Copula model selection using image recognition[*]

Andreas Tsanakas[1] and Rui Zhu[†1]

[1]Bayes Business School (formerly Cass), City, University of London

October 27, 2021

Abstract

The choice of a copula model from limited data is a hard but important task. Motivated by the visual patterns that different copula models produce in smoothed density heatmaps, we consider copula model selection as an image recognition problem. We extract image features from heatmaps using the pre-trained AlexNet, and present workflows for model selection that combine image features with statistical information. We employ dimension reduction via Principal Component and Linear Discriminant Analyses, and use a Support Vector Machine classifier. Simulation studies show that the use of image data improves the accuracy of the copula model selection task, particularly in scenarios where sample sizes and correlations are low. This finding indicates that transfer learning can support statistical procedures of model selection.

Keywords: copula, dependence modelling, image recognition, model selection, classification, transfer learning

1 Introduction

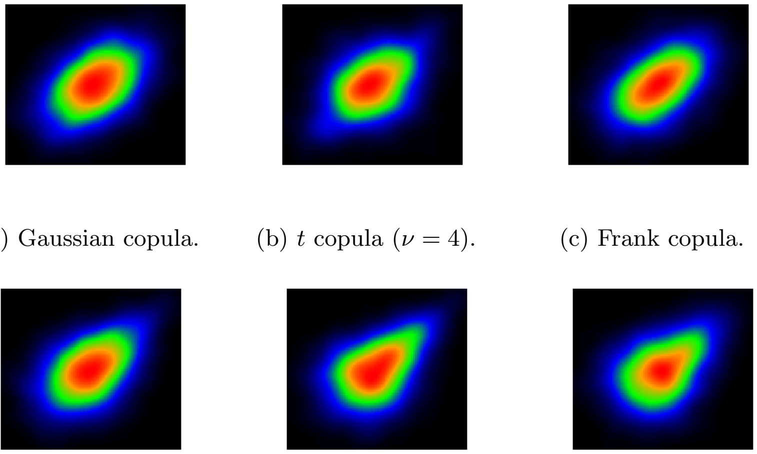
Copulas are dependence modelling tools of fundamental importance in actuarial and financial risk management (Frees and Valdez, 1998, Denuit et al., 2006, McNeil et al., 2015), and fields beyond (e.g. Genest and Favre, 2007). An extensive literature has emerged on specifically actuarial applications of copula modelling, in credibility (Frees and Wang, 2005), stochastic reserving (Shi and Frees, 2011, Shi, 2014, Abdallah et al., 2015), and claims modelling (Czado et al., 2012, Shi and Valdez, 2014, Hu et al., 2021, Tzougas and Pignatelli di Cerchiara, 2021).

[*]The authors are grateful to Marius Hofert, Jinghao Xue, George Tzougas, and Tim Verdonck for helpful comments.

[†]Corresponding author. Email: rui.zhu@city.ac.uk

At the same time, the problem of choosing a copula model based on datasets that are of limited size is a non-trivial task, as evidenced in various strands of the literature, indicatively including: seminal work on copula goodness-of-fit (Genest et al., 2009); the study of practical problems arising in insurance risk management (Shaw et al., 2010); the impact of copula choice on portfolio risk (McNeil et al., 2015, Sec. 11.1.5); and the consideration of dependence uncertainty in a regulatory framework (Embrechts et al., 2014).

The different properties of alternative copula models are often visualised by joint density contour plots or heatmaps. For example, in Figure 1 we show heatmap images of smoothed bivariate densities for six well-known copula models, with standard Normal margins. It is obvious that different copula models have heatmaps with different patterns, reflecting for example different degrees of skewness. This observation motivates our research question: *Do images of smoothed joint densities convey useful information that can improve the accuracy of copula model selection procedures?*



(a) Gaussian copula. (b) t copula ($\nu = 4$). (c) Frank copula.

(d) Gumbel copula. (e) Joe copula. (f) Pareto copula.

Figure 1: Examples of heatmap images of smoothed bivariate densities for different copula models ($n = 2000, \tau = 0.3$).

Our paper seeks to address this question, in the context of small data sizes and bivariate copula models. Small data sizes, e.g. from $n = 100$ to 250, make the copula model selection problem hard, hence an improvement offered by image data would be welcome. Furthermore, when sample size is small, it is natural to focus on the simplest models, given the likely lack of statistical power to detect more complex model features. We note that bivariate copulas can be used to hierarchically build complex multivariate dependence structures (Aas et al., 2009); for a selection tool projecting multivariate copulas to two dimensions see Michiels and De Schepper

(2013). In the case where datasets are richer, the problem the modeller faces is not so much one of selecting between different models, but more one of designing a model that is flexible enough to reflect idiosyncratic features of the data, see e.g. Hofert et al. (2021).

Here, we treat bivariate copula density heatmaps as RGB images and exploit the spatial patterns present in the images to aid bivariate copula selection. The copula selection task is treated as an image recognition or classification task: we classify a given copula sample to an element of a model set, based on its density heatmap image.

One vital challenge in image recognition is to obtain good representations of the images that can summarise well their distinct spatial patterns and thus make the recognition task easier; this is known as representation learning. Deep neural models have been demonstrated to be effective for representation learning, especially in the machine vision community (Bengio et al., 2013). We utilise a deep convolutional neural network, the AlexNet pre-trained by the ImageNet dataset (Krizhevsky et al., 2012), to extract image features with strong representation abilities. This is an example of *transfer learning*, that is, the use of knowledge from addressing a particular problem, to a new task (Pan and Yang, 2009, Zhuang et al., 2020).

Instead of using the extracted image features to train a classifier directly, we propose three additional amendments on them. First, the AlexNet image features are high-dimensional. To avoid potential problems induced by high dimensionality, principal component analysis (PCA) (Wold et al., 1987) is applied to reduce the dimensions of the extracted image features. Second, to further enhance the representation ability of the features, summary statistics are concatenated with image features to provide a more complete description of copula samples. Lastly, we aim to make these features more discriminative via linear discriminant analysis (LDA), which projects the concatenated features to a low-dimensional subspace where the observations from the same classes are grouped close together, while those from different classes are pushed apart (Yang and Jin, 2006). Hence, the recognition task becomes easier on this discriminative subspace. The features extracted by LDA are used as the final representations of the copula samples to train the classifier. Support vector machine (SVM) is chosen as the classifier for the image recognition task, because it is proven to be effective on various real-world applications (Tzotsos and Argialas, 2008, Islam et al., 2017, Sheykhmousa et al., 2020).

We test the performance of the proposed image recognition approach to copula selection via simulation studies. We consider the six copula models of Figure 1 and evaluate the classification accuracy of our approach in different scenarios, comparing to the statistical benchmarks of AIC and BIC. First, we consider model selection when all training and testing instances arise for copula pseudo-samples with the same sample size and underlying rank correlation. While this is not a realistic setting, it allows us to explore the performance of image recognition for different problem parameters. We observe that image recognition consistently outperforms AIC/BIC, except when the underlying rank correlation is very high. The biggest improvement occurs in those scenarios of low sample size and correlation, where the copula selection problem is the hardest. Robustness checks in relation to the choice of marginal for generating images and the dimensions of PC/LDA spaces illustrate sensitivity to those assumptions, but also that the a priori choices we made are reasonable.

Subsequently, we consider a more realistic scenario, where a copula model needs to be selected for data with differing sample sizes and correlations, generated under any rotation of the six bivariate baseline copula models. In this scenario, the heatmap images from the same copula model can present very different patterns, leading to substantial within-class variations. For that reason, we propose to add a first data rotation step, based on sample statistics, with the aim of converting the data to be positively correlated and skewed. Then, in a second step, we apply the image recognition approach to images generated from the rotated data.

We find that this two-step image recognition approach dominates AIC for the copula model selection task, again except in the situation of high correlations. This motivates our final proposal to combine the two-step image recognition approach with AIC. In this combined approach, AIC values (calculated on the rotated data) are concatenated with the image features and statistical features before applying LDA. Experiments show that the combined approach improves on both AIC and image recognition-based model selection. We conclude the analysis with a sensitivity analysis of model predictions, by adapting the scenario weighting method proposed by Pesenti et al. (2019) for simulation models, to our case of a predictive model. The analysis demonstrates how sample skewness is important for distinguishing symmetric from asymmetric models, with image features providing additional information that allows a more granular classification to individual models.

This paper is organised as follows. In Section 2, we give preliminaries on copula modelling. Section 3 introduces the image recognition approach, with fixed correlation and sample sizes. In Section 4 we discuss the two-step approach to copula model selection for more general datasets. Experimental results are summarised within each section. Section 5 presents our concluding remarks.

2 Copulas

2.1 Copula models and their properties

Consider continuous random variables X, Y with marginal distributions F, G and joint distribution H, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *copula* of (X, Y) is a distribution on $[0, 1]^2$

with uniform marginals, such that

$$H(x,y) = C(F(x), G(y)). \tag{1}$$

Denote $U = F(X),\ V = G(Y)$ and also $\bar{U} = 1 - U,\ \bar{V} = 1 - V$. Then it follows from (1) that C is the joint distribution of (U, V), that is,

$$C(u,v) = \mathbb{P}(U \leqslant u, V \leqslant v), \quad (u,v) \in [0,1]^2. \tag{2}$$

Analogously, the joint distribution of $(\bar{U}, \bar{V})$ is called the *survival copula* of (X, Y) and denoted by $\bar{C}$.

Definition (1) implies a separation of a random vector's marginal behaviour from its dependence structure, which has enabled copulas to be widely employed as multivariate modelling tools. For detailed treatments of copulas, including applications in insurance and financial risk management, see Nelsen (2007), Denuit et al. (2006), McNeil et al. (2015). We note that the copulas of discontinuous random vectors are not uniquely defined – in such a case the variables $U,\ V$ as constructed above are not uniform. However, we can always uniquely determine a copula for X, Y via (2), with uniform variables U, V constructed via the generalised distributional transform of Rüschendorf and de Valk (1993).

