Technische Universität München
Zentrum Mathematik
Lehrstuhl für Mathematische Statistik

Nonparametric estimation in simplified vine copula models

Thomas Werner Nagler

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:		Prof. Dr. Silke Rolles
Prüfer der Dissertation:	1.	Prof. Claudia Czado, Ph.D.
	2.	Prof. Dr. Irène Gijbels (nur schriftliche Beurteilung)
	3.	Prof. Harry Joe, Ph.D.

Die Dissertation wurde am 02.08.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 18.10.2018 angenommen.

Zusammenfassung

Nichtparametrische Dichteschätzer mit mehreren Variablen leiden unter einem bekannten Phänomen. Der *Fluch der Dimensionen* besagt, dass die Konvergenzgeschwindigkeit notwendigerweise abnimmt, wenn die Anzahl der Variablen steigt. Wir zeigen, dass man diesem Fluch entkommen kann, wenn man ein vereinfachtes Vine-Copula-Modell für die Abhängigkeit zwischen den Variablen annimmt. Wir vergleichen bestehende Methoden zur nichtparametrischen Dichteschätzung in solchen Modellen in einer großen Simulationsstudie und identifizieren die wichtigen Einflussfaktoren für deren Genauigkeit. Im Anschluss erweitern wir die Anwendbarkeit der vorgestellten Methoden in zweierlei Hinsicht. Wir stellen einen Ansatz vor, um allgemeine Regressionsprobleme mithilfe von copula-basierten Schätzungsgleichungen zu lösen. Wir leiten dessen asymptotische Eigenschaften unter allgemeinen Annahmen her und veranschaulichen seine vielseitige Einsetzbarkeit mit theoretischen und simulations-basierten Beispielen. Weiterhin diskutieren wir eine generische Methode, um nichtparametrische Funktionenschätzer auf Daten mit diskreten Variablen anwendbar zu machen. Durch Hinzufügen von künstlichem Rauschen kann man das Schätzproblem in ein rein stetiges Equivalent verwandeln. Abschlieend zeigen wir anhand eines ausführlichen Beispiels, dass dies nicht notwendigerweise die Effizienz der Schätzmethode beeinträchtigt.

Abstract

Practical applications of nonparametric density estimators with multiple variables 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 compare existing methods for the estimation of simplified vine copula densities in an extensive simulation study and identify the driving factors for their performance. We further extend the applicability of this technique in two ways. We introduce an approach to estimate regression functions using estimating equations based on the copula density. The method's asymptotic properties are derived under broad conditions and its versatility is illustrated with theoretical and simulated examples. Finally, we discuss a generic technique to make nonparametric function estimators applicable to discrete and mixed data types. By adding noise to the discrete variables, we can transform the problem into a purely continuous equivalent and show that this does not necessarily lead to a loss in efficiency.

Acknowledgments

I am deeply grateful to Prof. Claudia Czado for giving me the opportunity to pursue a PhD and the time and energy she put into my development as a researcher. She gave me the freedom to find my own path and supported me in every way possible. I also thank my fellow PhD students who made my office days so enjoyable (although I never got any work done with all the chatting). Another colleague that deserves special mention is my remote office mate Dr. Thibault Vatter. He has been an amazing collaborator and good friend — although being first hundreds, then thousands kilometers apart.

I further thank my dearest friends Alex, Basti, and Michi for all the fun we had and being by my side in good and bad times. And most importantly, I am deeply indebted to my family and Jojo for their love and support. This would not have been possible without you.

