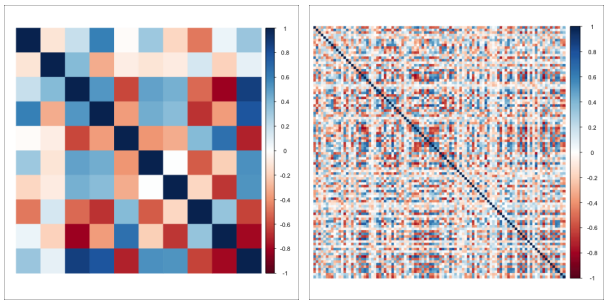
Zorgati, I., F. Lakhal, and E. Zaabi (2019). Financial contagion in the subprime crisis context: A copula approach. *The North American Journal of Economics and Finance 47*, 269–282.

Tables

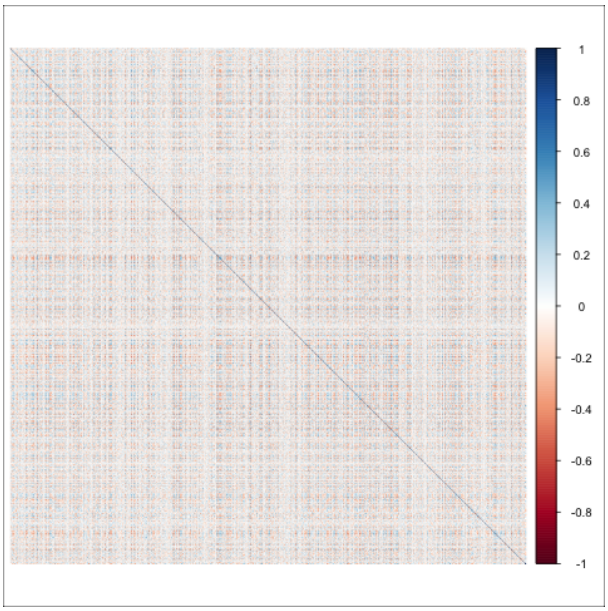
Table 1: Mean (s.d.) time of evaluation of $\hat{P}$ estimators, milliseconds