In risk management, the specific properties of different copula families are important. Assume that X and Y represent losses, such that high (joint) outcomes are associated with adverse events. Then, beyond considering (rank) correlation measures, it is important to model the extent to which X and Y will jointly achieve high values. A typical way in which the literature considers the propensity of joint extremes is via the coefficients of upper and lower tail dependence (e.g. McNeil et al., 2015, Sec. 7.2.4):

$$\begin{aligned}
\lambda_U(U,V) &:= \lim_{p \to 1} \mathbb{P}\left(Y > G^{-1}(p) \mid X > F^{-1}(p)\right) = \lim_{p \to 1} \frac{\bar{C}(1-p, 1-p)}{1-p}, \\
\lambda_L(U,V) &:= \lim_{p \to 0} \mathbb{P}\left(Y \leqslant G^{-1}(p) \mid X \leqslant F^{-1}(p)\right) = \lim_{p \to 0} \frac{C(p,p)}{p}.
\end{aligned}$$

Models for which λ_U or λ_L are non-zero are, respectively, called upper or lower tail dependent.

While tail dependence is an asymptotic property, a distinct issue is the skewness or asymmetry of a copula. A copula is symmetric if $C(1-u, 1-v) = C(u,v)$. Various measures of bivariate skewness have been proposed by Rosco and Joe (2013). Here we focus on the moment-based measure ζ, defined for $k \in (1, \infty)$ as:

$$\zeta(U,V;k) = \mathbb{E}\left[|U+V-1|^k \operatorname{sign}(U+V-1)\right].$$

Implications of the choice of the parameter k are briefly explored in Rosco and Joe (2013).

A further property of bivariate copulas relates to the extent that observations are concentrated in the four corners of the unit box, even when correlation is low, leading to spider-like pattern. This property, which distinguishes, e.g., a t from a Gaussian copula model, is termed *arachnitude* in Shaw et al. (2010), see also Androschuck et al. (2017), Genest et al. (2019). We measure arachnitude in the way proposed by Shaw et al. (2010), that is, as

$$\xi(U,V) := \rho\left((U-0.5)^2,(V-0.5)^2\right),$$

where ρ is the Pearson (product-moment) correlation.

In this paper we consider six bivariate copula models:

1. The *Gaussian copula* is probably the most widely used copula model. It is symmetric and tail independent.

2. The *t copula* is symmetric but both upper- and lower-tail dependent. It admits a degree of freedom parameter ν; λ_L, λ_U decrease in ν, while for $\nu \to \infty$, the t copula reduces to a Gaussian.

3. The *Frank copula* is symmetric and tail independent.

4. The *Gumbel copula*, commonly used in risk management, is positively skewed and upper tail dependent.

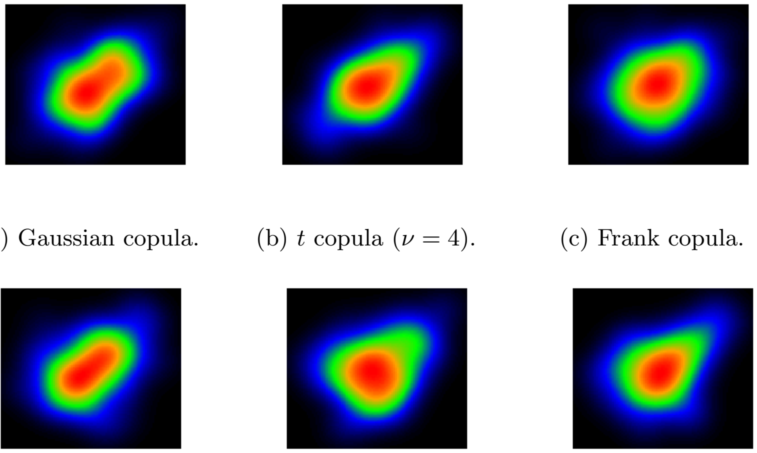
5. The *Joe copula* is positively skewed and upper tail dependent.

6. The *Pareto copula* (or Clayton survival copula), is positively skewed and upper tail dependent.

We do not provide technical detail on these models, as they are all well known and exhaustively discussed in the literature (Nelsen, 2007, Denuit et al., 2006, McNeil et al., 2015). The Gaussian and t models are, respectively, the copulas of bivariate Normal and t distributions; the remaining four models belong to the family of Archimedean copulas. All models, except the t copula, have a single parameter, which can be calibrated to (e.g. Kendall's) rank correlation or estimated by MLE. The t model has the degrees of freedom ν as an additional parameter.

The properties of different copula families are illustrated in Figure 1, which shows heatmaps of smoothed bivariate densities, each derived from samples of size $n = 2000$, with underlying Kendall rank correlation $\tau = 0.3$. The distinct patters of the different models are visible. At the same time there is substantial similarity between some of the resulting heatmaps (e.g. Gaussian and t; Joe and Pareto), which indicates that selecting the correct model from data is not a trivial task. This is of course even more challenging for smaller datasets. We show heatmaps from the same six copula families in Figure 2, but this time generated from bivariate samples of size

$n = 100$. While the general patterns observed in Figure 1 are preserved, they are substantially noisier.



(a) Gaussian copula.

(b) t copula ($\nu = 4$).

(c) Frank copula.

(d) Gumbel copula.

(e) Joe copula.

(f) Pareto copula.

Figure 2: Examples of heatmap images of smoothed bivariate densities for different copula models ($n = 100, \tau = 0.3$).

2.2 Estimation and model selection

Consider a sample from (X, Y), $(x_1, y_1), \ldots, (x_n, y_n)$. Realisations of the random vector (U, V) are not directly observable. As a result it is common to construct copula *pseudo-observations* (e.g. Genest et al., 2009). Let $r_i(\mathbf{z})$ be the rank of observation z_i in a univariate sample $\mathbf{z} = (z_1, \ldots, z_n)$. Then, the pseudo-observations are given by

$$u_i = \frac{r_i(\mathbf{x})}{n+1}, \quad v_i = \frac{r_i(\mathbf{y})}{n+1}.$$

From the pseudo-observations (u_i, v_i), $i = 1, \ldots, n$, we can readily estimate skewness and arachnitude, denoting the corresponding estimates by $\hat{\zeta}$, $\hat{\xi}$. Furthermore, we denote the sample version of Kendall's rank correlation as $\hat{\tau}$.

For a parametric family of copulas $C^{(m)}(\cdot;\theta)$, $\theta \in \Theta_m$, the parameters θ can be estimated by maximum likelihood estimation, treating the pseudo-observations as if they are a random sample from $C^{(m)}(\cdot;\theta)$. If we are considering a family of copula models $\{C^{(m)},\ m \in \mathcal{M}\}$, likelihood methods also offer model selection criteria. Let $c^{(m)}$ be the bivariate density corresponding to copula $C^{(m)}$ and $\hat{\theta}^{(m)}$ the (k_m-dimensional) estimate of the corresponding model parameter.

The Akaike and Bayes Information criteria are given by:

$$\mathrm{AIC}^{(m)} = 2k_m - 2\sum_{i=1}^{n} \log c^{(m)}\left(u_i, v_i; \hat{\theta}^{(m)}\right),$$

$$\mathrm{BIC}^{(m)} = k_m \log n - 2\sum_{i=1}^{n} \log c^{(m)}\left(u_i, v_i; \hat{\theta}^{(m)}\right).$$

The selected model is then the one with the lowest $\mathrm{AIC}^{(m)}$ or $\mathrm{BIC}^{(m)}$. A cross-validated log-likelihood criterion is formulated by Grønneberg and Hjort (2014, eq. (42)); see Jordanger and Tjøstheim (2014) for a simulation study. Other model selection criteria can be constructed using goodness-of-fit statistics; for example Kularatne et al. (2021) employ the Cramer-von Mises statistic, the use of which (including variations) in copula goodness-of-fit testing has been thoroughly explored by Genest et al. (2009). Bayesian copula selection is discussed in Huard et al. (2006).

In this paper, we use as statistical benchmarks for model selection AIC and BIC.[1]

3 An image recognition-based approach to copula model selection

In this section we introduce a new methodology, which uses image recognition to select a suitable bivariate copula from a pseudo-sample, by classifying its density heatmap image to one of the six models we consider. The recognition process is designed to incorporate rich information that can well represent the samples and is discriminative to make the classification task easier. The generation of the heatmap images is introduced first, and then the image recognition approach is discussed in detail. Subsequently, experimental results are shown to demonstrate the effectiveness of this approach for copula model selection.

We note that in the present section we apply the classification / copula model selection framework to simulated data with very benign features, with all images in any given dataset derived from pseudo-samples with the same sample size and underlying correlation. This is of course an unrealistic testing environment, with classes that are more homogeneous than in any practical application. Nonetheless, the setting of this section allows us to evaluate whether image recognition can be effective as a copula selection tool (and under which conditions). The restrictive assumptions of this section are relaxed in the two-step approach of Section 4.

[1] In early experiments we have found that these outperform both the Cramer-von Mises statistic and the cross-validated log-likelihood of Grønneberg and Hjort (2014) with 10-fold cross-validation. Hence we do not report on those statistics in the paper.

3.1 Generating heatmap images of smoothed bivariate densities

Here we outline how the image datasets are generated, on which classifiers are trained to perform the copula selection task. Each image is a smoothed bivariate density heatmap, generated from a simulated pseudo-sample from (U, V), drawn from a given copula specification $C^{(m)}(\cdot;\theta),\ \theta \in \Theta^{(m)},\ m \in \mathcal{M}$.

For each image dataset that we generate the following hold:

a) The dataset contains $R = 20,000$ images.

b) Each image is derived from a bivariate sample of size n, drawn from one of the 6 copula families we consider in this paper, $\mathcal{M} = \mathcal{M}_s \cup \mathcal{M}_a$, where $\mathcal{M}_s = \{\text{Gaussian, } t\text{, Frank}\}$ and $\mathcal{M}_a = \{\text{Gumbel, Joe, Pareto}\}$ contain symmetric and asymmetric (positively skewed) models respectively. Each dataset contains approximately the same number of images from each copula family.

c) In each dataset all images are generated from simulated samples with a fixed sample size $n \in \{100, 150, 200, 250\}$ and (population) Kendall $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Hence we have $4 \times 5 = 20$ datasets, corresponding to different (n, τ) combinations, each containing R images.

d) All images are generated from bivariate samples that have positive empirical rank correlation and positive empirical skewness. This means that samples that display negative rank correlation are rejected and not used to produce images in the dataset. Furthermore, samples that display negative sample skewness, when the underlying copula model has positive skewness, are also rejected.[2]

e) Images are generated based on pseudo-observations from the copula samples, since the latter would not be available in practice. Furthermore, to generate the heatmap images, we transform pseudo-observations to have a standard normal marginal distribution. This transformation is employed only for image generating purposes and reflects no assumption of normality for the marginal distribution of the underlying data.