Contents

1 Introduction 1

1.1 Motivation 1

1.2 Agenda of the thesis 2

2 Preliminaries 5

2.1 Copulas 5

2.2 Vine copulas 6

2.3 Stochastic convergence 9

2.4 Empirical processes 13

3 Nonparametric estimators of simplified vine copula densities 19

3.1 Introduction 19

3.2 Implementation 20

3.3 Description of the simulation study design 29

3.4 Results 32

3.5 Illustration with real data 40

3.6 Conclusion 43

4 Evading the curse of dimensionality with simplified vine copulas 45

4.1 Introduction 45

4.2 Nonparametric density estimators based on simplifed vine copulas 48

4.3 Main results 49

4.4 On an implementation as kernel estimator 53

4.5 Simulations 57

4.6 Application: filtering noise in telescope imaging 65

4.7 Discussion 68

4.8 Proofs 69

5 Solving estimating equations with copulas 79

5.1 Introduction 79

5.2 Copula-based solutions to estimating equations 80

5.3 Asymptotic theory 83

5.4 Simulations 93

5.5 Discussion 97

5.6 Proofs 100

6 The jittering technique for nonparametric function estimation 107

6.1 Introduction 107

6.2 Jittering mixed data 108

6.3 Nonparametric function estimation via jittering 112

6.4 Application: diagnosis of retinopathy 122

6.5 Discussion 123

7 Asymptotic analysis of the jittering kernel density estimator 127
7.1 Introduction . . . 127
7.2 The estimator . . . 128
7.3 Asymptotic analysis in the general setting . . . 129
7.4 A closer look at the univariate discrete setting . . . 133
7.5 Minimax rate optimality . . . 136
7.6 Simulation experiments . . . 138
7.7 Conclusion . . . 140
7.8 Proofs . . . 141

8 Conclusion 147
8.1 Summary . . . 147
8.2 Outlook . . . 147

References 149

"Data is the sword of the 21st century."
— Jonathan Rosenberg

1 Introduction

1.1 Motivation

With continuing technological advances, our society has become increasingly data-centric. This is most apparent in our economy, where companies across all industries are amassing data on their customers and products. A similar development is taking place in scientific research. The majority of natural and social sciences are more and more driven by data. For many areas, this constitutes a big paradigm shift, and not few are still struggling to adapt their methods and best practices. Even in our private lives, data plays an increasing role. Not only do others collect and exploit data about our habits and lifestyle; smart phones and watches make it easy to record data about ones own biorhythm, health, or happiness.

Collecting data is rarely an end in itself. The larger purpose is to learn from the data in the hope to improve revenue, knowledge, or well-being. Data collection has become so cheap that a typical data set contains information on numerous quantities, many of which are interconnected. Thus, the dependence between variables is key for understanding the data and learning from them.

The most general statistical concept for dependence is the *copula* and dates back to at least Sklar (1959), with similar ideas already put forth by Hoeffding (1940). A copula allows to separate the joint behavior of variables into two parts: the individual behavior of the variables (which is captured by the *marginal distributions*), and their dependence (which is captured by the copula). Conversely, if one knows the marginal distributions and copula, one has full knowledge of the joint behavior of the variables. Due to the simplicity and power of this concept, copulas have seen tremendous interest among researchers and practitioners alike. Numerous different copula models have been developed and applied in widespread domains (for reviews, see, e.g., Salvadori and De Michele, 2007, Elidan, 2013, Aas, 2016).

Most commonly, copulas are used in a *parametric* fashion: the copula is assumed to belong to a family of functions which is characterized by a finite number of parameters. However, a small number of parameters drastically limits the flexibility — and, thereby, the type of dependence that can be reflected by the model. One of the most promising classes that emerged are vine copulas (Joe, 1996, Bedford and Cooke, 2001, Aas et al., 2009). Vine copulas are hierarchical models that build the dependence structure from bivariate building blocks, called *pair-copulas*. Each pair-copula captures the dependence between a (conditional) pair of variables. Because every pair-copula can be parametrized differently, vine copulas allow each pair to have a different strength and type of dependence.

But also vine copula models are limited in flexibility by the number of parameters for each pair. In some situations, there is no parametric model that reflects the dependence of a pair adequately. For such cases, a *nonparametric* philosophy is more appropriate. In this philosophy, we start with minimal assumptions on a distribution's shape, and "let the data speak for themselves". In my Master's thesis (Nagler, 2014), I developed tools to estimate the dependence under a nonparametric vine copula model. The main motivation behind the research presented in this thesis is to better understand such tools and make them more widely applicable.

1.2 Agenda of the thesis

The content of this thesis is based on five research papers:

- Nagler, T. (2018a). Asymptotic analysis of the jittering kernel density estimator. *Mathematical Methods of Statistics*, 27(1):32–46.
- Nagler, T. (2018b). A generic approach to nonparametric function estimation with mixed data. *Statistics & Probability Letters*, 137:326–330.
- Nagler, T. and Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89.
- Nagler, T., Schellhase, C., and Czado, C. (2017). Nonparametric estimation of simplified vine copula models: comparison of methods. *Dependence Modeling*, 5(1):99–120.
- Nagler, T. and Vatter, T. (2018b). Solving estimating equations with copulas. *arXiv:1801.10576*.

