Journal of Multivariate Analysis 172 (2019) 180–192

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Model selection in sparse high-dimensional vine copula models with an application to portfolio risk

T. Nagler *, C. Bumann, C. Czado

Fakultät für Mathematik, Technische Universität München, Boltzmanstraße 3, 85748 Garching, Germany

ARTICLE INFO

Article history:
Received 29 January 2018
Available online 19 March 2019

AMS 2010 subject classifications:
62F07

Keywords:
BIC
Model selection
Sparsity
Value-at-Risk
Vine copula

ABSTRACT

Vine copulas allow the construction of flexible dependence models for an arbitrary number of variables using only bivariate building blocks. The number of parameters in a vine copula model increases quadratically with the dimension, which poses challenges in high-dimensional applications. To alleviate the computational burden and risk of overfitting, we propose a modified Bayesian information criterion (BIC) tailored to sparse vine copula models. We argue that this criterion can consistently distinguish between the true and alternative models under less stringent conditions than the classical BIC. The criterion suggested here can further be used to select the hyper-parameters of sparse model classes, such as truncated and thresholded vine copulas. We present a computationally efficient implementation and illustrate the benefits of the proposed concepts with a case study where we model the dependence in a large portfolio.

1. Introduction

Following the 2008 financial crisis, academic research and public media identified unrealistic mathematical models of the dependence between financial assets as one of its key causes [15,31]. In the aftermath, modeling these dependencies became a topic of intense study in finance and statistics. One of most promising tools that has emerged are vine copulas [2,5]. Vine copula models construct the dependence structure from bivariate building blocks, called pair-copulas. Each pair-copula captures the conditional dependence between a pair of variables. Because each pair-copula can be parametrized differently, vine copulas allow each pair to have a different strength and type of dependence. This flexibility is the key reason why vine copulas have become so popular for modeling dependence between financial assets; for a review, see [1].

The flexibility of vine copulas comes at the cost of a large number of model parameters. A vine copula on d variables consists of $d(d-1)/2$ pair-copulas, and each pair-copula can have multiple parameters. In financial applications, the number of model parameters quickly exceeds the number of variables. Suppose each pair-copula can have up to two parameters. Then a model for 50 assets has 2450 possible parameters; a model for 200 has almost 20,000. By contrast, five years of daily stock returns consist of roughly 1250 observations.

In such situations, there are two major challenges that we want to address: the risk of overfitting and the computational burden to fit thousands of parameters. Both issues make it desirable to keep the model sparse. A vine copula model is called sparse when a large number of pair-copulas are the independence copula. The key question is: which of the pair-copulas should we set to independence?

* Corresponding author.
E-mail addresses: thomas.nagler@tum.de (T. Nagler), cczado@ma.tum.de (C. Czado).

https://doi.org/10.1016/j.jmva.2019.03.004
0047-259X/

The most common strategy is to select the pair-copula family by either AIC or BIC [4,32]. The AIC is known to have a tendency to select models that are too large, which conflicts with the need for sparsity; see [12]. In contrast, the BIC is able to consistently select the true model, but only when the number of possible parameters grows sufficiently slowly with the sample size n. As explained above, this assumption should be seen critically in a high-dimensional context. Another problem is that the BIC is derived under the assumption that all models are equally likely. Under this assumption, we expect only half of the pair-copulas to be independence, a model we would not consider sparse.

We propose a new criterion called modified Bayesian Information Criterion for Vines (mBICV) that addresses both issues and is specifically tailored to vine copula models. It is based on a modification of the prior probabilities that concentrates on sparse models and is motivated by statistical practice. This modification turns out to be enough to relax the restrictions on the rate at which d diverges. The idea behind the mBICV is similar to other modified versions of the BIC that were developed for linear models; see [42].

The mBICV is useful for two things: selecting pair-copulas individually and selecting hyper-parameters of sparse model classes. One such class are truncated vine copulas [8,25]. A vine copula model is called truncated if all pair-copulas that are conditioned on more than M variables are set to independence. We further propose an alternative class of sparse models called thresholded vine copulas. They induce sparsity by setting a pair-copula to independence when the encoded strength of dependence falls short of a threshold. This idea has been around for a long time and heavily used, but the threshold was commonly tied to the p-value of an independence test [14]. In the more general form, the threshold is a free hyper-parameter. Such classes of sparse models give a computational advantage, since a substantial number of pair-copulas never have to be estimated.

2. Background on vine copulas

This section sets up the notation for vine copula models. For a more thorough introduction, see [23].

2.1. Copulas

Copulas are models for the dependence between random variables. By the theorem of Sklar [33], one can express any multivariate distribution F in terms of the marginal distributions $F_1, \dots, F_d$ and a function C, called the copula. One has, for all $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$,

$$F(x_1, \dots, x_d) = C\big\{F_1(x_1), \dots, F_d(x_d)\big\}. \tag{1}$$

If F is the distribution of a continuous random vector $\boldsymbol{X}$, then $C\colon [0,1]^d \to [0,1]$ is the distribution of $\boldsymbol{U} = (F_1(X_1), \dots, F_d(X_d))$. This suggests a two-step approach to estimation. Suppose we have observed data $x_{1,j}, \dots, x_{n,j}$ for all $j \in \{1, \dots, d\}$. Then we can proceed as follows: first, obtain estimates of the marginal distributions, say $\widehat{F}_1, \dots, \widehat{F}_d$; second, set $\widehat{u}_{i,j} = \widehat{F}_j(x_{i,j})$ and estimate C based on $\widehat{u}_{1,j}, \dots, \widehat{u}_{n,j}$ for all $j \in \{1, \dots, d\}$. Assuming that F is absolutely continuous, a formula similar to (1) can be derived for the density f of F. For all $\boldsymbol{x} \in \mathbb{R}^d$, one has

$$f(x_1, \dots, x_d) = c\big\{F_1(x_1), \dots, F_d(x_d)\big\} \times \prod_{k=1}^{d} f_k(x_k),$$

where c is the density of C and called the copula density, and $f_1, \dots, f_d$ are the marginal densities.

2.2. Vine copulas

Vine copula models are based on the idea of decomposing the dependence into a cascade of dependencies between (conditional) pairs due to Joe [22]. However, the decomposition is not unique. Bedford and Cooke [6] introduced a graphical model, called a regular vine, that organizes all valid decompositions. A regular vine on d variables consists of a sequence of linked trees $T_m = (V_m, E_m)$ with $m \in \{1, \dots, d-1\}$ with edge sets E_m and node sets V_m satisfying some conditions; see, e.g., [14].

A vine copula model identifies each edge in the vine with a bivariate copula. This is best explained by an example. Fig. 1 shows a regular vine tree sequence on five variables. The nodes in the first tree represent the random variables $U_1, \dots, U_5$. The edges are identified with a bivariate copula, called a pair-copula, where the edge (j_e, k_e) describes the dependence between U_{j_e} and U_{k_e}. In the second tree, the nodes are just the edges of the first tree. The edges are annotated by $(j_e, k_e; D_e)$ and describe the dependence between U_{j_e} and U_{k_e} conditional on U_{D_e}. In subsequent trees, the number of conditioning variables increases.

Bedford and Cooke [5] showed that this naturally leads to a decomposition of the copula density c, viz.

$$c(\boldsymbol{u}) = \prod_{m=1}^{d-1} \prod_{e \in E_m} c_{j_e,k_e;D_e}\big\{G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}),\ G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e});\ \boldsymbol{u}_{D_e}\big\}, \tag{2}$$

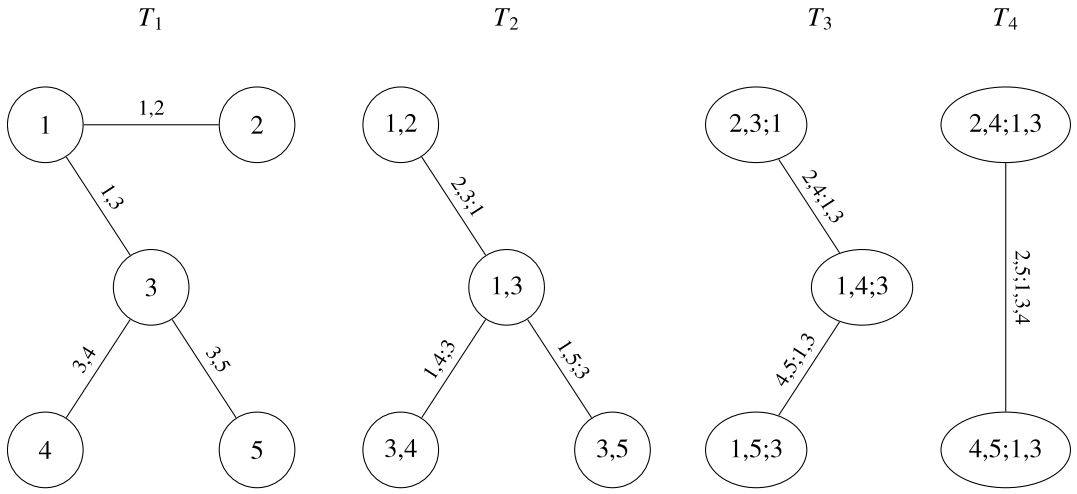


Fig. 1. Example of a regular vine tree sequence.