The precise process by which images are generated is given in Algorithm 1, where we suppress the subscript i for quantities estimated from the ith simulated sample. All calculations are carried out in **R**. For random number generation we use package *copula* (Jun Yan, 2007, Hofert et al., 2020). For AIC/BIC calculations we use the package *VineCopula* (Schepsmeier et al., 2021). Joint densities are estimated on a 100×100 grid on $[-3, 3]^2$ using an axis-aligned

[2] This process takes away from realism, by removing some noise from the generated datasets. Nonetheless, we pursue this strategy for image generation in order to maintain consistency with the two-step approaches of Section 4.

bivariate normal kernel, by the function `kde2d` of the package *MASS*, with bandwidths set to 1.3 times the values given by the heuristic in Venables and Ripley (2002, eq. (5.5)). The plots are produced by the `image` function and saved as png files.

3.2 The image recognition approach

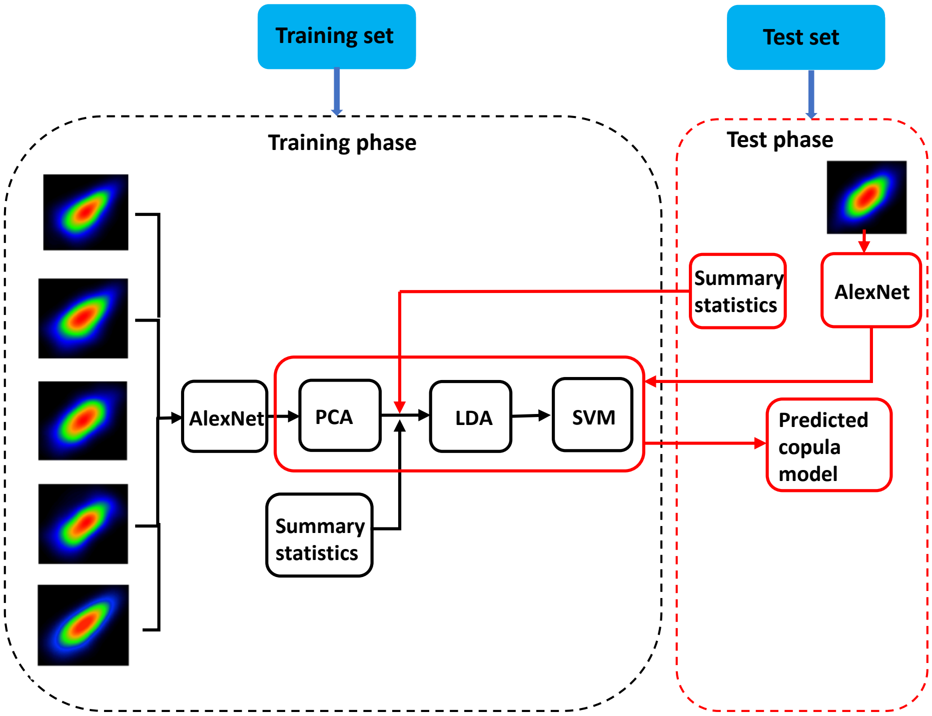


Figure 3: The workflow of the image recognition approach for copula model selection.

Now we present the image recognition approach with the complete workflow shown in Figure 3. Similarly to all classification tasks, the image recognition approach consists of a training phase to extract features and train a classifier, and a test phase to predict a copula model for a test sample. In Figure 3, the training phase is presented by the black flow while the test phase is presented by the red flow.

3.2.1 The training phase

The training phase contains two parts: 1) the representation learning part to extract features with strong representation and discrimination abilities and 2) the classification part to train an effective classifier.

Algorithm 1: Algorithm for generating the image dataset, with given fixed n, τ.

```
$i \leftarrow 0$;
while $i < R$ do
    Choose randomly copula family $C^{(m)},\ m \in \mathcal{M}$;
    if $m$ is the $t$ copula then
        Choose randomly degrees of freedom $\nu$ from $\{3, 4, \ldots, 10\}$;
        Work out parameters $\theta$ from $(\tau, \nu)$;
    else
        Work out parameter $\theta$ from $\tau$;
    end
    Simulate $n$ pairs of observations from $(U, V) \sim C^{(m)}(\cdot; \theta)$;
    Transform simulated observations to pseudo-observations $(u_j, v_j)_{j=1,\ldots,n}$;
    From $(u_j, v_j)_{j=1,\ldots,n}$ calculate sample statistics $\hat{\tau}$, $\hat{\zeta}$, $\hat{\xi}$;
    if $\hat{\tau} \geqslant 0$ then
        if $\hat{\zeta} \geqslant 0$ then
            $KeepData \leftarrow$ TRUE
        else
            if $m \in \mathcal{M}_s$ then
                Rotate pseudo-observations, $u_j \leftarrow 1 - u_j,\ v_j \leftarrow 1 - v_j,\ j = 1, \ldots, n$;
                $\hat{\zeta} \leftarrow -\hat{\zeta}$;
                $KeepData \leftarrow$ TRUE;
            else
                $KeepData \leftarrow$ FALSE;
            end
        end
    else
        $KeepData \leftarrow$ FALSE
    end
    if $KeepData =$TRUE then
        $i \leftarrow i + 1$;
        Calculate $\mathrm{AIC}^{(l)}$ and $\mathrm{BIC}^{(l)}$ from $(u_j, v_j)_{j=1,\ldots,n}$, for each $l \in \mathcal{M}$ ;
        Save $\hat{\tau}$, $\hat{\zeta}$, $\hat{\xi}$, $\mathrm{AIC}^{(l)}$, $\mathrm{BIC}^{(l)}$;
        Transform pseudo-observations to normal $x_j \leftarrow \Phi^{-1}(u_j),\ y_j \leftarrow \Phi^{-1}(v_j),\ j = 1, \ldots, n$;
        Estimate joint density of $(x_j, y_j)$. Create heatmap and save image;
    end
end
```

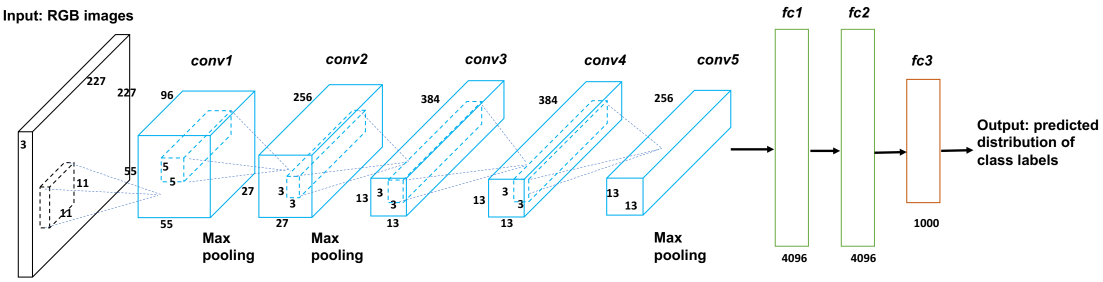


Figure 4: The architecture of the AlexNet.

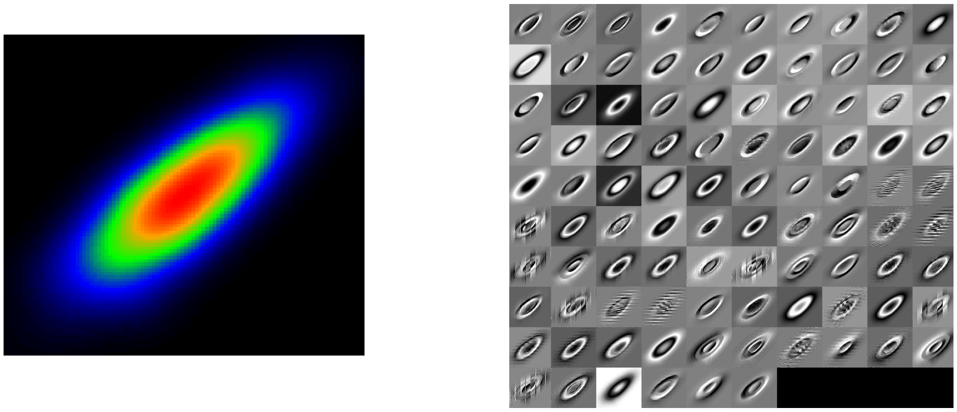


Figure 5: Left: The heatmap image of a Gaussian copula sample. Right: The activations of the first convolutional layer of AlexNet for the sample.

Representation learning In the representation learning part, two types of features are extracted from the training copula samples: pure image features and statistical features. The pure image features are extracted from the training heatmap images by the pretrained AlexNet, a deep convolutional neural network that has been trained on the ImageNet dataset with 1000 classes of high-resolution images (Deng et al., 2009). Given the complex nature of the images in the ImageNet dataset and the competitive classification performance of AlexNet, we believe that this pretrained network can provide good representations of our heatmap images of relatively simple patterns.

The architecture of AlexNet is depicted in Figure 4, with an input layer of RGB images, five convolutional layers and three fully connected layers. The output from the last fully connected layer is the predicted distribution of the class labels, i.e. the predicted probabilities for each of the 1000 classes considered. Note that we do not use the classification output of AlexNet, as the 1000 classes used are not relevant to our copula selection task; however the representation ability of the network allows us to use it for feature extraction from our heatmap images. To demonstrate the good representation ability of the pretrained AlexNet, we show the activations of a Gaussian copula sample for all 96 channels of the first convolutional layer in Figure 5, with each square presenting the activations of one channel. The bright pixels reflect high activations, which means that they make substantial contributions to the extracted features. It is obvious that the contour shapes of the Gaussian sample can be well captured by the first convolutional layer.