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

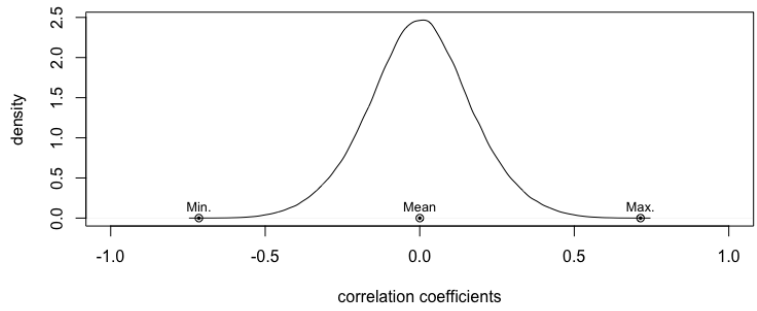
Figures



(a) $p = 10$ (b) $p = 100$

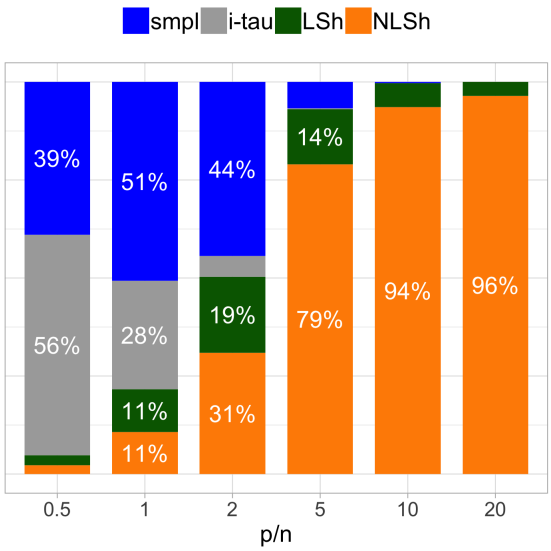


(c) $p = 1000$

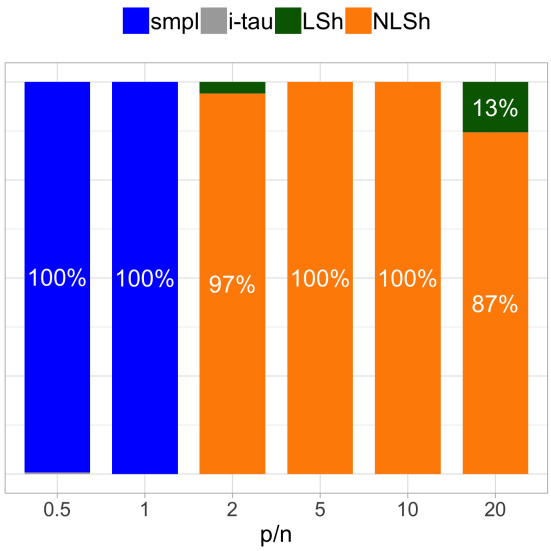


(d) distribution of correlation coefficients for the $p = 1000$ matrix

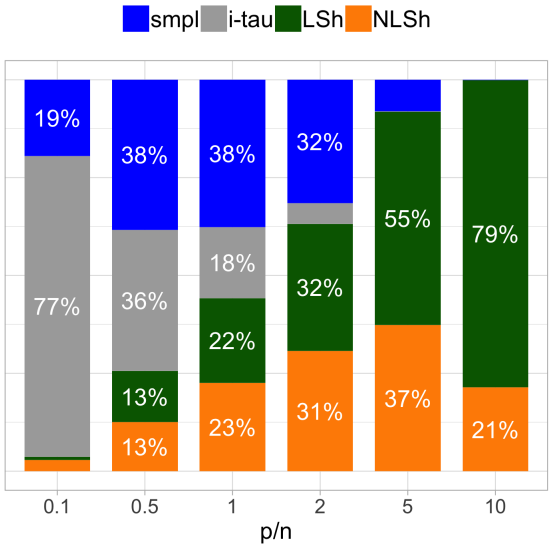
Figure 1: True correlation matrices P of arbitrary structure



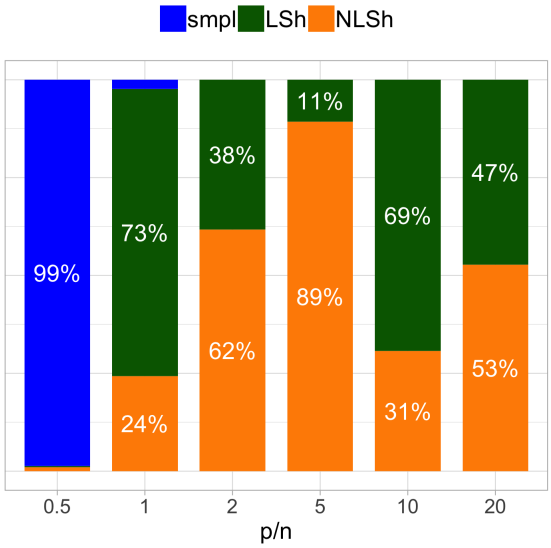
(a) $p = 1000$, Gaussian copula, arbitrary P