where $\boldsymbol{u}_{D_e} = (u_\ell)_{\ell \in D_e}$ and $G_{j_e|D_e}$ is the conditional distribution of $U_{j_e}|\boldsymbol{U}_{D_e} = \boldsymbol{u}_{D_e}$. The pair-copula density $c_{j_e,k_e;D_e}$ is the joint density of the bivariate conditional random vector

$$\left(G_{j_e|D_e}(U_{j_e}|\boldsymbol{U}_{D_e}),\, G_{k_e|D_e}(U_{k_e}|\boldsymbol{U}_{D_e})\right) \;|\; \boldsymbol{U}_{D_e} = \boldsymbol{u}_{D_e}.$$

Eq. (2) holds for any copula density, but note that the functions $c_{j_e,k_e;D_e}$ take a third argument $\boldsymbol{u}_{D_e}$. This indicates that, in general, the conditional dependence between U_{j_e} and U_{k_e} may be different for different values of $\boldsymbol{u}_{D_e}$. To increase tractability of the models, it is common to ignore this influence and simplify the model to

$$c(\boldsymbol{u}) = \prod_{m=1}^{d-1} \prod_{e \in E_m} c_{j_e,k_e;D_e}\left\{G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}),\; G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e})\right\}.$$

In this case the pair copula density $c_{j_e,k_e;D_e}\left\{G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}),\, G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e})\right\}$ encodes the partial dependence between U_{j_e} and U_{k_e} conditional on $\boldsymbol{U}_{D_e}$ [20,35]. This so-called simplifying assumption is often valid for financial data and even if it is violated, it may serve as a useful approximation to the truth. For further discussions, see [3,24,26,28,34,36,37], and references therein. Since tractability is vital in very high dimensions, we will assume that the simplifying assumption holds for the remainder of this paper.

3. A modified BIC for sparse vine copula models

Consider a vine copula model with density $c^{\boldsymbol{\eta}}$ that is characterized by a finite-dimensional parameter $\boldsymbol{\eta}$. We denote the number of parameters in the model by ν. To simplify our arguments, we assume that there is only one parameter per pair-copula, i.e., $\boldsymbol{\eta} = (\eta_e)_{e \in E_m, m=1,\dots,d-1}$, and that η_e is zero if and only if the pair-copula at edge e is the independence copula given, for all $u_1, u_2 \in [0,1]$ by $\Pi(u_1, u_2) = u_1 u_2$. However, we formally distinguish between the number of non-zero parameters ν and the number of non-independence pair-copulas q to maintain validity of the formulas in the general context.

The maximal model contains $q_{\max} = d(d-1)/2$ non-independence copulas. A sparse vine copula model contains many independence copulas and, thus, only few non-zero parameters. Let $q = \#\{\eta_e \colon \eta_e \neq 0, e \in \{E_1, \dots, E_{d-1}\}\}$ be the number of non-independence copulas in the model. We speak of a sparse model if $q \ll q_{\max}$ or, asymptotically, $q/q_{\max} \to 0$ as $d \to \infty$.

3.1. The BIC and why it is inappropriate for sparse models

Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n \in [0,1]^d$ be a random sample and $\widehat{\boldsymbol{\eta}}$ be the estimated parameter. The Bayesian information criterion (BIC) is defined as

$$\mathrm{BIC}(\widehat{\boldsymbol{\eta}}) = -2\ell(\widehat{\boldsymbol{\eta}}) + \widehat{\nu} \ln n,$$

where $\ell(\widehat{\boldsymbol{\eta}}) = \ln c^{\widehat{\boldsymbol{\eta}}}(\boldsymbol{u}_1) + \cdots + \ln c^{\widehat{\boldsymbol{\eta}}}(\boldsymbol{u}_n)$ denotes the log likelihood and $\widehat{\nu}$ is the number of non-zero parameters in $\widehat{\boldsymbol{\eta}}$. The lower the BIC, the more favorable the model.

The BIC of a vine copula model decomposes to

$$\mathrm{BIC}(\widehat{\boldsymbol{\eta}}) = -2\sum_{m=1}^{d-1}\sum_{e\in E_m}\sum_{i=1}^{n}\Big[\ln c^{\widehat{\eta}_e}_{j_e,k_e;D_e}\big\{G^{\widehat{\boldsymbol{\eta}}_{D_e}}_{j_e|D_e}(u_{i,j_e}|\boldsymbol{u}_{i,D_e}),\ G^{\widehat{\boldsymbol{\eta}}_{D_e}}_{k_e|D_e}(u_{i,k_e}|\boldsymbol{u}_{i,D_e})\big\} + \widehat{\nu}_e \ln n\Big] = -2\sum_{m=1}^{d-1}\sum_{e\in E_m}\{\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) + \widehat{\nu}_e \ln n\},$$

where $\widehat{\nu}_e$ is the number of parameters for edge e and

$$\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) = \sum_{i=1}^{n} \ln c^{\widehat{\eta}_e}_{j_e,k_e;D_e}\big\{G^{\widehat{\boldsymbol{\eta}}_{D_e}}_{j_e|D_e}(u_{i,j_e}|\boldsymbol{u}_{i,D_e}),\ G^{\widehat{\boldsymbol{\eta}}_{D_e}}_{k_e|D_e}(u_{i,k_e}|\boldsymbol{u}_{i,D_e})\big\}.$$

This decomposition suggests that the global BIC can be minimized by sequentially minimizing the BIC of individual pair-copulas. But there is a subtle issue: the log likelihood ℓ_e for an edge e in edge set E_m of tree T_m, with $m \in \{2, \dots, d-1\}$, also depends on estimated parameters $\widehat{\boldsymbol{\eta}}_{D_e}$ from previous trees. In general, step-wise and global optima may correspond to different models. But both criteria select the same model as $n \to \infty$. This motivates the common practice of selecting the pair-copulas individually in sequential estimation.

The BIC is derived from a Bayesian argument stating that the posterior log probability of a model is proportional to

$$-2\ell(\widehat{\boldsymbol{\eta}}) + \widehat{\nu}\ln n - 2\ln\psi(\widehat{\boldsymbol{\eta}}) + O_p(n), \tag{3}$$

where $\psi(\widehat{\boldsymbol{\eta}})$ is the prior probability of the model $c^{\widehat{\eta}}$ and the $O_p(n)$ term is independent of the model choice. The simpler form of the BIC is then a consequence of the assumption that all models are equally likely a priori. When $q_{\max}$ (hence d) is fixed, it is well known that the BIC selects the true model with probability tending to 1 as $n \to \infty$; see, e.g., [12]. In high-dimensional vine copula models, $q_{\max}$ is often of the same order or much larger than n. In such situations, "large n, fixed d" asymptotics are at least questionable.

Since all models are equally likely, we expect the number of parameters in the model to be $q_{\max}/2$. This stands in contrast to the sparsity assumption $q/q_{\max} \to 0$. In fact, one can show with arguments similar to those in Section 4 that the BIC cannot distinguish the true model from an incorrect one when $q_{\max} \geq \sqrt{n \ln n}$ or equivalently $d \geq \sqrt[4]{n \ln n}$. In terms of the number of variables d, this restriction is much more severe for vine copula models compared to (generalized) linear models, where the BIC has been studied most extensively [42].

3.2. A modified criterion

Several authors have considered modifications of the BIC to make it more suitable for high-dimensional problems [7,10,42]. The unifying idea is to adjust the prior probabilities in Eq. (3) such that sparse models are more likely than dense models. We follow the same path and assign each pair-copula a prior probability ψ_e of not being independence.