We input the training heatmap images to the pretrained AlexNet, and extract 4096 features from the second fully connected layer, 'fc2' in Figure 4, which is the last feature extraction layer before working out predicted probabilities. This layer provides high-level abstract features that can well represent the images. To be precise, each training heatmap image is represented by a 4096-dimensional vector $\mathbf{x}_i^M \in \mathbb{R}^{4096\times 1}$ $(i = 1, 2, \ldots, N)$, where N is the number of training copula samples, and the image features of the whole training set is denoted as $\mathbf{X}^M = [\mathbf{x}_1^M, \mathbf{x}_2^M, \ldots, \mathbf{x}_N^M]^T \in \mathbb{R}^{N\times 4096}$. (Note that N here is different to the value of R in Algorithm 1, as the generated images for each dataset will be subsequently split into training and test samples.)

The high number of image features can lead to potential problems in classification, e.g. the curse of dimensionality and high computational cost. Thus, we reduce the number of image features before training a classifier. This dimension reduction step is achieved by principal component analysis (PCA), which is a widely adopted unsupervised dimension reduction method that can provide low-dimensional yet effective representations of the original high-dimensional data (Wold et al., 1987). PCA projects data from the original feature space to a low-dimensional subspace, spanned by the first few principal components (PCs) that can explain most of the variation in data. This can be achieved by applying the reduced singular value decomposition (SVD) on the column-centred $\mathbf{X}^M$:

$$(\mathbf{X}^M)^c = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{3}$$

where $(\mathbf{X}^M)^c \in \mathbb{R}^{N\times 4096}$ is the column-centred $\mathbf{X}^M$ derived by extracting column means from $\mathbf{X}^M$, $\mathbf{U} \in \mathbb{R}^{N\times r}$ and $\mathbf{V} \in \mathbb{R}^{4096\times r}$ contain left and right singular vectors, $\mathbf{D} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with singular values $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_r \geqslant 0$. The first q ($q \leqslant r$) columns in $\mathbf{V}$, i.e. the first few PCs, are selected to construct the PC subspace. In this paper, we set $q = 150$ to explain 99.9% of the variation in $\mathbf{X}^M$. After PCA, the image features of the training set become

$$\mathbf{X}^{MP} = (\mathbf{X}^M)^c\mathbf{V}_{150} \in \mathbb{R}^{N\times 150}, \tag{4}$$

where $\mathbf{V}_{150} \in \mathbb{R}^{4096\times 150}$ is $\mathbf{V}$ with the first 150 columns. Thus the image features now lie in a 150-dimensional PC subspace.

Besides the pure image features extracted from AlexNet, we propose to utilise some additional statistical features to enrich the description of the training copula samples. Three summary statistics, Kendall's rank correlation, skewness and arachnitude are chosen as the statistical features. The statistical features of each sample are denoted as $\mathbf{x}_i^S = (\hat{\tau}_i, \hat{\zeta}_i, \hat{\xi}_i) \in \mathbb{R}^{3\times 1}, i = 1, 2, \ldots, N$.

The low-dimensional image features and three statistical features are then concatenated to provide a representation for the ith copula sample, $\mathbf{x}_i = [(\mathbf{x}_i^{MP})^T, (\mathbf{x}_i^S)^T]^T \in \mathbb{R}^{153\times 1}$, where $\mathbf{x}_i^{MP} \in \mathbb{R}^{150\times 1}$ is the ith observation in $\mathbf{X}^{MP}$. The feature matrix to represent all training copula samples is denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N\times 153}$. Before simply feeding this feature matrix to a classifier, we extract more compact and discriminative information to reflect the differences between classes better and make the classification process easier. For that purpose, we apply linear discriminant analysis (LDA) on $\mathbf{X}$. LDA is a well known supervised dimension reduction method that projects data to a subspace such that between-class variation is maximised while within-class variation is minimised (Yang and Jin, 2006). The classification task is easier on this subspace because the instances from the same class are pulled close together while those from different classes are pushed further away. LDA finds such discriminative subspace by solving the following optimisation problem:

$$\max_{\mathbf{W}} \frac{\det(\mathbf{W}^T\mathbf{S}_B\mathbf{W})}{\det(\mathbf{W}^T\mathbf{S}_W\mathbf{W})}, \tag{5}$$

with $\mathbf{S}_W = \sum\limits_{k=1}^{K} \sum\limits_{i \text{ in class } k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$ and $\mathbf{S}_B = \sum\limits_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T$. Here K is the number of classes, $\mathbf{W} \in \mathbb{R}^{153\times(K-1)}$ contains the bases of the linear discriminant subspace,

$\mathbf{x}_i \in \mathbb{R}^{153\times 1}$ denotes each sample in $\mathbf{X}$, $\boldsymbol{\mu}_k \in \mathbb{R}^{153\times 1}$ is the class mean of the k-th class and $\boldsymbol{\mu} \in \mathbb{R}^{153\times 1}$ is the overall mean of $\mathbf{X}$. As the optimisation problem (5) involves class information, $\mathbf{W}$ summarises the discriminative information between classes. By projecting $\mathbf{X}$ on the linear discriminant subspace, we have

$$\mathbf{X}^P = \mathbf{X}\mathbf{W} \in \mathbb{R}^{N\times(K-1)}. \tag{6}$$

LDA can provide at most a $(K-1)$-dimensional subspace. In this paper, six copula models are considered, thus the LDA subspace is at most five-dimensional. Here we take all five discriminative dimensions provided by LDA.

Classification Support vector machine (SVM) is chosen as the classifier for its efficiency in many real-world applications (Tzotsos and Argialas, 2008, Islam et al., 2017, Sheykhmousa et al., 2020). The training set to train SVM contains N pairs of observations $\{\mathbf{x}_i^P, m_i\}_{i=1}^N$, where $\mathbf{x}_i^P \in \mathbb{R}^{5\times 1}$ is the feature vector of the ith observation obtained in the representation learning part and $m_i \in \mathcal{M}$ is the corresponding copula model.

SVM aims to find a separating hyperplane $f(\mathbf{x}) = \phi(\mathbf{x}_i^P)^T\mathbf{w} + b$ for classification by maximising the margin M between two classes:

$$\begin{aligned} &\max_{\mathbf{w},b} \;\; M \\ \text{s.t.} \;\; & y_i(\phi(\mathbf{x}_i^P)^T\mathbf{w} + b) \geqslant M(1-\psi_i) \;\; \forall i, \\ & \psi_i \geqslant 0 \;\; \forall i, \;\; \sum_{i=1}^N \psi_i \leqslant C, \end{aligned} \tag{7}$$

where $\mathbf{w}$ and b defines the separating hyperplane, $\phi(\cdot)$ is a function that projects $\mathbf{x}_i^P$ to a reproducing kernel Hilbert space, ψ_i is the slack variable that allows violations of the training observations to the margins, and C is a predefined positive integer that controls the trade-off between the goodness-of-fit of the training set to the classifier and the generalisation ability of the classifier on unseen data. The solutions $\mathbf{w}^*$ and b^* are then used to classify a test observation $\mathbf{x}$: if $f(\mathbf{x}) = \phi(\mathbf{x})^T\mathbf{w}^* + b^*$ is positive, then $\mathbf{x}$ belongs to the positive class; otherwise, $\mathbf{x}$ belongs to the negative class.

Note that this formulation of SVM can only be used for binary classification. To apply it in our case with six classes, we adopt the one-versus-one strategy (Friedman et al., 2009). We apply SVM to pairs of classes, which means that $\binom{K}{2}$ SVM classifiers are trained. For the test observation $\mathbf{x}$, we obtain $\binom{K}{2}$ classification results. A majority vote is then applied to these classification results to determine the class of $\mathbf{x}$, i.e. the class with the highest vote is selected.

3.2.2 The test phase

In the test phase, given one observed copula sample, we first extract its image features $\mathbf{x}_t^M \in \mathbb{R}^{4096\times 1}$ from the pretrained AlexNet and project them to the PC subspace constructed in the training phase:

$$\mathbf{x}_t^{MP} = \mathbf{V}_{150}^T(\mathbf{x}_t^M)^c \in \mathbb{R}^{150\times 1}, \tag{8}$$

where $(\mathbf{x}_t^M)^c$ is the centred $\mathbf{x}_t^M$ by the column means of $\mathbf{X}^M$. The image features are then concatenated with the summary statistics $\mathbf{x}_t^S \in \mathbb{R}^{3\times 1}$ to form the feature vector of the test copula sample, $\mathbf{x}_t = [(\mathbf{x}_t^M)^T, (\mathbf{x}_t^S)^T]^T \in \mathbb{R}^{153\times 1}$. We then project this vector to the linear discriminant subspace to obtain the discriminative features for classification:

$$\mathbf{x}_t^P = \mathbf{W}^T\mathbf{x}_t \in \mathbb{R}^{5\times 1}. \tag{9}$$

The copula model is then selected by applying the trained SVM on $\mathbf{x}_t^P$, based on the procedure discussed in section 3.2.1.

