Journal of Multivariate Analysis 151 (2016) 69–89

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas

Thomas Nagler*, Claudia Czado

Department of Mathematics, Technische Universität München, Boltzmanstraße 3, 85748 Garching, Germany

HIGHLIGHTS

- A nonparametric density estimator based on simplified vine copulas is proposed.
- The convergence rate of the proposed estimator is independent of dimension.
- Simulations illustrate a huge gain in accuracy if the model assumption is true.
- When it is violated, the estimator is still favorable if the dimension is not small.

ARTICLE INFO

Article history:
Received 2 December 2015
Available online 29 July 2016

AMS subject classifications:
primary 62G07
secondary 62G20

Keywords:
Asymptotic
Classification
Copula
Dependence
Kernel density
Vine

ABSTRACT

Practical applications of nonparametric density estimators in more than three dimensions suffer a great deal from the well-known curse of dimensionality: convergence slows down as dimension increases. We show that one can evade the curse of dimensionality by assuming a simplified vine copula model for the dependence between variables. We formulate a general nonparametric estimator for such a model and show under high-level assumptions that the speed of convergence is independent of dimension. We further discuss a particular implementation for which we validate the high-level assumptions and establish asymptotic normality. Simulation experiments illustrate a large gain in finite sample performance when the simplifying assumption is at least approximately true. But even when it is severely violated, the vine copula based approach proves advantageous as soon as more than a few variables are involved. Lastly, we give an application of the estimator to a classification problem from astrophysics.

1. Introduction

Density estimation is one of the most important problems in nonparametric statistics. Most commonly, nonparametric density estimators are used for exploratory data analysis, but find many further applications in fields such as astrophysics, forensics, or biology [7,4,32]. Many of these applications involve the estimation of multivariate densities. However, most applications so far focus on two- or three-dimensional problems. Furthermore, the persistent interest amongst practitioners is contrasted by a falling tide of methodological contributions in the last two decades.

A probable reason is the prevalence of the *curse of dimensionality*: due to sparseness of the data, nonparametric density estimators converge more slowly to the true density as dimension increases. Put differently, the number of observations required for sufficiently accurate estimates grows excessively with the dimension. As a result, there is very little benefit from the ever-growing sample sizes in modern data. Section 7.2 in [44] illustrates this phenomenon for a kernel density

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

http://dx.doi.org/10.1016/j.jmva.2016.07.003
0047-259X/© 2016 Elsevier Inc. All rights reserved.

estimator when the standard Gaussian is the target density: to achieve an accuracy comparable to $n = 50$ observations in one dimension, more than $n = 10^6$ observations are required in ten dimensions.

In general, this issue cannot be solved: Stone [48] proved that any estimator $\widehat{f}$ that is consistent for the class of p times continuously differentiable d-dimensional density functions converges at a rate of at most $n^{-p/(2p+d)}$. More precisely,

$$\widehat{f}(\boldsymbol{x}) = f(\boldsymbol{x}) + O_p(n^{-r}),$$

for all densities f of this class and some $r > 0$, implies that $r \leq p/(2p+d)$. The curse of dimensionality manifests itself in the d in the denominator. It implies that the optimal convergence rate necessarily decreases in higher dimensions. Thus, to evade the curse of dimensionality, all we can hope for is to find subclasses of densities for which the optimal convergence rate does not depend on d. One such subclass is the collection of density functions corresponding to independent variables, which can be estimated as a simple product of univariate density estimates. But the independence assumption is very restrictive. We also want the subclass to be rich and flexible. We will show that simplified vine densities are such a class and provide a useful approximation even when the simplifying assumption is severely violated.

1.1. *Nonparametric density estimation based on simplified vine copulas*

We introduce a nonparametric density estimator whose convergence speed is independent of the dimension. The estimator is build on the foundation of a simplified vine copula model, where the joint density is decomposed into a product of marginal densities and bivariate copula densities, see, e.g., [12] and Section 3.9 in [29].