The contents of these papers have been revised and many chapters contain new materials, corrections, and additional arguments or clarifications.

We start with a brief introduction to the foundational tools and concepts of this thesis in Chapter 2. The next two chapters try to deepen our understanding about nonparametric estimators for vine copula models. Chapter 3 (which is based on Nagler et al., 2017) surveys and extends existing nonparametric estimators of simplified vine copula densities. The methods are compared in an extensive simulation study. We identify and discuss several factors driving the relative performance of the estimators: dimension, sample size, strength and type of dependence.

The empirical assessment is complemented by theoretical results in Chapter 4, which is based on Nagler and Czado (2016). It follows up on an conjecture in Nagler (2014) that simplified vine copula models allow to estimate a joint density without the *curse of dimensionality*. This curse is a well-known phenomenon in the world of nonparametric function estimation (see, e.g. Scott, 2008): as the number of variables increases, the convergence rate of estimators slows down. This issue is so severe that fully nonparametric estimators are rarely used with more than a hand full of variables. Under a simplified vine copula model, the estimation of any d-dimensional density boils down to the estimation of one-dimensional marginals and two-dimensional pair-copulas. Hence, there is hope that its density can be estimated at a faster convergence rate. This is established by the chapter's main result: Theorem 4.1 gives high level conditions under which the density is estimated with a rate equivalent to a two-dimensional problem, irrespective of d. We discuss the assumptions in the context of a specific implementation as a kernel estimator and establish its asymptotic normality. After confirming the theoretical findings by simulations, the estimator is applied to a classification problem from astrophysics.

The unifying goal for the remaining chapters is to extend the applicability of nonparametric vine copula estimators. Another commonality is that they take a more high-level view instead of focusing on vine copula models, giving the results and concepts a larger scope.

Since the copula contains all information about the dependence, it also characterizes regression relationships between a variable of interest and explanatory variables. This idea was seized by many researchers in the past, especially in the context of mean and quantile regression. Chapter 4 (based on Nagler and Vatter, 2018b) frames such approaches as solutions of estimating equations and generalizes them in a unified framework. The main results establish consistency (Theorem 5.1), asymptotic normality (Theorem 5.2), and validity of the bootstrap (Theorem 5.3) for copula-based Z-estimators. The results are formulated under high-level conditions that allow for parametric, semiparametric, and fully nonparametric estimators of the copula density and marginal distributions. We further illustrate the versatility of such estimators through theoretical and simulated examples.

A big drawback of the methods and theory of the previous chapters is that they are only valid for continuous random variables. This is problematic: in most domains, data contain discrete variables, such as counts or categories (e.g., gender, age group, alive/dead). The reason for this drawback is twofold. Subtle technical and conceptual issues arise when copulas are used to model discrete data (see, e.g., Genest and Neslehova, 2007). With some additional effort, vine copula models can be extended to work with discrete data; see Panagiotelis et al. (2012), Stöber et al. (2015), Panagiotelis et al. (2017), and Joe (2014a, Section 3.9.5). However, most techniques for nonparametric function estimation are only valid if data are continuous. A simple but slightly awkward solution is discussed in Chapter 6, which is largely based on Nagler (2018b). A common trick among practitioners is to make the discrete variables continuous by adding a small amount of noise (often called *jittering*). We formalize this trick and show that it allows for valid estimates under suitable conditions on the noise density. Several examples will show that this trick is easily and broadly applicable, e.g., to density estimation, regression, and classification. In particular, it bypasses all obstacles related to the use of copulas for discrete data.

Since we are adding noise to the data, a natural concern is that jittering estimators will lose in efficiency. This is investigated for a simple example in Chapter 7, which is based on Nagler (2018a). We give an in-depth analysis of the jittering kernel density estimator, which reveals several appealing properties. We show that the estimator is asymptotically normal and unbiased for discrete variables (Theorem 7.1), as well as strongly consistent (Theorem 7.2). It further converges at minimax-optimal rates, which are established as a by-product of our analysis (Theorem 7.3 and Theorem 7.4). To understand the effect of adding noise, we further study asymptotic efficiency and finite sample bias and conduct a small simulation study.