More precisely, we assume that the indicators $I_e = \mathbf{1}(\eta_e \neq 0)$ are independent Bernoulli variables with mean ψ_e and propose to choose $\psi_e = \psi_0^m$ for any edge in tree T_m. The resulting prior probability of a vine copula model $c^{\boldsymbol{\eta}}$ is

$$\psi(\boldsymbol{\eta}) = \prod_{m=1}^{d-1} \psi_0^{m q_m}(1-\psi_0^m)^{d-m-q_m},$$

where q_m is the number of non-independence copulas in tree m. Now Eq. (3) suggests the criterion

$$\mathrm{mBICV}(\widehat{\boldsymbol{\eta}}) = -2\ell(\widehat{\boldsymbol{\eta}}) + \widehat{\nu}\ln n - 2\sum_{m=1}^{d-1}\big\{\widehat{q}_m \ln\psi_0^m + (d-m-\widehat{q}_m)\ln(1-\psi_0^m)\big\},$$

where $\ell(\widehat{\boldsymbol{\eta}}) = c^{\widehat{\eta}}(\boldsymbol{u}_1) + \cdots + c^{\widehat{\eta}}(\boldsymbol{u}_n)$ is the log-likelihood and $\widehat{q}_m$ is the number of non-independence copulas in tree m of the fitted model $\widehat{\boldsymbol{\eta}}$. The mBICV decomposes similarly to the BIC,viz.

$$\mathrm{mBICV}(\widehat{\boldsymbol{\eta}}) = \sum_{m=1}^{d-1}\sum_{e\in E_m}\Big[-2\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) + \widehat{\nu}_e \ln n - 2\big\{\widehat{I}_e \ln\psi_0^m + (1-\widehat{I}_e)\ln(1-\psi_0^m)\big\}\Big] = \sum_{m=1}^{d-1}\sum_{e\in E_m}\mathrm{mBICV}_e\big(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}\big), \tag{4}$$

suggesting that the mBICV is a suitable criterion for sequential selection of pair-copulas.

The sequence ψ_0^m is decreasing in m, which implies that higher-order pairs are more likely to be (conditionally) independent. Hence, mBICV penalizes non-independence copulas more severely compared to BIC, but only in higher trees. In the first tree, the prior probability of a non-independence copulas is ψ_0. If $\psi_e > 0.5$, the mBICV is more likely to select a non-independence model. For $\psi_0 = 0.9$, this means that the first seven trees are more favorable to non-independence compared to BIC.

This property is motivated by statistical practice. All popular structure selection algorithms try to capture strong dependence relationships in the first few trees [9,14,24,27]. Although there is no guarantee, this typically leads to models where there is a lot of dependence in the first few trees and relatively little dependence in higher trees. Further, the parameters of a vine copula are typically estimated sequentially starting from the first tree. Because estimation errors

accumulate over the trees, estimates in higher trees are less reliable. Moreover, the less reliable the estimates are, the more conservatively we should choose our model.

Under the prior used for the mBICV, q_m is a binomial experiment with $d-m$ trials and success probability ψ_0^m. Therefore, the expected number of non-independence copulas in tree T_m is $\mathrm{E}(q_m) = (d-m)\psi_0^m$ and the expected total number of non-independence copulas in the entire vine-copula model is

$$\mathrm{E}(q) = \sum_{m=1}^{d-1}(d-m)\psi_0^m = O\left(d\sum_{m=1}^{d-1}\psi_0^m\right) = O(d).$$

Recalling that $q_{\max} \sim d^2$, we obtain $\mathrm{E}(q)/q_{\max} \to 0$ as $d \to \infty$ and, hence, expect the model to be sparse.

A referee noted that there are other choices for ψ_e that correspond to sparse models. For example, $\psi_e = \psi_0/m$ leads to $\mathrm{E}(q) = O(d \ln d)$ which induces sparsity at a slightly slower rate. In that sense, our choice $\psi_e = \psi_0^m$ is somewhat arbitrary. In the following theoretical arguments the two choices are essentially equivalent. However, given that sparser models are also computationally more tractable, inducing sparsity at a faster rate seems reasonable.

4. Asymptotic properties

We will argue that the mBICV can consistently distinguish between a finite number of models provided that $d = o(\sqrt{n})$, which is a less stringent condition compared to BIC. This is only a weak form of consistency: when $d \to \infty$, it does not automatically imply that the criterion finds the best among all possible candidate models. Because the number of possible candidates grows rapidly with d, results on this stronger form of consistency are considerably more difficult to derive and have only recently emerged in the simpler context of linear models [17,38,39].

Unfortunately, asymptotic results for parameter estimates in vine copula models of diverging size are not yet available. For the Gaussian copula, Xue and Zou [41] showed that the parameters can be estimated consistently when d diverges, although at a rate slower than $O_p(n^{-1/2})$. The Gaussian copula is a special case of a vine copula, where all pair-copulas are Gaussian and η_e equals the partial correlation $\rho_{j_e,k_e;D_e}$. Some general results for (joint) maximum-likelihood estimation were derived in [16], but they do not cover the sequential estimation method that is typically used for vine copula models.

This lack of results makes it impossible to formally establish consistency of the mBICV. Instead, we give a semi-formal argument that explains why we expect the mBICV to be consistent when $d = o(\sqrt{n})$ and substantiate this claim by a small simulation experiment. We will consider the decomposition in Eq. (4) and approximate the probabilities of selecting a wrong model for a fixed edge in tree T_m with $m < \infty$. The resulting probabilities of selecting the wrong model for a given edge are then aggregated and shown to vanish asymptotically. Since we do not approximate the error probabilities uniformly in m and neglect lower-order terms when applying the Central Limit Theorem, the argument is only heuristic and should be taken with a grain of salt.

4.1. An informal argument

Our null hypothesis is that the true pair-copula at edge $e \in E_m$ corresponds to independence, i.e., $\eta_e = 0$. Denote by $\alpha_{n,e}$ the probability of a type I error, i.e., of selecting the non-independence model $\eta_e \neq 0$ although the true model has $\eta_e = 0$, viz.

$$\alpha_{n,e} = \Pr\left\{\mathrm{mBICV}_e\left(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}\right) < \mathrm{mBICV}_e\left(0, \widehat{\boldsymbol{\eta}}_{D_e}\right) \mid \eta_e = 0\right\},$$

and by $\beta_{n,e}$ the probability of a type II error, i.e., of selecting the independence model $\eta_e = 0$ although the true model has $\eta_e \neq 0$, viz.

$$\beta_{n,e} = \Pr\left\{\mathrm{mBICV}_e\left(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}\right) > \mathrm{mBICV}_e\left(0, \widehat{\boldsymbol{\eta}}_{D_e}\right) \mid \eta_e \neq 0\right\}.$$

In Appendix we argue that for any $e \in E_m$ with $m < \infty$, the error probabilities are approximately

$$\alpha_{n,e} = O\left(\frac{\psi_0^m}{\sqrt{n \ln n}}\right), \quad \beta_{n,e} = O\left(\frac{e^{-n\mathrm{MI}_e^2}}{\sqrt{n}\mathrm{MI}_e}\right), \tag{5}$$

where

$$\mathrm{MI}_e = \int_{[0,1]^2} c^{\eta_e}_{j_e,k_e;D_e}(u,v) \ln c^{\eta_e}_{j_e,k_e;D_e}(u,v) du dv,$$

is the mutual information associated with edge e.

The type I error probability $\alpha_{n,e}$ is increasing in ψ_0^m, the prior probability of a non-independence model in tree T_m. Lower values shift our expectations to sparser models and the probability of overfitting decreases. In particular, since ψ_0^m is decreasing in the tree level m, so is $\alpha_{n,e}$. This is related to the fact that the mBICV expects more independence copulas at higher tree levels. The type II error probability $\beta_{n,e}$ decreases in the mutual information MI_e which confirms our intuition that stronger dependence relationships are easier to detect. The rate for $\beta_{n,e}$ is unlikely to hold uniformly in m, since our argument assumes $\|\widehat{\boldsymbol{\eta}}_{D_e} - \boldsymbol{\eta}_{D_e}\|_2 = O_p(n^{-1/2})$, which is only valid when m is fixed. However, $\beta_{n,e}$ vanishes much faster