First, we separate the marginal densities and the copula density (which captures the dependence between variables). Let $(X_1, \dots, X_d) \in \mathbb{R}^d$ be a random vector with joint distribution F and marginal distributions $F_1, \dots, F_d$. Provided densities exist, Sklar's Theorem [45] allows us to rewrite the joint density f as the product of a copula density c and the marginal densities $f_1, \dots, f_d$: for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$f(\boldsymbol{x}) = c\big\{F_1(x_1), \dots, F_d(x_d)\big\} \times f_1(x_1) \times \cdots \times f_d(x_d),$$

where c is the density of the random vector $\big(F_1(X_1), \dots, F_d(X_d)\big) \in [0, 1]^d$. In order to estimate the joint density f, we can therefore obtain estimates of the marginal densities $f_1, \dots, f_d$ and the copula density c separately, and then plug them into the above formula. With respect to the curse of dimensionality, nothing is gained (so far) since estimation of the copula density is still a d-dimensional problem.

A crucial insight is that any d-dimensional copula density can be decomposed into a product of $d(d-1)/2$ bivariate (conditional) copula densities [5]. Equivalently, one can build arbitrary d-dimensional copula densities by using $d(d-1)/2$ building blocks (so-called *pair-copulas*). Following this idea, the flexible class of *vine copula* models – also known as *pair-copula-constructions (PCCs)* – were introduced in [1] and have seen rapidly increasing interest in recent years. For instance, a three-dimensional joint density can be decomposed as

$$\begin{aligned} f(x_1, x_2, x_3) = {} & c_{1,2}\big\{F_1(x_1), F_2(x_2)\big\} \times c_{2,3}\big\{F_2(x_2), F_3(x_3)\big\} \times c_{1,3;2}\big\{F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2);\ x_2\big\} \\ & \times f_1(x_1) \times f_2(x_2) \times f_3(x_3), \end{aligned}$$

where $c_{1,3;2}$ is the joint density corresponding to the conditional random vector $\big(F_{1|2}(X_1|X_2), F_{3|2}(X_3|X_2)\big)\big|X_2 = x_2$. Note that the copula of the vector depends on the value x_2 of the conditioning variable X_2. To reduce the complexity of the model, it is usually assumed that the influence of the conditioning variable on the copula can be ignored. In this case, the conditional density $c_{1,3;2}$ collapses to an unconditional – and most importantly, two-dimensional – object, and one speaks of the *simplifying assumption* or a simplified vine copula model/PCC. For general dimension d, a similar decomposition into the product of d marginal densities and $d(d-1)/2$ pair-copula densities holds.

Some copula classes where the simplifying assumption is satisfied are given in [47]. An important special case is the Gaussian copula. It is the dependence structure underlying a multivariate Gaussian distribution and can be fully characterized by $d(d-1)/2$ partial correlations. Note that under a multivariate Gaussian model, conditional correlations and partial correlations coincide. This property is in direct correspondence to the simplifying assumption which states that all conditional copulas collapse to partial copulas. When the Gaussian copula is represented as a vine copula, it consists of $d(d-1)/2$ Gaussian pair-copulas where the copula parameter of each pair corresponds to the associated partial correlation. In a general simplified vine copula model, we replace each Gaussian pair-copula by an arbitrary bivariate copula. Such models are extremely flexible and encompass a wide range of dependence structures. The class of simplified vine distributions is even more flexible, because it allows to couple a simplified vine copula model with arbitrary marginal distributions.

Under the simplifying assumption, a d-dimensional copula density can be decomposed into $d(d-1)/2$ unconditional bivariate densities. Consequently, the estimation of a d-dimensional copula density can be subdivided into the estimation of $d(d-1)/2$ two-dimensional copula densities. Intuitively, we expect that the convergence rate of such an estimator will be equal to the rate of a two-dimensional estimator and, thus, there is no curse of dimensionality. This is formally established by our main result: Theorem 1.

Nonparametric estimation of simplified vine copula densities has been discussed earlier using kernels [34] and smoothing splines [30]. However, both contributions lack an analysis of the asymptotic behavior of the estimators. We treat the more