Chapter 8 offers concluding remarks and discusses future directions of research.

2 Preliminaries

This chapter introduces the central concepts in this thesis: copulas, vine copulas, stochastic convergence, and empirical processes. It is primarily meant to refresh the readers' memory and provide references to more extensive treatments.

2.1 Copulas

Copulas are mathematical objects that encode the stochastic dependence between random variables. For an extensive introduction to copulas, the reader is referred to the monographs of Nelsen (2006) and Joe (2014b).

Sklar's theorem (Sklar, 1959) states that any multivariate distribution function F can be split into its marginal distributions $F_1, \dots, F_d$ and a *copula* C:

$$F(x_1, \dots, x_d) = C\big(F_1(x_1), \dots, F_d(x_d)\big) \tag{2.1}$$

Conversely, one can combine arbitrary marginal distributions and copulas to obtain a valid multivariate distribution. Further, if F admits a density with respect to the Lebesgue measure, we can differentiate the above equation to get

$$f(x_1, \dots, x_d) = c\big(F_1(x_1), \dots, F_d(x_d)\big) \times \prod_{k=1}^{d} f_k(x_k), \tag{2.2}$$

where c and $f_1, \dots, f_d$ are the probability density functions corresponding to C and $F_1, \dots, F_d$ respectively.

Now suppose that there is a random vector $\boldsymbol{X}$ with joint distribution F. If the distributions F_i are continuous, C is the joint distribution of the random vector $\boldsymbol{U} = (U_1, \dots, U_d) = \big(F_1(X_1), \dots, F_d(X_d)\big)$. Note that $U_1, \dots, U_d$ are defined as the *probability integral transforms* of $X_1, \dots, X_D$ and, hence, uniformly distributed on the unit interval.

Given an *iid* sequence of random variables $\boldsymbol{X}_i$, $i = 1, \dots, n$ (acting as observations), this suggests a two-step procedure for the estimation of C: first, estimate the marginal distributions by $\widehat{F}_1, \dots, \widehat{F}_d$; then estimate C based on the *pseudo-observations* $\widehat{\boldsymbol{U}_i} = \big(\widehat{F}_1(X_{i,1}), \dots, \widehat{F}_d(X_{i,d})\big)$.

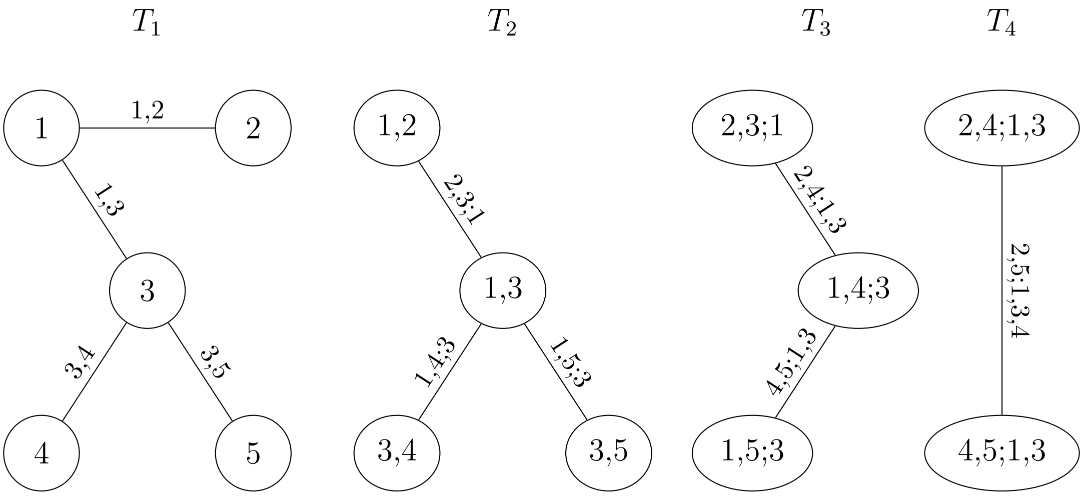


Figure 2.1: Example of a regular vine tree sequence.

2.2 Vine copulas

Vine copula models build a full d-variate dependence model using a collection of bivariate building blocks. A thorough introduction to vine copulas can be found in Aas et al. (2009) Czado (2010), and Joe (2014b).