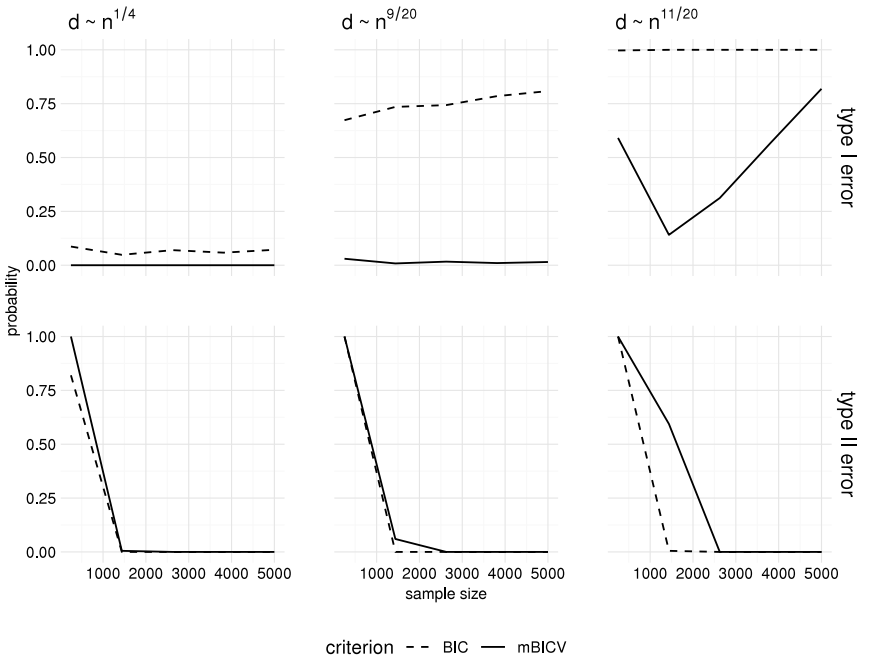


Fig. 2. Family-wise error rates with simultaneously increasing sample size and dimension in three asymptotic regimes.

than $\alpha_{n,e}$ and we conjecture that this is still the case when taking $m \to \infty$. The qualitative interpretation remains valid in any case.

The error probabilities in Eq. (5) can now be aggregated over the whole vine. Consider the family-wise error rates (FWER), i.e., the probability of selecting the wrong model for at least one edge in the model. Define α_n as the FWER of a type I error and similarly β_n for the type II error. Using Bonferroni's inequality, we get

$$\alpha_n \leq \sum_{m=1}^{d-1} \sum_{e \in E_m : \eta_e = 0} \alpha_{n,e}, \quad \beta_n \leq \sum_{m=1}^{d-1} \sum_{e \in E_m : \eta_e \neq 0} \beta_{n,e}.$$

By expanding the sum and using Eq. (5), we obtain

$$\alpha_n \leq \frac{1}{\sqrt{n \ln n}} \sum_{m=1}^{d-1} (d-m)\psi_0^m = O\left(\frac{d}{\sqrt{n \ln n}}\right).$$

Similarly, we get $\beta_n = O(d^2 e^{-nMI_e^2}/\sqrt{nMI_e}) = o(\alpha_n)$. Combining the results for type I and type II errors suggests that the mBICV is consistent when $d = o\left(\sqrt{n \ln n}\right)$. However, our approximation for the type I error requires $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 = o_p(1)$. Each non-zero parameter can be estimated at best at rate $O_p(n^{-1/2})$ which suggests that $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 = O_p(d/\sqrt{n})$. Hence, we must strengthen our condition for consistency of the mBICV to $d = o(\sqrt{n})$.

4.2. Simulation

Since the arguments above are only heuristic, we conduct a small simulation experiment to check whether the condition $d = o(\sqrt{n})$ is reasonable. We consider a simple setup where the true model contains only Gaussian pair-copulas. Half of the pair-copulas in every tree are set to independence ($\eta_e = 0$), all others have parameter $\eta_e = 0.2$. Note that this model corresponds to the implicit prior probabilities of the BIC and is not sparse. We simulate data sets of increasing sample size n and let the dimension d grow with n at three different rates. We consider five sample sizes and simulate 1000 data sets each. We always fit two models: one selects the copula families based on BIC, the other based on mBICV.

Fig. 2 shows the probabilities of selecting at least one pair-copula incorrectly, differentiating between type I (α_n) and type II (β_n) errors. The three columns correspond to the three asymptotic regimes $d \sim n^{1/4}$, $d \sim n^{9/20}$, and $d \sim n^{11/20}$, respectively. First, we observe that the type II error probabilities vanish fast for both criteria. This is in line with our asymptotic arguments that suggested that type II errors are much easier to avoid. The type I error probability is the main factor determining consistency of the criteria.

In the first regime, $d \sim n^{1/4}$, we expect both BIC and mBICV to consistently select the true model as n becomes large. The type I error probability of the mBICV is zero for every sample size under consideration. The error probability of the

BIC decreases slowly, which is also in line with our theoretical considerations: the BIC is consistent when $d = o(\sqrt[4]{n \ln n})$ and $d \sim n^{1/4}$ is close to the border of that range. In the second regime, we still expect the mBICV to be consistent since $d \sim n^{9/20} = o(\sqrt{n})$, but not the BIC. This is supported by the simulations: while the type I error probability vanishes for mBICV, it is increasing for the BIC. Finally, we expect both criteria to be inconsistent for $d \sim n^{11/20}$ and, indeed, the type I error probabilities tend to one for both criteria. In summary, the simulations support our hypothesis that the mBICV is consistent when $d = o(\sqrt{n})$.

5. Special classes of sparse vine copula models and their selection

The mBICV can be used to decide which pair-copulas in the model are set to be independence copulas. But in order to calculate the mBICV, one first needs to estimate a model. In high-dimensional vine copula models, there is a huge number of pair-copulas. Estimating all of them will be computationally demanding. This predicament is solved by focusing on sparse model classes that set a large proportion of pair-copulas to be independence copulas before a model is fit. Below we discuss two such classes, truncated and thresholded vine copula models. Both classes have a sparsity inducing hyper-parameter that can be selected using mBICV.

5.1. Truncated vine copulas

Truncated vine copula models induce sparsity by setting all pair-copulas after a certain tree level $M \in \{1, \dots, d-1\}$ to be the independence copula. The lower the truncation level M, the higher the degree of sparsity: an M-truncated vine copula model allows for only $M(2d - M - 1)/2$ non-independence copulas. Since the density of the independence copula is 1 everywhere, the density of an M-truncated vine copula model can be written as

$$c(\boldsymbol{u}) = \prod_{m=1}^{M} \prod_{e \in E_m} c_{j_e,k_e;D_e}\left\{G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}),\ G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e})\right\}.$$

The logic behind truncated vine copulas is closely related to the structure selection heuristic of Dißmann et al. [14]. Its goal is to capture most of the dependence in the first couple of trees. If this allows to capture all dependence in the first M trees, the truncated model arises naturally. If this is not the case, one can at least hope that the dependence in higher trees is practically irrelevant.

5.2. Thresholded vine copulas

The idea of thresholded regular vines is different. The ultimate goal is to set all conditional pair-copulas that are practically irrelevant to be independence copulas. Arguably, 'practical relevance' is a vague concept and we need to rely on a proxy measure for it. A natural choice are measures for the strength of dependence. One such measure is Kendall's τ. It is a measure of concordance and, as such, is a functional of the copula only. In particular, it does not depend on the marginal distributions [19,30]. In the remainder of this article we take Kendall's τ as our target measure, although other choices are equally valid.

Similar to the truncation level M before, thresholded vine copulas have a hyper-parameter θ, called threshold. Denote the Kendall's τ associated with the pair-copula $c_{j_e,k_e;D_e}$ by τ_e and define $E_m^\theta = \{e \in E_m : |\tau_e| > \theta\}$ for any edge set E_m in the vine. Then the density of a thresholded vine copula model becomes

$$c(\boldsymbol{u}) = \prod_{m=1}^{d-1} \prod_{e \in E_m^\theta} c_{j_e,k_e;D_e}\left\{G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}),\ G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e})\right\}.$$

While it would be possible to use a different threshold in each tree, we will only consider the simple case where θ is constant across all tree levels.

The number of non-independence copulas in the model can be controlled by the threshold parameter θ. But in contrast to the truncated model, that number also depends on the actual dependence in the random vector $\boldsymbol{U}$. To illustrate this, we fix θ and consider a few interesting boundary cases:

(a) If $|\tau_e| > \theta$ for all $e \in E_m$ with $m \in \{1, \dots, d-1\}$, the thresholded model is equal to the full model.
(b) If for $M \in \{1, \dots, d-2\}$, one has $|\tau_e| > \theta$ for all $e \in E_m$ with $m \in \{1, \dots, M\}$ and $|\tau_e| \leq \theta$ for all $e \in E_m$, $m \in \{M+1, \dots, d-1\}$, the thresholded model is equal to the M-truncated model.
(c) If $|\tau_e| \leq \theta$ for all $e \in E_m$ with $m \in \{1, \dots, d-1\}$, the thresholded model contains only independence pair-copulas and is thus equal to the independence model in which the d variables are mutually independent.