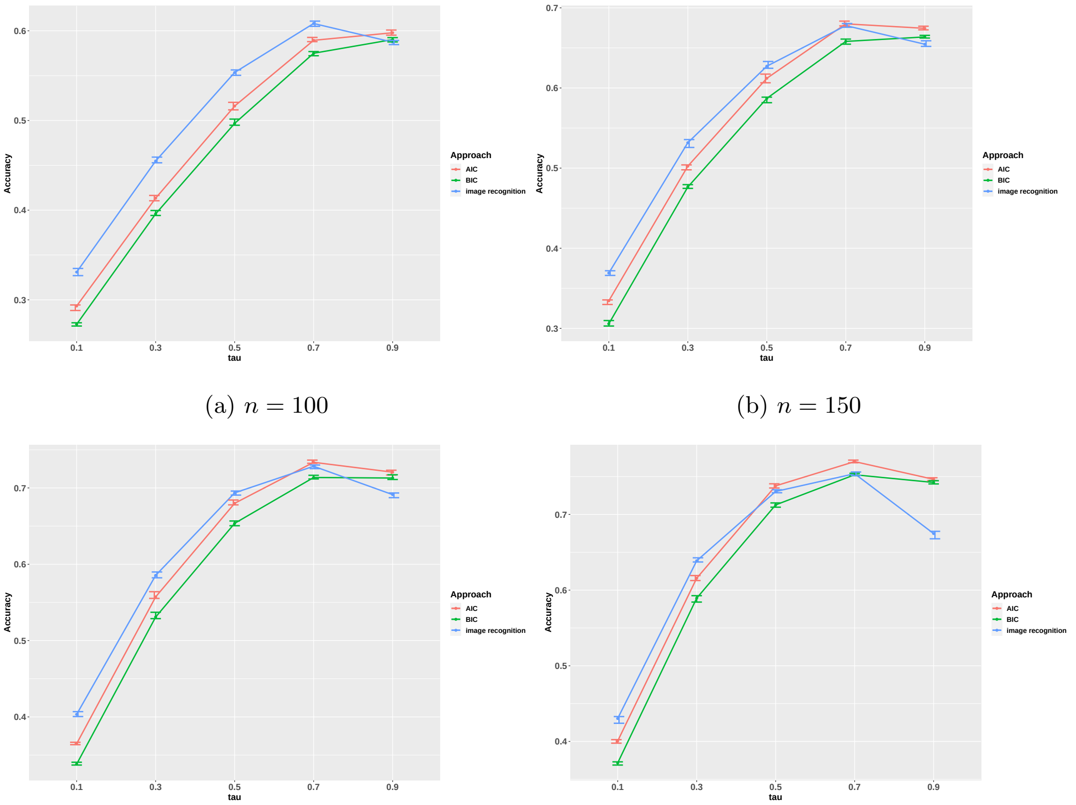
3.3 Classification results on copula samples with fixed n and τ

3.3.1 Experimental settings

For each dataset with fixed n and τ, we randomly select 70% of the $R = 20,000$ images to form the training set, with the remaining images used as a test set; hence the training sample size is $N = 0.7 \times R = 14,000$. For SVM, the radial basis function (RBF) kernel is chosen as the kernel function. The hyperparameters associated with the SVM classifier are tuned by 10-fold cross-validation on the training set. The classification accuracies of the image recognition approach are recorded. Furthermore, we record the accuracy by which the copula model is selected when using AIC or BIC as a criterion. To make the results more reliable, the training/test random split process is repeated 20 times. Thus for each combination of n and τ, we record 20 classification accuracies for each copula model selection method.

3.3.2 Classification results

The classification results are shown in Figure 6, with each plot presenting the accuracies of the three approaches with a fixed value of n and all values of τ. For each plot, the horizontal axis represents the values of τ, while the vertical axis represents the classification accuracy. The mean classification accuracies of the image recognition approach are shown by blue curves, those of AIC are shown by red curves, and those of BIC are shown by green curves. The two short bars associated with each point in Figure 6 are the lower quartile and upper quartile of the 20 recorded accuracies for each method.



(a) $n = 100$

(b) $n = 150$

(c) $n = 200$

(d) $n = 250$

Figure 6: Classification accuracies for fixed n and τ.

It is clear that, generally, the classification accuracy increases as τ increases for each value of n, and also increases as n increases for each value of τ. This pattern makes sense because the underlying model can be better described by samples with larger n's. The models with larger τ's are also easier to classify, as the characteristics of the models are presented clearer when τ is large. The image recognition approach can beat both AIC and BIC when τ is small, i.e. less than 0.7, and this is more obvious when n is small. This is encouraging, as the improvement provided by the image recognition approach to select copula models occurs for samples that are difficult to classify, i.e. those with small n and τ. However, when τ is large, the image recognition approach performs obviously worse than AIC and BIC; moreover, beyond $\tau = 0.7$, the classification accuracies start decreasing. A plausible reason for this is that, when τ is very large, the heatmap images of different copula models exhibit very similar visual patterns and image recognition cannot easily distinguish between them.

We also note the consistently better performance of AIC compared to BIC. The only difference between those two statistics is the higher penalty for additional parameters that BIC assigns. As there is only one model with 2 parameters (the t copula), the BIC seems to systematically mistake the t copula for another model (most commonly the Gaussian).

To sum up, the preliminary experimental results on copula samples with fixed n and τ demonstrate the potential effectiveness of image recognition to select copula models, especially when n and τ are relatively small.

3.4 Robustness tests

Here we summarise the results of two robustness checks, seeking to evaluate the extent to which classification performance is impacted by potentially arbitrary decisions in the design of our copula selection process.

First, in order to generate heatmaps, a choice of marginal distribution is necessary. (As we are only investigating dependence effects, this choice does not reflect any assumption regarding the marginal distributions of the actual data one may be modelling; in a sense, it is a hyperparameter choice). So far, all heatmap images are generated from bivariate samples with Normal margins. Here, we additionally consider Cauchy, Laplace and Uniform margins. For the $R = 20,000$ bivariate copula samples we generated with $n = 250$ and $\tau = 0.5$, we produce heatmaps using each of those additional margins by slightly modifying Algorithm 1. Subsequently, we extract features and train a SVM to classify those heatmaps in the case of each margin, following the same process and experiment settings as in Section 3.3. The results are summarised in Figure 7. Clearly, the Normal and Laplace margins have very similar medians and interquartile ranges, indicating that classification performance is similar for those two marginal choices. The Cauchy and Uniform margins show worse classification accuracies, with medians lower than those of Normal and Laplace margins by around 1.5%. This shows that the choice of margin has a noticeable effect on the final results and that the choice of a Normal margin proved to be a beneficial one.

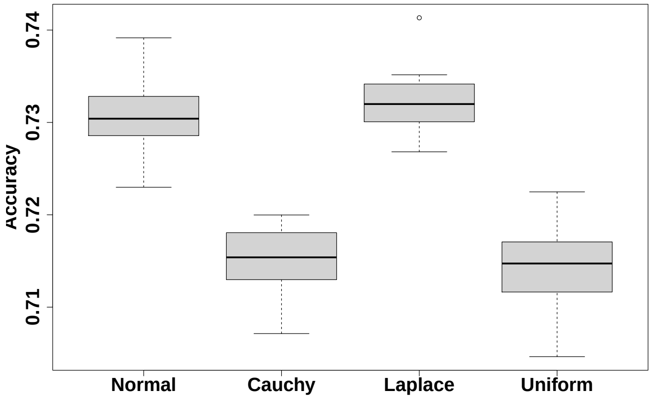


Figure 7: The classification accuracies for different margins with $n = 250$ and $\tau = 0.5$.

Second, the dimensions of the PC and LD subspaces can affect the final classification performance, because they determine the amount of information to be included in the low-dimensional subspaces. Setting the dimensions to small numbers may result in low classification accuracies because of the loss of vital information for classification, while setting them to large numbers close to the original feature dimensions fails to achieve dimension reduction. In Figure 8, we show the surface curve of the classification accuracies for different dimensions of the PC and LD subspaces to classify copula samples with $n = 250$ and $\tau = 0.5$. Five dimensions of the PC subspace are tested, {10,50,70,100,150}, while three dimensions of the LD subspace are tested, {3,4,5}. As expected, when the dimensions of the subspaces are low, e.g. the dimension of the PC subspace is 10 and that of the LD subspace is 3 or 4, the classification accuracies are just around 71%. However, when the dimension of the PC subspace is higher than 50 and that of the LD subspace is set to the maximum number of five, we can observe the highest classification accuracies of more than 73%. These results demonstrate that our choices of 150-dimensional PC subspace and 5-dimensional LD subspace are sensible.

Following the analysis of this section, we will continue using Normal margins, 150 PCs and a 5-dimensional LD subspace in the rest of the paper.

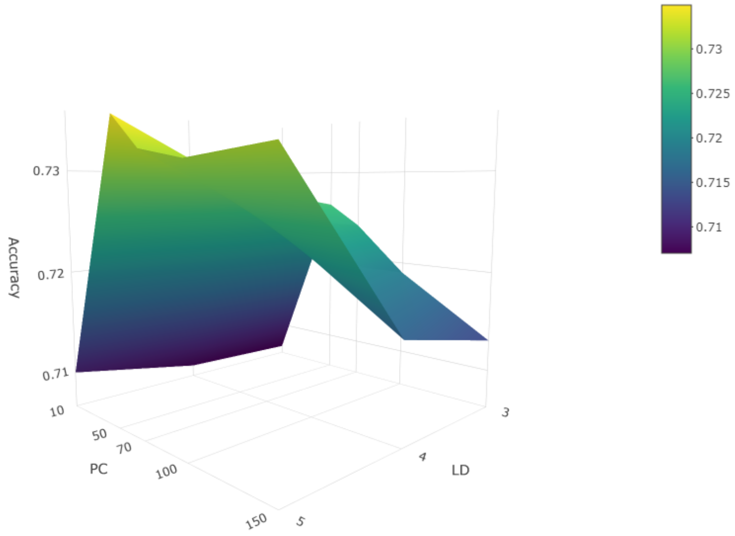


Figure 8: Classification accuracies for different dimensions of the PC and LD subspaces to classify copula samples with $n = 250$ and $\tau = 0.5$.

4 Selecting (rotated) copula models with variable n and τ

Encouraged by the results of Section 3, we here address the more realistic scenario where n and τ are not fixed. Furthermore, we allow τ to be negative. In particular, for asymmetric copula models, we consider all rotations, such that each model in $\mathcal{M}_a$ has now four distinct versions. We believe that this addresses a realistic modelling scenario, as the classification approach does not assume anything a priori about the sign of either the correlation or the skewness of the underlying model.

In this section we consider three distinct approaches:

1. **Copula selection with AIC.** This approach requires us to consider an enlarged set of candidate models, which also includes the rotations of the positively correlated and skewed models in $\mathcal{M}$.

2. **Image recognition with a two-step approach.** In the first step, sample statistics are used to assess the sign of correlation and skewness – essentially trying to detect if elements of $\mathcal{M}$ have been rotated. Then, the pseudo-samples are transformed to have positive correlation and skewness. In the second step, heatmap images are generated from the transformed samples and classified to models in $\mathcal{M}$.

3. **Combining image recognition with AIC.** The same approach as in 2. is followed,

with the AIC of the models in $\mathcal{M}$ added as a feature in the second step of the process. The motivation for this is to not miss out on any information encoded in likelihood-based statistical criteria, which is not present in image features.