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



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



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

Figure 2: Shares of simulations in which each estimator returns the best KLIC

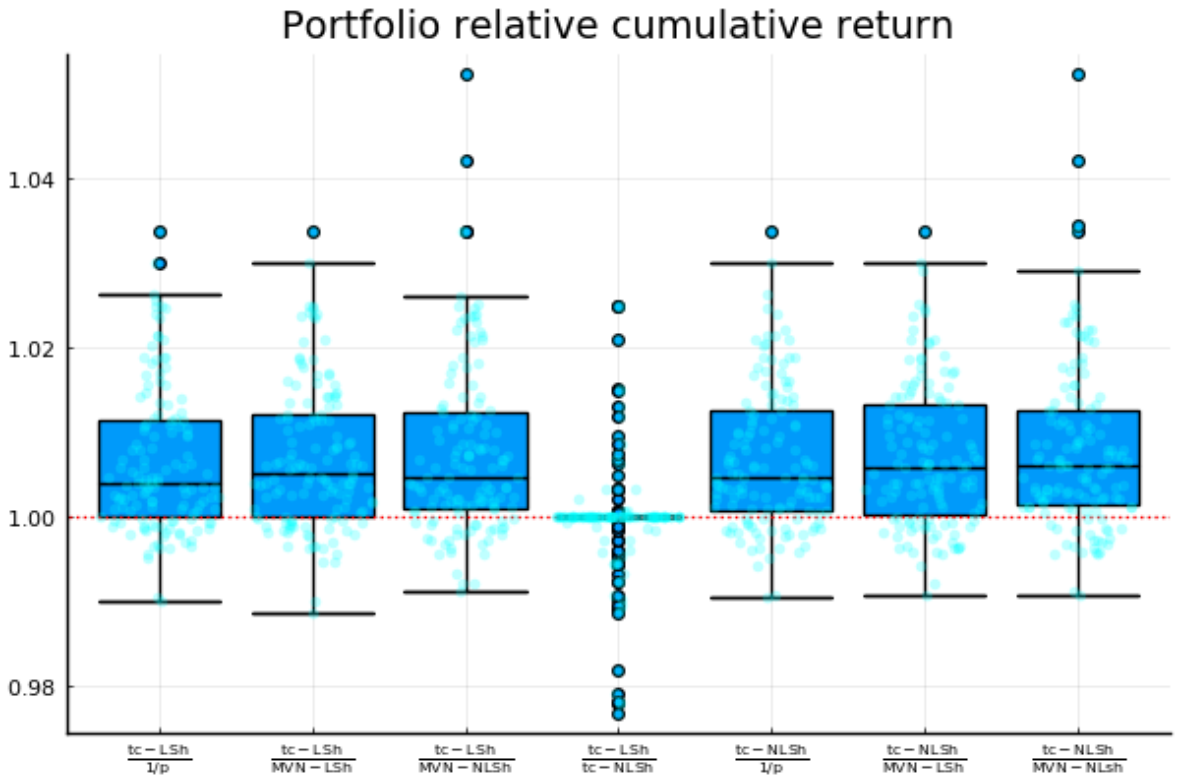


Figure 4: Relative returns of model-based portfolios across the sets of assets

A Quality of approximation of a copula correlation parameter with pseudo-observations correlation

The results of the study heavily rely on the approximation (2.8). It suggests using simple correlation of pseudo-observations $\mathcal{U}$ from either Gaussian or t copula as a reliable 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. if the copula's matrix parameter

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

the approximation (2.8) 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 either Gaussian or t copula (in the case of t copula, the parameter of degrees of freedom ν varies in $\{2, 4, 8, 10, 16\}$) and evaluate the error of this approximation for different values of ρ. The results are summarized in Figure 5.