The thresholded model generalizes the truncated model: whenever a model is truncated at level M, it can be represented as a thresholded model with $\theta < \min_{e \in E_1, \dots E_M : \tau_e \neq 0} |\tau_e|$. If a model is M-truncated, one has $\tau_e = 0$ for all $e \in E_m$ with $m > M$. If we choose, e.g., $\theta = \min_{e \in E_1, \dots E_M : \tau_e \neq 0} |\tau_e|/2$, then $\theta < |\tau_e|$ for all non-independence copulas up to tree M and $\theta > |\tau_e|$ for all remaining pairs. Hence, the thresholding operation keeps all non-independence copulas of the

truncated model and sets all others to be independence copulas. If the original model is not truncated, the thresholded model can still adapt to the sparsity patterns of the truth and is therefore more flexible.

The thresholding idea is not new. Several authors used a similar approach, but tied the value of θ to the critical value of a significance test for an empirical version of Kendall's τ [8,13,14]. The idea is to perform a statistical test for independence and set the pair-copula to independence whenever the null hypothesis cannot be rejected. This procedure is only heuristic: in general, the test based on the empirical Kendall's τ is not consistent for the null hypothesis of independence and there is no correction for multiple testing. Treating the threshold as a free hyper-parameter brings additional flexibility and allows to tailor the threshold to a specific application.

5.3. Automatic hyper-parameter selection based on mBICV

The mBICV allows us to compare models for various thresholds and decide which is the best. The natural way to select the best model is to fit several models for a fine grid of θ values and select the one with lowest mBICV. This strategy can be extremely time-consuming in high dimensions, where a single fit of the full model often takes hours. An advantage of sparse vine copula models is that only a fraction of the pair-copulas needs to be estimated; all others are set to be independence copulas. This motivates the following strategy to automatically select the threshold parameter θ:

1. Start with a large threshold and fit the model.
2. Reduce the threshold and fit a new model.
3. If the new model improves the mBICV, continue with Step 2. Stop if there is no improvement.

This approach has several computational advantages. Since we start with a large threshold, only a few pair-copulas have to be estimated in the first iteration. For all following iterations, only a few pair-copulas will change from one iteration to the next. By keeping the result from the previous fit in memory, we can re-use most of the fitted pair-copulas and only need to estimate a few additional parameters. Upon termination of the algorithm, most of the non-independent pair-copulas have only been estimated once. Thus, the overall time spent on model fitting and selection will be comparable to the time required for fitting only the mBICV-optimal model. Depending on the dimension d and the level of sparsity in the data, this can be several times faster than fitting the full model.

The same strategy can be used for selecting the truncation level: start with a low truncation level and gradually add more tree levels until the mBICV fails to improve. It is similarly straightforward to select the threshold and the truncation level simultaneously by using an outer loop for the threshold and an inner loop for the truncation level.

5.4. Implementation

We propose to reduce the threshold in a data-driven manner. For the initial threshold, we choose the maximum of all pair-wise absolute empirical Kendall's τ. In this case, the initial model consists of only independence copulas and mBICV $= 0$. To reduce the threshold from one iteration to the next, we choose the threshold such that about 5% of the previously independent pairs may become non-independent in the new model. To be more precise, let θ_k denote the threshold in iteration k, $\mathcal{T}_k = \{\tau_e : |\tau_e| \leq \theta_k\}$ and $N_k = |\mathcal{T}_k|$. Then we set θ_{k+1} to the $\lceil 0.05 N_k \rceil$th largest value in $\mathcal{T}_k$.

To check if a pair-copula from the last iteration can be re-used, we need to check if the pseudo-observations have changed. In very high-dimensional models it is also important to make efficient use of memory. We should not store the pseudo-observations of each conditional pair in the vine. Because of the large number of pair-copulas, this will quickly exceed the memory of common computers. Instead, one can store a summary statistic (like a weighted sum) of the pseudo-observations and only check if this summary statistic has changed.

An open-source implementation of the selection algorithm in C++ is available as part of the `vinecopulib` library and its interface to R [29].

6. Case study: Modeling dependence in a large stock portfolio

To illustrate the concepts introduced in the previous section, we use the new methodology to model the dependence between a large number of stocks. Our data set contains daily stock returns from the S&P 100 constituents from the period of January 1, 2010 to December 31, 2016 ($n = 1756$). Only stocks that were traded over the whole period are included, leaving us with $d = 96$.

6.1. Modeling

Models for the marginal time series

A stylized fact about stock returns is that the squared time-series exhibit strong inter-serial dependence. A popular model that takes this fact into account is the ARMA-GARCH model [18]. Let $x_{t,k}$ be the return of stock k at time t. The ARMA(1, 1)-GARCH(1, 1) model for this time series is

$$x_{t,k} = \mu_k + \phi_k x_{t-1,k} + \psi_k a_{t-1,k} + a_{t,k}, \quad a_{t,k} = \sigma_{t,k} \epsilon_{t,k},$$

$$\sigma_{t,k}^2 = \omega_k + \beta_k \sigma_{t-1,k}^2 + \alpha_k \epsilon_{t-1,k}^2,$$

where $\epsilon_{t,k}$ are iid Student t variables with zero mean, unit variance, and ν_k degrees of freedom. The parameters of this model can be estimated with maximum likelihood, for example using the R package `fGarch` [40]. In addition to parameter estimates, this also gives us estimates $\widehat{\epsilon}_{t,k}$ and $\widehat{\sigma}_{t,k}$ of the unobserved innovation process $\epsilon_{t,k}$ and volatility process $\sigma_{t,k}$.

Dependence model

The cross-sectional dependence between stocks is modeled by a vine copula for the residual time series $\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \dots, \epsilon_{t,d})$ with $t \in \{1, \dots, T\}$. We assume that the dependence in $\boldsymbol{\epsilon}_t$ is induced by a vine copula model C. Define $u_{t,k} = \Psi(\epsilon_{t,k}; \nu_k)$, where $\Psi(\cdot; \nu_k)$ is the cumulative distribution function of a Student t distribution with zero mean, unit variance, and ν_k degrees of freedom. Then the variables $\boldsymbol{u}_t = (u_{t,1}, \dots, u_{t,d})$ with $t \in \{1, \dots, T\}$ are independent observations from a random vector with distribution C.

In practice, we do not observe $\boldsymbol{u}_t$. Suppose that $\widehat{\boldsymbol{\epsilon}}_t$ are the observed residual series and $\widehat{\nu}_k$ are the estimated degrees of freedom parameters of the fitted ARMA-GARCH models. Then the statistics $\widehat{u}_{t,k} = \Psi(\widehat{\epsilon}_{t,k}; \widehat{\nu}_k)$ with $k \in \{1, \dots, d\}$ and $t \in \{1, \dots, T\}$ act as pseudo-observations of the copula C. Based on these, a thresholded vine copula model can be estimated for any fixed value of the threshold θ. We only allow for parametric pair-copulas, and choose the family of each pair-copula from the Gaussian, Student t, Clayton, Gumbel, Frank, and Joe families by the mBICV criterion. The prior probability ψ_0 is set to 0.9 and the vine structure is selected by the algorithm of Dißmann et al. [14].

6.2. Illustration of the new concepts

To assess the influence of the sparsity hyper-parameters in thresholded and truncated vine copula models, we will first estimate the marginal and dependence models using the full data from January 1, 2010 to December 31, 2016. This allows us to make a direct connection between the hyper-parameters and the mBICV, which helps us to understand their relation. In particular, we fit vine copula models to the pseudo-observations $\widehat{\boldsymbol{u}}_1, \dots, \widehat{\boldsymbol{u}}_T$ for a range of values for the hyper-parameters and investigate how they affect the mBICV, out-of-sample likelihood, model sparsity, and computation time. The results are shown in Fig. 3(a), where thresholded models are in the left and truncated models in the right column. To separate the effect of the selection criterion and hyper-parameters, Fig. 3(b) shows the same results when the BIC is used for selection.

Optimal hyper-parameters and models

The mBICV is shown in the upper panel of Fig. 3(a). The mBICV is a convex function of the threshold θ and has a minimum at $\theta \approx 0.05$. The value $\theta_{\text{auto}} \approx 0.043$ is the threshold selected by the automatic algorithm proposed in Section 5.3 and is shown as a vertical dotted line. The second row shows the out-of-sample log-likelihood obtained based on 5-fold cross-validation. It has a peak at $\theta = 0.045$, which is close to the mBICV-optimal parameter. This suggests that the mBICV finds just the right balance between model complexity and fit. Similarly, the mBICV is a convex function of the truncation level with a minimum at $M = 25$. The truncation level selected by the automatic procedure is $M_{\text{auto}} = 26$. The cross-validated log-likelihood has a maximum at $M = 28$, which is only slightly larger.