The performance of the three copula model selection approaches is assessed on a test set. In contrast to Section 3, for the two-step approaches of this section we cannot validate the result of the classification algorithm on subsets of the training set. In the training set, all copula samples are positively correlated and skewed, to make the within-class variation smaller. However, this is not a realistic testing scenario. Thus, in the test set, we consider copula samples that may have negative or positive correlation and skewness. Before we discuss each of the three approaches in more depth, we describe the construction of this test set.

4.1 Test set

We produce a test set of $S = 10,000$ bivariate copula pseudo-samples. Each pseudo-sample is generated from a copula in $\mathcal{M}$, with randomly (uniformly) chosen $n \in [100, 250]$ and $\tau \in [0.1, 0.9]$. Before simulating each pseudo-sample, the underlying copula model is rotated by 0, 90, 180, or 270 degrees counter-clockwise. Hence, we deal with an enlarged model set, denoted by $\mathcal{M}'$. Let m_r represent a model m in $\mathcal{M}$, rotated by r degrees, such that $m_0 = m$. For symmetric models $m \in \mathcal{M}_s$, we also have $m_{180} = m_0,\ m_{270} = m_{90}$. Then, we let $\mathcal{M}' = \mathcal{M}'_s \cup \mathcal{M}'_a$, where $M'_s = \{m_r :\ m \in \mathcal{M}_s,\ r \in \{0, 90\}\}$ and $M'_a = \left\{m_r :\ m \in \mathcal{M}_a,\ r \in \{0, 90, 180, 270\}\right\}$. The data and copula model specification are saved, as well as the rank correlation of the rotated model τ_r.

The generation of the test set is outlined in Algorithm 2.

4.2 Copula selection approaches

4.2.1 Copula selection with AIC

For each instance i in the test set, the model in $\mathcal{M}'$ is chosen with the smallest AIC, as calculated in the penultimate step of Algorithm 2. The classification is successful if the chosen model is identical to the underlying model $m_r \in \mathcal{M}'$.

4.2.2 Image recognition with a two-step approach

When we allow for models with negative correlation and skewness, the classification task becomes harder. One can either assign a different class for each element in the enlarged model space $\mathcal{M}'$ – thus moving from 6 to 18 classes – or within each of the 6 classes in $\mathcal{M}$ accommodate

Algorithm 2: Algorithm for generating the test set, with variable n, τ and model rotations.

```
$i \leftarrow 0$;
while $i < S$ do
    Choose randomly $m \in \mathcal{M}$, $n \in [100, 250]$, $\tau \in [0.1, 0.9]$;
    if $m$ is the $t$ copula then
        Choose randomly degrees of freedom $\nu$ from $\{3, 4, \ldots, 10\}$;
        Work out parameters $\theta$ from $(\tau, \nu)$;
    else
        Work out parameter $\theta$ from $\tau$;
    end
    Choose randomly $r \in \{0, 90, 180, 270\}$;
    if $r \in \{0, 180\}$ then
        $\tau_r \leftarrow \tau$;
    else
        $\tau_r \leftarrow -\tau$;
    end
    Simulate $n$ pairs of observations from $(U, V) \sim C^{(m_r)}(\cdot; \theta)$;
    Transform simulated observations to pseudo-observations $(u_j, v_j)_{j=1,\ldots,n}$;
    Calculate $\mathrm{AIC}^{(l)}$, $l \in \mathcal{M}'$;
    Save $m, r, \tau_r$, $(u_j, v_j)_{j=1,\ldots,n}$, and $\mathrm{AIC}^{(l)}$, $l \in \mathcal{M}'$;
    $i \leftarrow i + 1$;
end
```

model rotations – resulting in non-homogeneous classes. We address this challenge pragmatically, by using the sample statistics $\hat{\tau}$, $\hat{\zeta}$ to infer the rotation of the underlying model. Thus, as a first step pseudo-observations are transformed to have positive correlation and skewness. Subsequently, heatmap images are created from the transformed samples. As a second step, these heatmaps are classified, to select one of the 6 copula models in $m \in \mathcal{M}$. Thus in the image recognition stage we avoid the need to consider too many or very heterogeneous classes.[3]

Figure 9 depicts the workflow of the two-step image recognition approach. The two steps are discussed in more detail below.

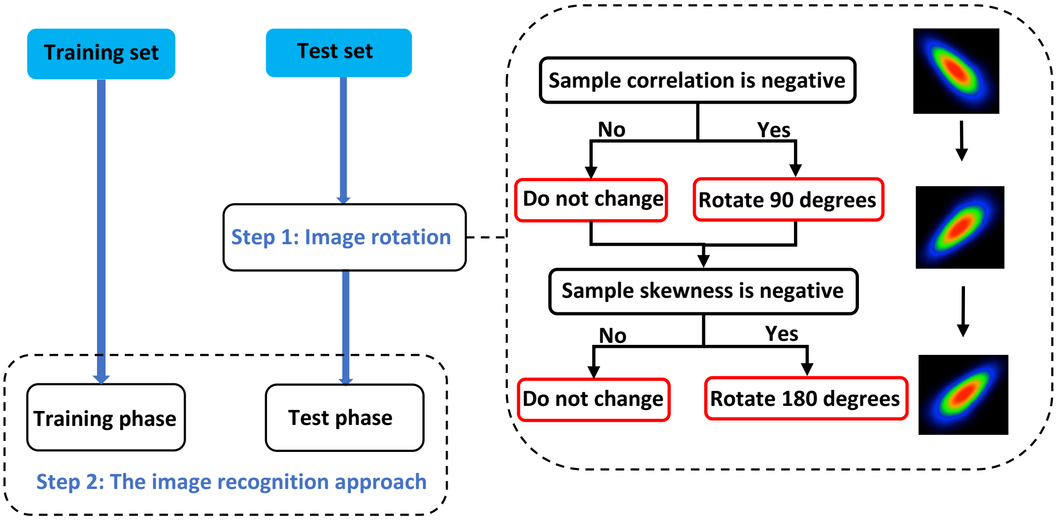


Figure 9: The workflow of the two-step approach for copula model selection.

Step 1

The pre-processing of samples in the test-set to arrive at homogeneous image classes, corresponding to the first step discussed above, is outlined in Algorithm 3. The algorithm first checks whether the rank correlation is negative – if so the data are rotated by 90 degrees counterclockwise. Then skewness is checked – if negative, the data are rotated by another 180 degrees. The quantity s represents the degree of data rotation to achieve a positive correlation and skewness; hence $\hat{r} = 360 - s$ is an estimate of the rotation r under which the pseudo-observations (u_j, v_j) were simulated. The statistics and images are exported after being calculated on the

[3] Note this two-step procedure gives the rationale for the rejections of training samples with negative skewness or correlation, in Algorithm 1, which will be adapted to this section. Importantly, in the test set of Algorithm 2, no such rejection of images takes place.

transformed sample (note that the saved AIC values are only used in the combined approach of Section 4.2.3). The process is illustrated on the right of Figure 9, which shows an example of the heatmap image of a 90-degree rotated Pareto copula.

At the end of Algorithm 3 we assess whether the first step has been performed successfully, in that the rotation applied to the data before producing the heatmap is consistent with the rotation of the underlying copula model. Symmetric and asymmetric models are treated differently, to reflect that for symmetric models $m_0 = m_{180},\ m_{90} = m_{270}$, which in practice requires us only to test whether the sign of the rank correlation τ_r of the copula model m_r matches that of the pseudo-observations, $\hat{\tau}$.

Step 2 In the second step, images generated by Step 1 are classified to models in $\mathcal{M}$. The classification approach proceeds analogously with what was discussed in Section 3.2. The only difference is that, when generating a training sample of $R = 20,000$ according to Algorithm 1, rather than using fixed n and τ, for each i these are now randomly chosen in $[100, 250]$ and $[0.1, 0.9]$ respectively.

Once again we randomly split the $R = 20,000$ to a training set containing with 70% of the samples and a validation set with 30% of the samples. A SVM with RBF kernel is trained based on training sets of size $0.7R = 14,000$. The training/validation split process is repeated 20 times.

Finally, the 2-step approach is applied to classify the samples in the test set. For each test sample, we obtain 20 classification results based on the 20 classifiers trained in the training phase. A majority vote is applied to these results to get the final decision. In other words, the copula model with the highest vote by the 20 results is selected for the test copula sample.

We count a test sample as correctly classified only if both steps in the classification process are successful. In other words, for a sample to be classified correctly we need both to be true:

1. In the first step, the data were rotated in a way consistent with the underlying copula model; the variable $FirstStep$ in Algorithm 3 has the value TRUE.

2. In the second step, the classifier identifies the correct copula model out of the six models in $\mathcal{M}$. In other words, if the model underlying a test set was $m_r \in \mathcal{M}'$, the prediction of the classifier is $m \in \mathcal{M}$.