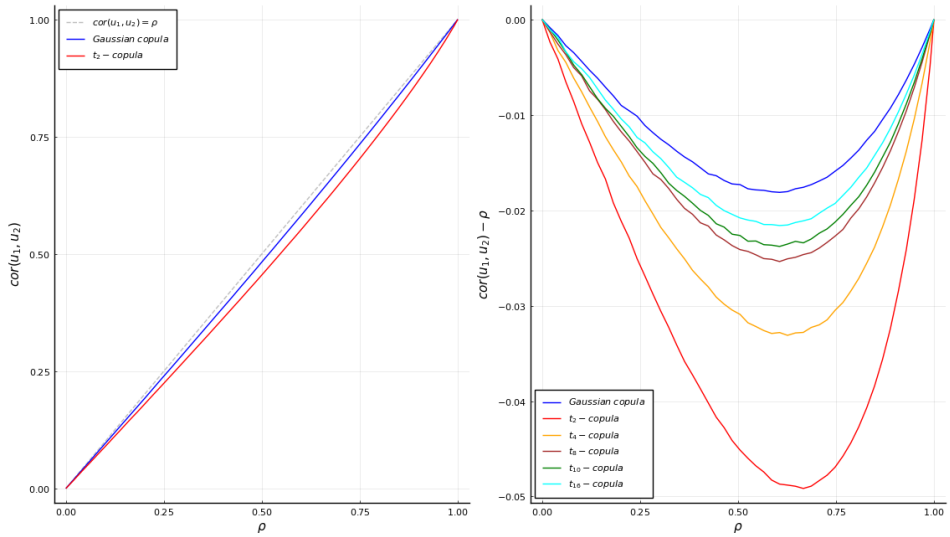


Figure 5: Approximation of copula parameter by pseudo-observations correlation

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, $\{S_t^i\}_{i=1,\dots,p,t=1,\dots,T}$. We call *a portfolio* a p-dimensional vector of shares, $\alpha = (\alpha_1, \dots, \alpha_p)$, such that $\forall i = 1, \dots, p\ \alpha_i \geq 0$, 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:

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

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 2.3.1 & 2.3.2), 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 $t = n + 1, ..., T$, $\{e^i_t(b)\}_{i=1,...,p,t=n+1,...,T,b=1,...,B}$, and use the fitted ARMA-EGARCH specifications to calculate the corresponding trajectories of future stock prices, $\{\widehat{S}^i_t(b)\}_{...}$. We use these simulated data to calculate the simulation analogs of the portfolio value (B.1) and return (B.2) as

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

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

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: $\alpha^*_{MVN-LSh}$, $\alpha^*_{MVN-NLSh}$, α^*_{tc-LSh}, $\alpha^*_{tc-NLSh}$, 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 $M_1 \in \{tc{-}LSh,\ tc{-}NLSh\}$, $M_2 \in \{1/p, MVN{-}LSh, MVN{-}NLSh, tc{-}NLSh\}/M_1$.

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[7] 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

[7] 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

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.

C Technical remarks

C.1 Computational software

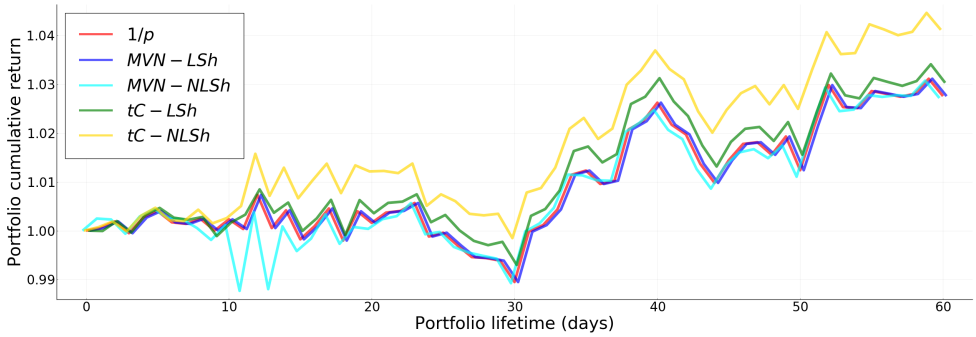
All the calculations for the simulation study were performed using R language (R Core Team, 2013). The packages *foreach* (Weston, 2019b) and *doParallel* (Weston, 2019a) were used to perform parallel computations. The package *copula* (Hofert et al., 2018; Jun Yan, 2007; Ivan Kojadinovic and Jun Yan, 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 implied to estimate and select ARMA-EGARCH models.

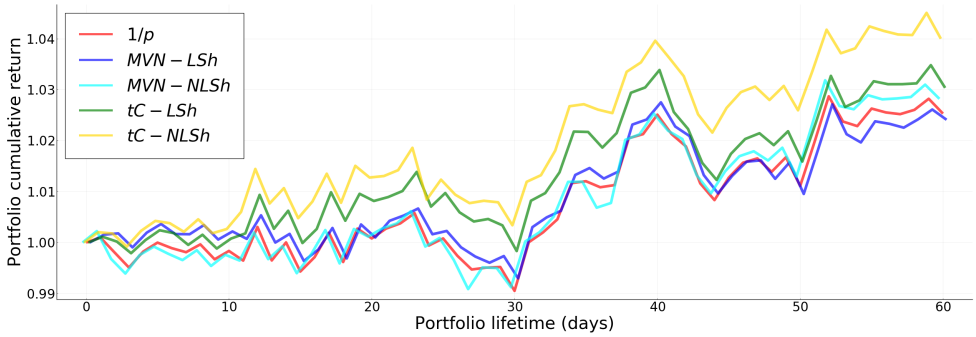
C.2 Evaluation time of estimators

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 1. The assessment of evaluation time was performed on an Intel(R) Core(TM)