Fig. 3(b) shows the same graphs but using BIC as a selection criterion. We observe that the BIC-optimal models and the ones selected automatically are sub-optimal in terms of cross-validated log-likelihood which confirms that the BIC is inappropriate for high-dimensional models.

Level of sparsity

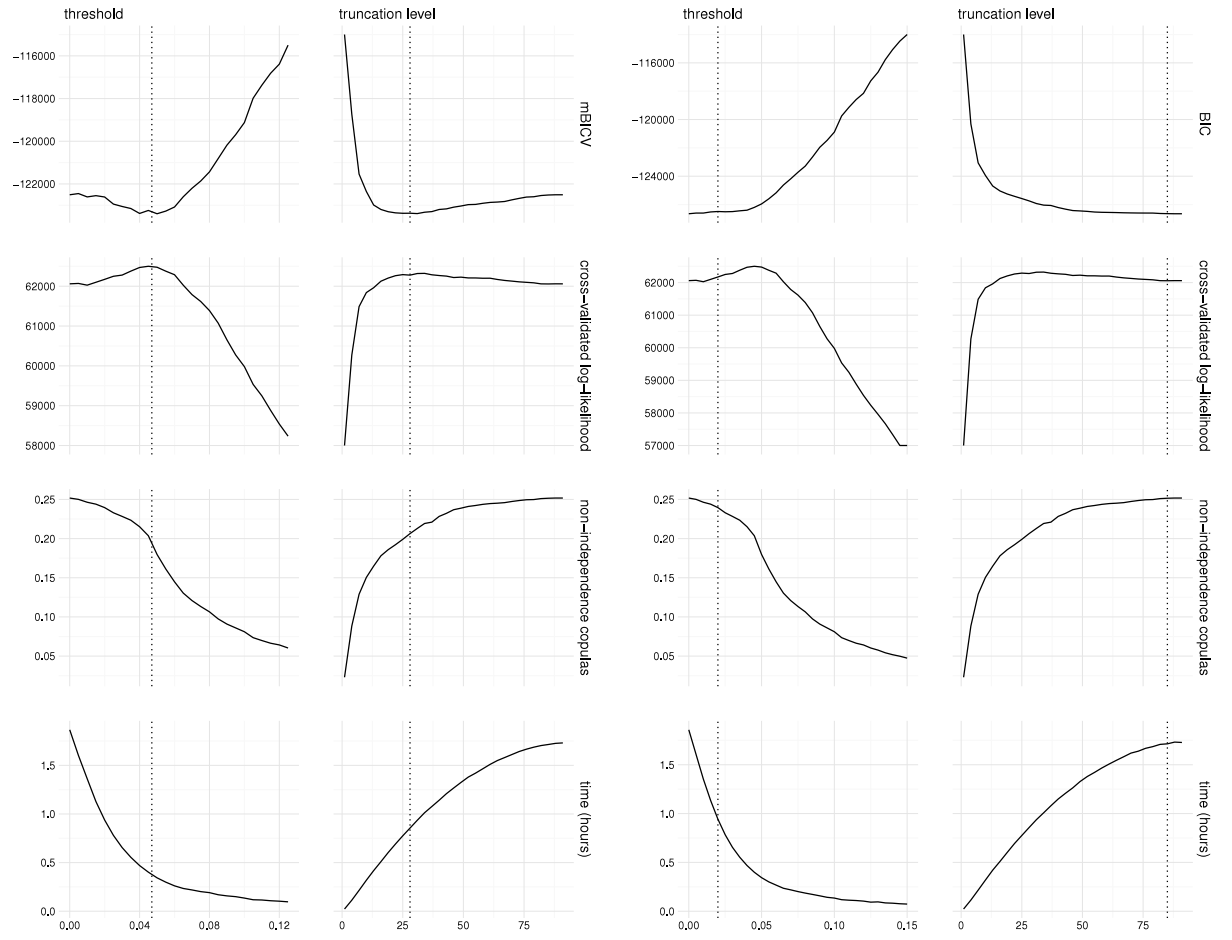
The sparsity of a vine copula model is characterized by the number of independence copulas. The third row in Fig. 3 shows the proportion of non-independence copulas in the model ($q/q_{\max}$) as function of the hyper-parameters. The proportion of non-independence copulas decreases with the threshold and increases with the truncation level. For both thresholded and truncated models, the mBICV selects models with fewer non-independence copulas (approximately 20%) compared to BIC (approximately 24%). The proportions are quite close comparing the thresholded and truncated models, but the truncated model is slightly larger. This again relates to the fact that thresholded models generalize truncated ones.

Computation time

As explained in Section 5, sparse vine copulas have a side benefit in terms of computation time, because a large number of pair-copulas are directly set to be independence copulas and thus never estimated. We can see this effect in the lower panels of Fig. 3, where the hyper-parameters are plotted against the time required to fit the model. All times reported here were recorded on a single thread of an 8-way Opteron (Dual-Core, 2.6 GHz) CPU with 64 GB RAM.

The larger the threshold, the less time it takes to fit the model. The full model ($\theta = 0$) takes almost two hours, the BIC-optimal thresholded model takes roughly one hour; the mBICV-optimal model takes only 25 min. However, there is an overhead for selecting the threshold. Finding the mBICV-optimal threshold and fitting the model takes 35 min in total, which is still considerably faster than fitting the full or BIC-optimal models. The difference between the two can be expected to increase with the dimension and level of sparsity.

Similarly, truncated models with a larger truncation level take more time to fit. The mBICV-optimal truncated model ($M = 26$) and the one selected by the automatically take a bit less than one hour; for truncated models, there is no



(a) selection based on mBICV (b) selection based on BIC

Fig. 3. The selection criterion, cross-validated log-likelihood, proportion of non-independence copulas, and time required to fit the model as functions of the sparsity hyper-parameters. The dotted lines indicate the values of the hyper-parameter that have been selected by the automatic selection procedure from Section 5.3.

overhead for selection. This is much less than for the full model ($M = 94$, 1.7 h) or the BIC-optimal one ($M = 88$, 1.7 h). But it is twice the time taken to fit the thresholded model. This can be explained by the fact that even in the first 25 trees there are many independence copulas that were ruled out in the optimal thresholded model.

We conclude that despite the overhead of threshold selection, the mBICV-optimal thresholded model is much faster to fit than BIC-optimal or truncated models.

6.3. Out-of-sample value-at-risk forecasts

We now analyze the models' ability to accurately forecast the Value-at-Risk out-of-sample. We fit the marginal and vine copula models on a training period $t \in \{t^* - T_{\text{train}}, \dots, t^*\}$, and predict the Value-at-Risk for subsequent days $t \in \{t^*+1, \dots, t^*+T_{\text{test}}\}$. To ensure that a reasonable amount of data is available, we choose $T_{\text{train}} = 1260$ (five years) and $T_{\text{test}} = 252$ (one year). After each year, the marginal and vine copula models are fit again to the previous five years of data. To make the results comparable to the in-sample analysis, we want to use the same period for evaluating out-of-sample forecasts. To achieve this, we augment the data by five more years of data for the period 2005–2009. Four more stocks have to be dropped, leaving us with $d = 92$.

The fitted models can be used to forecast the one-day-ahead Value-at-Risk (VaR) of an equally weighted portfolio of all $d = 96$ stocks. The portfolio return at time t is then $y_t = (x_{t,1} + \cdots + x_{t,d})/d$. The theoretical α-level portfolio VaR on day $t + 1$ is defined as the $(1 - \alpha)$-quantile of y_{t+1}, for $\alpha \in (0, 1)$, viz.

$$\text{VaR}_{t+1,\alpha} = \inf\{y \in \mathbb{R} \colon \Pr(y_{t+1} \leq y) \geq 1 - \alpha\}.$$

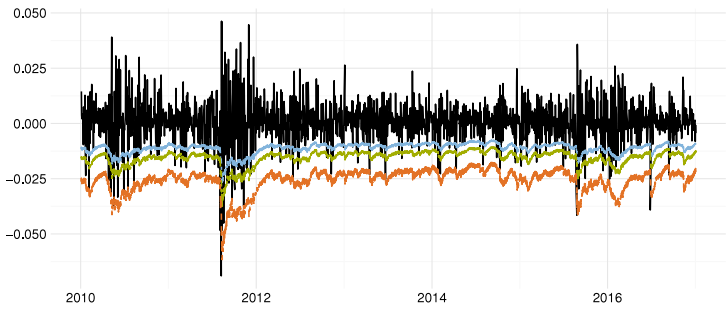


Fig. 4. Time series of an equally weighted portfolio of S&P 100 stocks, and VaR predictions at the 90%, 95%, and 99% levels.