4.2.3 Combining image recognition with AIC

The analysis of Section 3 has shown that image recognition and statistical approaches may be complementary, each being dominant in different ranges of n and τ. For that reason we

Algorithm 3: Algorithm for generating heatmap images from the test set, with variable fixed n, τ and model rotations. (First step of Section 4.2.2.)

```
$i \leftarrow 0$;
while $i < S$ do
    Read pseudo-observations $(u_j, v_j)_{j=1,\ldots,n}$, underlying model with $(m, r)$ such that $m_r \in \mathcal{M}'$,
     and rank correlation $\tau_r$, from the $i$th instance of the test set;
    Calculate $\hat{\tau}$ from $(u_j, v_j)_{j=1,\ldots,n}$;
    Initialise the degree to which data will be rotated, $s \leftarrow 0$;
    if $\hat{\tau} < 0$ then
        $s \leftarrow s + 90$;
        $v_j \leftarrow 1 - v_j,\ j = 1, \ldots, n$;
        $\hat{\tau}_s \leftarrow -\hat{\tau}$;
    end
    Calculate $\hat{\zeta}$ from $(u_j, v_j)_{j=1,\ldots,n}$;
    if $\hat{\zeta} < 0$ then
        $s \leftarrow s + 180$;
        $u_j \leftarrow 1 - u_j,\ v_j \leftarrow 1 - v_j,\ j = 1, \ldots, n$;
        $\hat{\zeta}_s \leftarrow -\hat{\zeta}$;
    end
    Estimate the copula rotation $\hat{r} \leftarrow 360 - s$;
    Calculate from $(u_j, v_j)_{j=1,\ldots,n}$, $\hat{\xi}_s$ and $\mathrm{AIC}_s^{(l)}$ for each $l \in \mathcal{M}$;
    Save $\hat{\tau}_s$, $\hat{\zeta}_s$, $\hat{\xi}_s$, $\mathrm{AIC}_s^{(l)}$;
    Transform pseudo-observations to normal $x_j \leftarrow \Phi^{-1}(u_j),\ y_j \leftarrow \Phi^{-1}(v_j),\ j = 1, \ldots, n$;
    Estimate joint density of $(x_j, y_j)$. Create heatmap and save image;
    if $m \in \mathcal{M}_s$ then
        if $sign(\hat{\tau}) = sign(\tau_r)$ then
            $FirstStep \leftarrow$ TRUE
        else
            $FirstStep \leftarrow$ FALSE
        end
    end
    if $m \in \mathcal{M}_a$ then
        if $\hat{r} = r$ then
            $FirstStep \leftarrow$ TRUE
        else
            $FirstStep \leftarrow$ FALSE
        end
    end
    $i \leftarrow i + 1$;
end
```

propose to combine the two approaches. We adapt the approach of Section 4.2.2, to integrate AIC information for different models in $\mathcal{M}$ in the second step of the process. Specifically:

1. We follow the Step 1 of Section 4.2.2 in exactly the same way.

2. In Step 2, we follow again the approach of 4.2.2, but now adding the AIC values for $l \in \mathcal{M}$ in both the training and testing phases. Specifically, for a training instance i denote the AIC values of different models by $\mathbf{x}_i^A = \{\text{AIC}_i^{(l)}\}_{l \in \mathcal{M}}$, $i = 1, 2, \ldots, N$ – these are extracted by Algorithm 1 (modified version with variable n, τ). The concatenated feature vector for each copula sample then becomes $\mathbf{x}_i = [(\mathbf{x}_i^{MP})^T, (\mathbf{x}_i^S)^T, (\mathbf{x}_i^A)^T]^T \in \mathbb{R}^{159 \times 1}$. The rest of the training phase follows exactly as described in the second step of Section 4.2.2. Similarly, a test sample is represented by $\mathbf{x}_t = [(\mathbf{x}_t^M)^T, (\mathbf{x}_t^S)^T, (\mathbf{x}_t^A)^T]^T \in \mathbb{R}^{159 \times 1}$, where $\mathbf{x}_t^A$ are the AIC values, extracted by Algorithm 3. $\mathbf{x}_t$ is then projected to the linear discriminant subspace constructed in the training phase and classified by SVM.

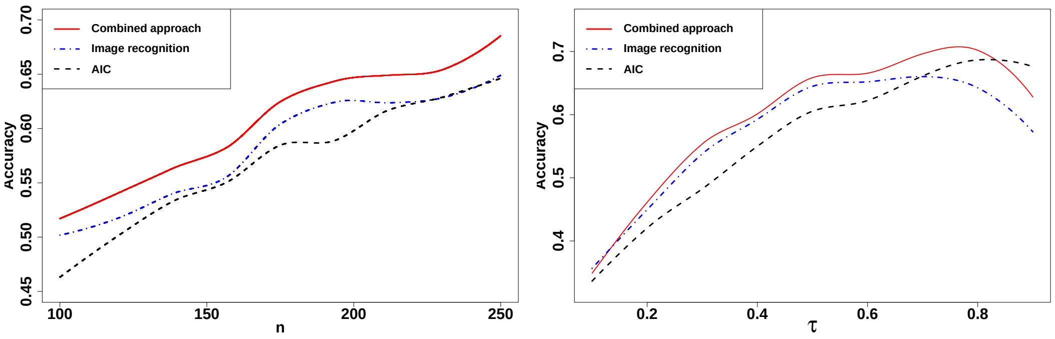
4.2.4 Classification results

First we compare the performance of the different copula model selection methods presented in Section 4.2 by calculating their classification accuracies on the test set. It is seen in Table 1 that the image recognition approach of Section 4.2.2 outperforms AIC. Furthermore, combining image recognition with AIC information, as in Section 4.2.3, leads to a better accuracy than either of those two approaches.

Table 1: Test classification accuracies of AIC, two-step image recognition approach, and combining image recognition with AIC.

	AIC	Image recognition	Combined
Accuracy	0.5688	0.5822	**0.6061**

To gain more insight into the test results, we plot smoothed curves of the average test classification accuracies of the three approaches against n and τ in Figures 10(a) and 10(b), respectively. The curves are generated by the function `locfit` of the package *locfit* with the smoothing parameter of 0.5. It is clear that the image recognition approach outperforms AIC, except for high correlations. Furthermore, the combined approach, including the information from both images and AIC, provides better results than simply using the image information in the two-step approach for all values of n and τ. Figure 10(a) shows that the combined approach is the best over all values of n while AIC is the worst, which is consistent with our conclusion in Table 1. However, Figure 10(b) presents a slightly different pattern for τ: AIC performs better



(a)

(b)

Figure 10: Fitted curves of the test classification accuracies of the three approaches against (a) n and (b) τ.

than both the combined and two-step approaches when the value of τ is larger than 0.8. This indicates that the heatmap image information does not help copula selection when τ is very high.

Table 2: Confusion matrix of the two-step image recognition approach on the test set.

True Predicted	Gaussian	t	Frank	Gumbel	Joe	Pareto
Gaussian	**1037**	391	181	179	27	50
t	287	**974**	51	204	23	18
Frank	197	102	**1360**	64	19	31
Gumbel	163	225	50	**868**	126	137
Joe	5	16	8	116	**1003**	803
Pareto	31	11	21	136	409	**580**
Accuracy	0.6029	0.5666	0.8144	0.5539	0.6241	0.3582

In Tables 2 and 3 respectively, we present the confusion matrices of the image recognition and combined approaches. The classification accuracies for each class are summarised in the bottoms of the two tables. For simplicity of exposition, we calculate the confusion matrices only for those testing instances where we had $FirstStep = \mathit{TRUE}$ (99.02% of instances). It is apparent that the Frank, Gaussian and Joe models are best predicted. Predictions for underlying Gumbel and t models are less accurate, while for Pareto models the predictions are the worst. Pareto models are confused with Joe by both approaches, due to the similarity between their heatmap

Table 3: Confusion matrix of the approach combining image recognition and AIC on the test set.

Predicted \ True	Gaussian	t	Frank	Gumbel	Joe	Pareto
Gaussian	**1181**	394	160	179	28	40
t	191	**1004**	53	203	24	19
Frank	171	97	**1390**	68	22	37
Gumbel	136	190	42	**870**	112	150
Joe	6	19	6	106	**979**	736
Pareto	35	15	19	141	442	**637**
Accuracy	0.6866	0.5841	0.8323	0.5552	0.6092	0.3935

images. Comparing Tables 2 and 3, we can see that the combined approach can provide better predictions on Gaussian and Pareto models, which demonstrates the effectiveness of involving the information provided by AIC.

4.3 Sensitivity analysis

We complete our discussion with a sensitivity analysis of the predictive model behind the image recognition approach of Section 4.2.2. Given the complexity of the workflow of Figure 9, and the concurrent use of image-based and statistical features, we are interested in understanding which of those features drive the predictions of each model. For this purpose we adapt the scenario weighting and reverse sensitivity framework developed by Pesenti et al. (2019) in the context of stress testing simulation models and implemented in the **R** package *SWIM* (Pesenti et al., 2021). This framework is well suited to situations where it is cumbersome or computationally expensive to repeatedly evaluate the prediction function on new observations.

We apply the sensitivity analysis on the test set, with $\mathbf{x}_t \in \mathbb{R}^{153}$ the feature vector of the tth sampling instance, for $t = 1, \dots, S$, where $S = 10,000$. Furthermore, for each testing instance we also consider the vector $\mathbf{y}_t \in \mathbb{R}^6$, where $y_{t,l}$ represents the number of votes obtained by the lth copula model as part of the majority voting procedure described in Section 3.2. Then, for each model $l = 1, \dots, 6$ we calculate a vector of weights in $\mathbb{R}^S$, such that, under re-weighting the sample $y_{1,l}, \dots, y_{S,l}$, the average number of votes for this model increases by 1. The vector of weights is selected by minimising the Kullback-Leibler divergence; specifically we solve the problem:

$$\begin{cases} \min_{\mathbf{w}\in\mathbb{R}^S} \frac{1}{S}\sum_{t=1}^S w_t \log(w_t) \quad \text{s. t.} \\ w_t > 0,\ t = 1,\dots,S \\ \frac{1}{S}\sum_{t=1}^S w_t = 1 \\ \frac{1}{S}\sum_{t=1}^S w_t y_{t,l} = \frac{1}{S}\sum_{t=1}^S y_{t,l} + 1. \end{cases}$$

The solution of this problem follows directly from Csiszár (1975); see Pesenti et al. (2019) for an overview of related work. The solution $\mathbf{w}^* \in \mathbb{R}^S$ applies a higher weight to those testing instances that drive the increase in the average vote for model l. Subsequently a sensitivity index for the ith feature can be defined as the normalised increase in the average of $x_{t,j},\ t = 1,\dots,S,\ j = 1,\dots,153$, over instances, arising from weighting by $\mathbf{w}^*$ – for more details see Pesenti et al. (2019).

The results of this analysis are shown in Figure 11, which plots the sensitivity of the majority vote for each of the models in $\mathcal{M}$ to the first 10 principal components of the heatmap images, as well as the statistical features $\hat{\tau}$ (tau) $\hat{\zeta}$ (skew), $\hat{\xi}$ (arach). It can be seen that the sensitivity to skewness $\hat{\zeta}$ is important for symmetric models (with a negative effect) and for the Pareto and Joe models (with a positive effect), consistently with the properties of these copulas. Beyond that, the main role in telling apart the different models is played by the image principal components; e.g. we can note the quite different patterns of PC1-PC10, for the 3 symmetric models on the left of the plot. On the other hand, for the Joe and Pareto models, which, as discussed, cannot be easily distinguished by the classification approach, the PC patterns are rather similar, confirming the information in the confusion matrix of Table 2.