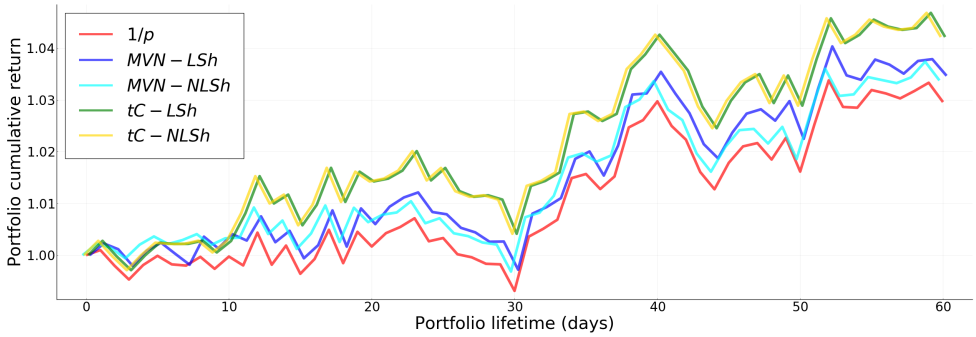
i7-7700K CPU @4.20GHz machine with 16GB of RAM running on Windows 10 Home edition. For assessing evaluation time, no parallel computing was used. The R package *microbenchmark* (Mersmann, 2019) was used to record the running time of the four estimators evaluation.



(a)



(b)



(c)

(d)

Figure 3: Examples of model-based portfolio cumulative return dynamics

Abstrakt

Kopule představují užitečný rámec pro modelování sdružených rozdělení, obzvlášť v případě vysokorozměrných dat. V současnosti jsou kopule aplikovány na problémy s potenciálně až několika stovkami proměnných a vyžadují velký výběr, aby odhad byl dostatečně přesný. V tomto článku využíváme metodu smrštění pro odhad velkých kovariačních matic Gaussovských a t kopulí, jejichž dimenze je výrazně vyšší než obvyklé dimenze v existující literatuře. Konkrétně využíváme smrštění kovariační matice dle Ledoita a Wolfa k odhadu vysokorozměrných matic Gaussovských a t kopulí s řádově až tisíci proměnnými a až 20násobně menšími výběry, než je běžné. Simulační studie ukazuje, že smrštěný odhad významně překonává tradiční odhady v případě nižších i vyšších dimenzí. Tento přístup také aplikujeme na problém alokace velkých portfolií.

Klíčová slova: Gaussovské kopule, t kopule, vysoká dimenze, velké kovariační matice, smrštění, alokace portfolia

The appendix to this working paper is available at https://www.cerge-ei.cz/working-papers/.

Working Paper Series
ISSN 1211-3298
Registration No. (Ministry of Culture): E 19443

Individual researchers, as well as the on-line and printed versions of the CERGE-EI Working Papers (including their dissemination) were supported from institutional support RVO 67985998 from Economics Institute of the CAS, v. v. i.

Specific research support and/or other grants the researchers/publications benefited from are acknowledged at the beginning of the Paper.

Published by
Charles University, Center for Economic Research and Graduate Education (CERGE)
and
Economics Institute of the CAS, v. v. i. (EI)
CERGE-EI, Politických vězňů 7, 111 21 Prague 1, tel.: +420 224 005 153, Czech Republic.
Printed by CERGE-EI, Prague
Subscription: CERGE-EI homepage: http://www.cerge-ei.cz

Phone: + 420 224 005 153
Email: office@cerge-ei.cz
Web: http://www.cerge-ei.cz

Editor: Byeongju Jeong

The paper is available online at http://www.cerge-ei.cz/publications/working_papers/.

ISBN 978-80-7343-506-6 (Univerzita Karlova, Centrum pro ekonomický výzkum a doktorské stadium)
ISBN 978-80-7344-601-7 (Národohospodářský ústav AV ČR, v. v. i.)