Table 1
Exceedance frequencies of out-of-sample Value-at-Risk forecast and p-values of the conditional coverage test of Christoffersen [11] for three models: the one selected by BIC, and thresholded and truncated models selected by the procedure described in Section 5.3.

α	BIC			θ_{auto}			M_{auto}		
	0.900	0.950	0.990	0.900	0.950	0.990	0.900	0.950	0.990
Exceedances	0.099	0.058	0.015	0.099	0.058	0.012	0.104	0.061	0.016
p-value	0.644	0.149	0.120	0.605	0.135	0.055	0.653	0.082	0.056

Our goal is to forecast $\text{VaR}_{t+1,\alpha}$, based on fitted marginal and vine copula models, and only using past observations of the processes $\widehat{\epsilon}_t$ and $\widehat{\sigma}_t$. We proceed as follows:

1. Simulate $\boldsymbol{u}_{t+1}^{(r)}$, $r \in \{1, \dots, R\}$ with $R = 10^6$, from the fitted vine copula model.
2. For all $k \in \{1, \dots, d\}$ and $r \in \{1, \dots, R\}$, set

$$\epsilon_{t+1,k}^{(r)} = \Psi^{-1}(u_{t+1}^{(r)}; \widehat{\nu}_k), \quad \widehat{\sigma}_{t+1,k} = \widehat{\omega}_k + \widehat{\beta}_k \widehat{\sigma}_{t,k}^2 + \widehat{\alpha}_k \widehat{\epsilon}_{t,k}^2,$$

$$x_{t+1,k}^r = \widehat{\mu}_k + \widehat{\phi}_k x_{t,k} + \widehat{\psi}_k \widehat{a}_{t,k} + \widehat{\sigma}_{t+1,k} \epsilon_{t+1,k}^{(r)}, \quad y_{t+1}^r = \frac{1}{d} \sum_{k=1}^{d} x_{t+1,k}^r.$$

3. The forecast $\widehat{\text{VaR}}_{t+1,\alpha}$ is the empirical α-quantile of $y_{t+1}^1, \dots, y_{t+1}^R$.

Note that although the models are fitted only on the returns in the training period $\{t^* - T_{\text{train}}, \dots, t^*\}$, also all observations in the period $\{t^* + 1, \dots, t\}$ enter the one-day-ahead prediction at time $t + 1$ through the residual $\widehat{\epsilon}_t$.

Fig. 4 shows the observed time series y_t along with VaR predictions from an illustrative model. In the following we focus on three models: thresholded and truncated models selected by the automatic procedure from Section 5 (θ_{auto} and M_{auto}), and the non-sparse model selected by BIC as a baseline. We consider three levels for the Value-at-Risk, $\alpha \in \{0.9, 0.95, 0.99\}$. The second row of Table 1 shows the frequencies with which the Value-at-Risk forecasts were exceeded. For an optimal model we expect these frequencies to be $1 - \alpha$, which, along with independence of exceedance indicators, is the null hypothesis of the conditional coverage test [11]. Corresponding p-values are listed in the third row.

We observe that all models show exceedance frequencies close to our expectation. This is also reflected in the p-values, which indicate that no model can be rejected. However, the two sparse models are not only less complex, they were also substantially faster to fit. These two benefits would become even more striking for portfolios of larger size.

7. Conclusion

This article was concerned with model selection in high-dimensional vine copula models. We proposed the mBICV as a selection criterion tailored to sparse vine copula models. It can be used to sequentially select individual pair-copula or automatically select hyper-parameters in sparse model classes. The benefits of the mBICV were illustrated by a case study modeling the dependence in a large stock portfolio. The mBICV-optimal sparse models were shown to produce valid out-of-sample forecasts for the Value-at-Risk and to be computationally more efficient than models selected by BIC.

We took a first step towards "large n, diverging d" asymptotics in vine copula models by arguing that the selection criterion is consistent when $d = o(\sqrt{n})$. This claim is supported by simulations, but our arguments are only informal due to a lack of results on the behavior of parameter estimates as $d \to \infty$. A formal study of such properties is an interesting direction for future research.

Acknowledgments

The first and third author were partially supported by the Deutsche Forschungsgemeinschaft (DFG grants CZ 86/5-1 and CZ 86/4-1). The authors gratefully acknowledge the compute and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de). The authors are especially grateful to two anonymous referees for many valuable comments and ideas.

Appendix. Approximation of error rates

A.1. Preliminaries

Since we consider an edge $e \in E_m$ for a fixed tree level $m < \infty$, we will argue as if the dimension d were fixed. In this case, the usual asymptotic properties of the sequential MLE were established in Hobæk Haff [21] for cases where the margins are estimated parametrically or based on ranks. In particular the estimator is asymptotically normal and satisfies $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 = O_p(n^{-1/2})$. This rate of convergence will only be used for type II error rates, while $o_p(1)$ is sufficient for type I error rates. We further use the following tail bound for the normal distribution

$$\Pr(Z > x) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp(-x^2/2), \tag{A.1}$$

where $Z \sim \mathcal{N}(0, 1)$.

The independence model, $c^0_{j_e,k_e;D_e} \equiv 1$, yields

$$\mathrm{mBICV}_e\big(0, \widehat{\boldsymbol{\eta}}_{D_e}\big) = -2\ln(1 - \psi_0^m).$$

Therefore, the alternative model $c^{\widehat{\eta}_e}_{j_e,k_e;D_e}$ is selected if

$$2\ell_e\big(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}\big) > \widehat{\nu}_e \ln n - 2\ln \psi_0^m + 2\ln(1 - \psi_0^m).$$

A.2. Type I error rates

We start by approximating the type I error $\alpha_{n,e}$. Let $\widehat{\eta}_e$ be the sequential MLE and $\eta_e = 0$. One has

$$\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) = \ell_e(\eta_e, \widehat{\boldsymbol{\eta}}_{D_e}) + \partial_{\eta_e}\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e})(\widehat{\eta}_e - \eta_e) + \partial^2_{\eta_e}\ell_e({\eta_e}^*, \widehat{\boldsymbol{\eta}}_{D_e})(\widehat{\eta}_e - \eta_e)^2,$$

where η_e^* lies between 0 and $\widehat{\eta}_e$. The first term on the right is zero because $c^0_{j_e,k_e;D_e} \equiv 1$ for $\eta_e = 0$. The second term is zero by the definition of the sequential MLE. Hence,

$$\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) = \partial^2_{\eta_e}\ell_e({\eta_e}^*, \widehat{\boldsymbol{\eta}}_{D_e})(\widehat{\eta}_e - \eta_e)^2 = E\{\partial^2_{\eta_e}\ell_e(\eta_e, \boldsymbol{\eta}_{D_e})\}(\widehat{\eta}_e - \eta_e)^2\{1 + o_p(1)\},$$

where the second equality follows from the Law of Large Numbers and the consistency of $\widehat{\boldsymbol{\eta}}$. Now $\mathrm{E}\{\partial^2_{\eta_e}\ell_e(0, \boldsymbol{\eta}_{D_e})\}^{1/2}(\widehat{\eta}_e - \eta_e)$ converges in distribution to a standard normal variable, $\mathcal{N}(0, 1)$, yielding

$$\alpha_{n,e} \approx 2\Pr\left\{Z > \sqrt{\widehat{\nu}_e \ln n - 2\ln \psi_0^m + 2\ln(1 - \psi_0^m)}\right\},$$

where $Z \sim \mathcal{N}(0, 1)$. Using (A.1) together with the fact that ψ_0^m lies in $(0, 1)$ and is decreasing in m yields

$$\alpha_{n,e} \approx \frac{2}{\sqrt{2\pi n}} \frac{\psi_0^m(1 - \psi_0^m)^{-1}}{\sqrt{\ln n - 2\ln \psi_0^m + 2\ln(1 - \psi_0^m)}} \leq \frac{2}{\sqrt{2\pi n}} \frac{\psi_0^m(1 - \psi_0)^{-1}}{\sqrt{\ln n + 2\ln(1 - \psi_0)}} = O\left(\frac{\psi_0^m}{\sqrt{n \ln n}}\right).$$