5 Conclusions

In this paper, we proposed approaches for selecting copula models by utilising image recognition of the density heatmap images obtained from copula samples. PCA-reduced AlexNet image features and summary statistics were utilised to represent each copula sample and a low-dimensional discriminative projection by LDA used to train an SVM for classification. Experimental results showed that the proposed image recognition approach can provide improved classification performance on copula samples with relatively low sample size and correlation, compared to AIC. When combining image recognition with AIC, performance improves further.

Hence, we can answer our research question in the affirmative: indeed, heatmap images do contain relevant information for copula model selection that can help improve on standard procedures, and we propose workflows to harness this information for model selection. Furthermore, we demonstrate the potential value of transfer learning in statistical applications. In this paper,

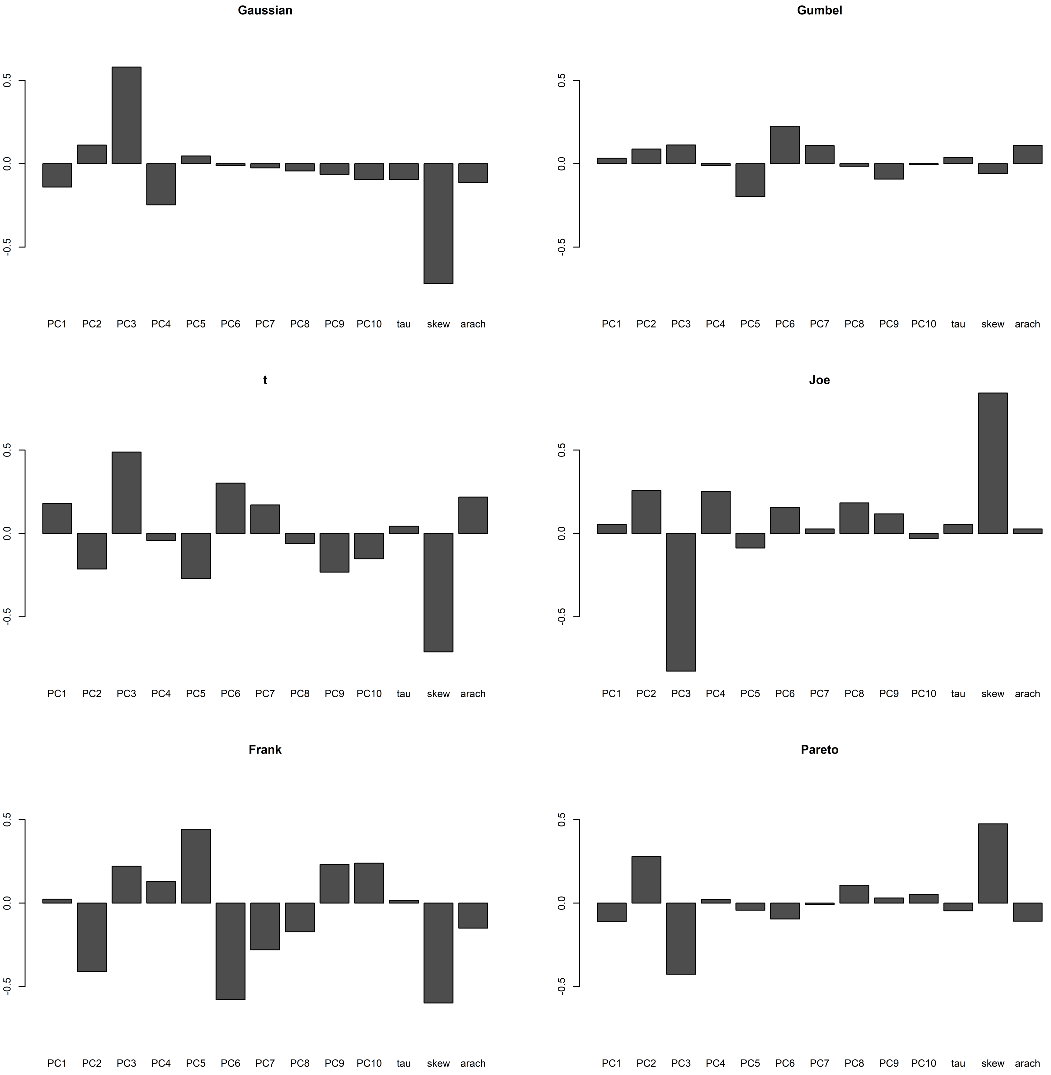


Figure 11: Sensitivity of majority vote for each copula model, to different features.

the knowledge obtained from the domain of natural images is transferred to the different domain of copula heatmap images, to assist in the new classification task of copula selection. This shows the potential of utilising machine learning algorithms that have been trained on a large amount of high-quality data to provide additional useful information for statistical applications, or to help in situations where collecting enough data is not an easy task (Pan and Yang, 2009).

The workflows we propose are more complex than evaluating likelihood-based criteria, at least where efficient implementation of the latter is available. We do not argue for replacing well-established criteria such as AIC, but provide a 'proof of concept' that image recognition can supplement standard statistical approaches. With this in mind, future work can consider designing an expert system with a broader scope, e.g. including a wider range of models and sample sizes, as well as handling multivariate dependence structures. For example, a copula sample with multivariate dependence structure can be partially represented by several two-dimensional heatmap images generated from pairwise bivariate dependence structures. In this way, each sample is represented by a set of images, which leads to the image set classification task in computer vision (Fukui and Maki, 2015).

References

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

Abdallah, A., Boucher, J.-P., and Cossette, H. (2015). Modeling dependence between loss triangles with hierarchical archimedean copulas. *ASTIN Bulletin*, 45(3):577–599.

Androschuck, T., Gibbs, S., Katrakis, N., Lau, J., Oram, S., Raddall, P., Semchyshyn, L., Stevenson, D., and Waters, J. (2017). Simulation-based capital models: testing, justifying and communicating choices. a report from the life aggregation and simulation techniques working party. *British Actuarial Journal*, 22(2):257–335.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158.

Czado, C., Kastenmeier, R., Brechmann, E. C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4):278–305.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2006). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. John Wiley & Sons.

Embrechts, P., Puccetti, G., Rüschendorf, L., Wang, R., and Beleraj, A. (2014). An academic response to Basel 3.5. *Risks*, 2(1):25–48.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25.

Frees, E. W. and Wang, P. (2005). Credibility using copulas. *North American Actuarial Journal*, 9(2):31–48.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics, New York.

Fukui, K. and Maki, A. (2015). Difference subspace and its generalization for subspace-based methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2164–2177.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368.

Genest, C., Nešlehová, J. G., Rémillard, B., and Murphy, O. A. (2019). Testing for independence in arbitrary distributions. *Biometrika*, 106(1):47–68.

Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213.

Grønneberg, S. and Hjort, N. L. (2014). The copula information criteria. *Scandinavian Journal of Statistics*, 41(2):436–459.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2020). *copula: Multivariate Dependence with Copulas*. R package version 1.0-1.

Hofert, M., Prasad, A., and Zhu, M. (2021). Quasi-random sampling for multivariate distributions via generative neural networks. *Journal of Computational and Graphical Statistics*, 30(3):647–670.

Hu, S., Murphy, T. B., and O'Hagan, A. (2021). mvClaim: an R package for multivariate general insurance claims severity modelling. *Annals of Actuarial Science*, 15(2):441–457.

Huard, D., Evin, G., and Favre, A.-C. (2006). Bayesian copula selection. *Computational Statistics & Data Analysis*, 51(2):809–822.

Islam, M., Dinh, A., Wahid, K., and Bhowmik, P. (2017). Detection of potato diseases using image segmentation and multiclass support vector machine. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE.

Jordanger, L. A. and Tjøstheim, D. (2014). Model selection of copulas: AIC versus a cross validation copula information criterion. *Statistics & Probability Letters*, 92:249–255.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Kularatne, T. D., Li, J., and Pitt, D. (2021). On the use of Archimedean copulas for insurance modelling. *Annals of Actuarial Science*, 15(1):57–81.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools – revised edition.* Princeton University Press.

Michiels, F. and De Schepper, A. (2013). A new graphical tool for copula selection. *Journal of Computational and Graphical Statistics*, 22(2):471–493.

Nelsen, R. B. (2007). *An Introduction to Copulas.* Springer.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pesenti, S. M., Bettini, A., Millossovich, P., and Tsanakas, A. (2021). Scenario Weights for Importance Measurement (SWIM) – an R package for sensitivity analysis. *Annals of Actuarial Science*, 15(2):458–483.

Pesenti, S. M., Millossovich, P., and Tsanakas, A. (2019). Reverse sensitivity testing: What does it take to break the model? *European Journal of Operational Research*, 274(2):654–670.

Rosco, J. and Joe, H. (2013). Measures of tail asymmetry for bivariate copulas. *Statistical Papers*, 54(3):709–726.

Rüschendorf, L. and de Valk, V. (1993). On regression representations of stochastic processes. *Stochastic Processes and their Applications*, 46(2):183–198.

Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., et al. (2021). Package 'vinecopula'. *R package version*, 2.4.2.

Shaw, R., Smith, A. D., and Spivak, G. (2010). Calibration and communication of dependencies with a case study based on market returns. URL: https://www.actuaries.org.uk/system/files/documents/pdf/b3.pdf. Presented to the Institute and Faculty of Actuaries, November 2010.

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., and Homayouni, S. (2020). Support vector machine vs. random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Shi, P. (2014). A copula regression for modeling multivariate loss triangles and quantifying reserving variability. *ASTIN Bulletin*, 44(1):85–102.

Shi, P. and Frees, E. W. (2011). Dependent loss reserving using copulas. *ASTIN Bulletin*, 41(2):449–486.

Shi, P. and Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, 55:18–29.

Tzotsos, A. and Argialas, D. (2008). Support vector machine classification for object-based image analysis. In *Object-Based Image Analysis*, pages 663–677. Springer.

Tzougas, G. and Pignatelli di Cerchiara, A. (2021). The multivariate mixed negative binomial regression model with an application to insurance a posteriori ratemaking. *Insurance: Mathematics and Economics*.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52.

Yang, L. and Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State University*, 2(2):4.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.