A.3. Type II error rates

For the type II error, we have

$$\beta_{n,e} = \Pr\big\{2\ell_e\big(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}\big) < \widehat{\nu}_e \ln n - 2\ln \psi_0^m + 2\ln(1 - \psi_0^m) \mid \eta_e \neq 0\big\}.$$

We again begin with a Taylor expansion, viz.

$$\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) = \ell_e(\eta_e, \boldsymbol{\eta}_{D_e}) + \nabla_{\boldsymbol{\eta}}\ell_e(\eta_e, \boldsymbol{\eta}_{D_e})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^\top \nabla^2_{\boldsymbol{\eta}}\ell_e({\eta_e}^*, \boldsymbol{\eta}^*_{D_e})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}).$$

Since $\mathrm{E}\{\nabla_{\boldsymbol{\eta}}\ell_e(\eta_e, \boldsymbol{\eta}_{D_e})\} = 0$ by definition, one has $\nabla_{\boldsymbol{\eta}}\ell_e(\eta_e, \boldsymbol{\eta}_{D_e})/n = O_p(n^{-1/2})$ by the Law of Large Numbers. Further, the quadratic term is of order $O_p(1)$ given that $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 = O_p(n^{-1/2})$, whence

$$\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) = \ell_e(\eta_e, \boldsymbol{\eta}_{D_e}) + O_p(1).$$

We apply the Central Limit Theorem and Slutsky's Lemma to the right-hand side, yielding

$$\frac{1}{\sqrt{n}}\{\ell_e(\widehat{\eta}_e, \widehat{\boldsymbol{\eta}}_{D_e}) - \sqrt{n}\mathrm{MI}_e\} \rightsquigarrow \mathcal{N}(0, \sigma_e^2),$$

where $\rightsquigarrow$ denotes convergence in distribution. Noting

$$\sqrt{n}\,\mathrm{MI}_e - \frac{\widehat{\nu}_e \ln n - 2\ln\psi_0^m + 2\ln(1-\psi_0^m)}{2\sqrt{n}\sigma_e} = \sqrt{n}\,\mathrm{MI}_e\{1 + o(1)\},$$

and an application of the tail bound (A.1) suggests

$$\beta_{n,e} \approx \Pr\left(Z > -\sqrt{n}\mathrm{MI}_e\right) = O\left(\frac{e^{-n\mathrm{MI}_e^2}}{\sqrt{n}MI_e}\right).$$

References

[1] K. Aas, Pair-copula constructions for financial applications: a review, Econometrics 4 (2016) 43.
[2] K. Aas, C. Czado, A. Frigessi, H. Bakken, Pair-copula constructions of multiple dependence, Insurance Math. Econom. 44 (2009) 182–198.
[3] E.F. Acar, C. Genest, J. Nešlehová, Beyond simplified pair-copula constructions, J. Multivariate Anal. 110 (2012) 74–90.
[4] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (1974) 716–723.
[5] T. Bedford, R.M. Cooke, Probability density decomposition for conditionally dependent random variables modeled by vines, Ann. Math. Artif. Intell. 32 (2001) 245–268.
[6] T. Bedford, R.M. Cooke, Vines: a new graphical model for dependent random variables, Ann. Statist. 30 (2002) 1031–1068.
[7] M. Bogdan, J.K. Ghosh, R.W. Doerge, Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci, Genetics 167 (2004) 989–999.
[8] E.C. Brechmann, C. Czado, K. Aas, Truncated regular vines in high dimensions with application to financial data, Canad. J. Statist. 40 (2012) 68–85.
[9] E.C. Brechmann, H. Joe, Truncation of vine copulas using fit indices, J. Multivariate Anal. 138 (2015) 19–33.
[10] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (2008) 759–771.
[11] P.F. Christoffersen, Evaluating interval forecasts, Internat. Econ. Rev. (1998) 841–862.
[12] G. Claeskens, N.L. Hjort, Model Selection and Model Averaging, Cambridge University Press, Cambridge, MA, 2008.
[13] C. Czado, U. Schepsmeier, A. Min, Maximum likelihood estimation of mixed c-vines with application to exchange rates, Stat. Modell. 12 (2012) 229–255.
[14] J. Dißmann, E.C. Brechmann, C. Czado, D. Kurowicka, Selecting and estimating regular vine copulae and application to financial returns, Comput. Statist. Data Anal. 59 (2013) 52–69.
[15] C. Donnelly, P. Embrechts, The devil is in the tails: actuarial mathematics and the subprime mortgage crisis, Astin Bull. 40 (2010) 1–33.
[16] J. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, Ann. Statist. 32 (2004) 928–961.
[17] Y. Fan, C.Y. Tang, Tuning parameter selection in high dimensional penalized likelihood, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (2013) 531–552.
[18] C. Francq, J.-M. Zakoian, GARCH Models: Structure, Statistical Inference and Financial Applications, Wiley, New York, 2011.
[19] C. Genest, J. Nevslehová, N. Ben Ghorbal, Estimators based on Kendall's tau in multivariate copula models, Austr. N. Z. J. Statist. 53 (2011) 157–177.
[20] I. Gijbels, M. Omelka, N. Veraverbeke, Partial and average copulas and association measures, Electron. J. Statist. 9 (2015) 2420–2474.
[21] I. Hobæk Haff, Parameter estimation for pair-copula constructions, Bernoulli 19 (2013) 462–491.
[22] H. Joe, Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters, in: L. Rüschendorf, B. Schweizer, M.D. Taylor (Eds.), Distributions with Fixed Marginals and Related Topics, Institute of Mathematical Statistics, Hayward, CA, 1996, pp. 120–141.
[23] H. Joe, Dependence Modeling with Copulas, Chapman & Hall/CRC, Boca Raton, FL, 2014.
[24] D. Kraus, C. Czado, Growing simplified vine copula trees: Improving Dißmanns algorithm, arXiv preprint arXiv:1703.05203 (2017).
[25] D. Kurowicka, Optimal truncation of vines, in: D. Kurowicka, H. Joe (Eds.), Dependence Modeling: Vine Copula Handbook, World Scientific Publishing, Hackensack, NJ, 2011.
[26] M.S. Kurz, F. Spanhel, Testing the simplifying assumption in high-dimensional vine copulas, arXiv:1706.02338, 2017.
[27] D. Müller, C. Czado, Selection of sparse vine copulas in high dimensions with the Lasso, Statist. Comput. 29 (2019) 269–287.
[28] T. Nagler, C. Czado, Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas, J. Multivariate Anal. 151 (2016) 69–89.
[29] T. Nagler, T. Vatter, `rvinecopulib`: High Performance Algorithms for Vine Copula Modeling, 2018. R package version 0.3.1.1.0.
[30] R.B. Nelsen, An Introduction to Copulas, second ed., Springer, New York, 2006.
[31] F. Salmon, Recipe for disaster: the formula that killed wall street, in: Wired.com, 2009.
[32] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[33] A. Sklar, Fonctions de ré répartition á n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8 (1959) 229–231.
[34] F. Spanhel, M.S. Kurz, The partial vine copula: A dependence measure and approximation based on the simplifying assumption, arXiv:1510.06971, 2017.
[35] F. Spanhel, M.S. Kurz, The partial copula: properties and associated dependence measures, Statist. Probab. Lett. 119 (2016) 76–83.
[36] J. Stöber, H. Joe, C. Czado, Simplified pair copula constructions: limitations and extensions, J. Multivariate Anal. 119 (2013) 101–118.
[37] T. Vatter, T. Nagler, Generalized additive models for pair-copula constructions, J. Comput. Graph. Statist. 27 (2018) 715–727.
[38] H. Wang, B. Li, C. Leng, Shrinkage tuning parameter selection with a diverging number of parameters, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 71 (2009) 671–683.
[39] T. Wang, L. Zhu, Consistent tuning parameter selection in high dimensional sparse linear regression, J. Multivariate Anal. 102 (2011) 1141–1151.
[40] D. Wuertz, with contributions from T. Setz, Y. Chalabi, C. Boudt, P. Chausse, M. Miklovac, `fGarch`: Rmetrics: Autoregressive Conditional Heteroskedastic Modelling, 2016. R package version 3010.82.1.
[41] L. Xue, H. Zou, Optimal estimation of sparse correlation matrices of semiparametric Gaussian copulas, Stat. Interface 7 (2014) 201–209.
[42] M. Zak-Szatkowska, M. Bogdan, Modified versions of the bayesian information criterion for sparse generalized linear models, Comput. Statist. Data Anal. 55 (2011) 2908–2924.