Vine copula models follow the idea of Joe (1996) that any d-dimensional copula can be expressed in terms of $d(d-1)/2$ bivariate (conditional) copulas. Because such a decomposition is not unique, Bedford and Cooke (2002) introduced a graphical method to organize the possibilities in terms of linked trees $T_m = (V_m, E_m)$, $m = 1, \dots, d-1$.

2.2.1 Tree representation

A sequence $\mathcal{V} := (T_1, \dots, T_{d-1})$ of trees is called a *regular vine (R-vine) tree sequence* on d elements if the following conditions are satisfied:

(i) T_1 is a tree with nodes $V_1 = \{1, \dots, d\}$ and edges E_1.

(ii) For $m \geq 2$, T_m is a tree with nodes $V_m = E_{m-1}$ and edges E_m.

(iii) (*Proximity condition*) Whenever two nodes in T_{m+1} are joined by an edge, the corresponding edges in T_m must share a common node.

The tree sequence is also called the *structure* of the vine. An example of an R-vine tree sequence for $d = 5$ is given in Figure 2.1. For the annotation of the edges in each tree we follow (Czado, 2010).

2.2.2 Vine decomposition of copula densities

Bedford and Cooke (2001) showed that *any* copula density can be represented as

$$c(\boldsymbol{u}) = \prod_{m=1}^{d-1} \prod_{e \in E_m} c_{j_e,k_e;D_e}\big\{G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}),\, G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e});\, \boldsymbol{u}_{D_e}\big\}, \tag{2.3}$$

where

- $\boldsymbol{u}_{D_e} := (u_\ell)_{\ell \in D_e}$ is a subvector of $\boldsymbol{u} = (u_1, \dots, u_d) \in [0,1]^d$,
- $(E_m)_{m=1}^{d-1}$ are the edges of an arbitrary R-vine tree sequence,
- $G_{j_e|D_e}$ is the conditional distribution of $U_{j_e}|\boldsymbol{U}_{D_e} = \boldsymbol{u}_{D_e}$,
- $c_{j_e,k_e;D_e}$ is the copula density associated with the conditional random vector

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

The set D_e is called *conditioning set* and the indices j_e, k_e form the *conditioned set*. In the first tree the conditioning set D_e is empty, and we define $G_j(u) := u$ for notational consistency. An *R-vine copula model* identifies each edge of the trees with a bivariate (conditional) copula model (a so-called *pair-copula*). Conversely, arbitrary (conditional) copula densities can be plugged into (2.3) to obtain a well-defined copula density.

Example 2.1. *The density of an R-vine copula corresponding to the tree sequence in Figure 2.1 is*

$$\begin{aligned} c(u_1, \dots, u_5) = {} & c_{1,2}(u_1, u_2) \times c_{1,3}(u_1, u_3) \times c_{3,4}(u_3, u_4) \times c_{3,5}(u_3, u_5) \\ & \times c_{2,3;1}(u_{2|1}, u_{3|1}; u_1) \times c_{1,4;3}(u_{1|3}, u_{4|3}; u_3) \times c_{1,5;3}(u_{1|3}, u_{5|3}; u_3) \\ & \times c_{2,4;1,3}(u_{2|1,3}, u_{4|1,3}; \boldsymbol{u}_{1,3}) \times c_{4,5;1,3}(u_{4|1,3}, u_{5|1,3}; \boldsymbol{u}_{1,3}) \\ & \times c_{2,5;1,3,4}(u_{2|1,3,4}, u_{5|1,3,4}; \boldsymbol{u}_{1,3,4}), \end{aligned}$$

where we used the abbreviation $u_{j_e|D_e} = G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e})$*.*

2.2.3 Simplified vine copula models

In (2.3), the pair-copula densities $c_{j_e,k_e;D_e}$ takes $\boldsymbol{u}_{D_e}$ as an argument and the functional form with respect to the first two arguments may be different for each value of $\boldsymbol{u}_{D_e}$. This conditional structure makes the model very complex and complicates estimation. To simplify matters, it is commonly assumed that the conditional copula is equal across all possible values of $\boldsymbol{u}_{D_e}$: we say that the *simplifying assumption* holds. In this case, (2.3) collapses to

$$c(\boldsymbol{u}) = \prod_{m=1}^{d-1} \prod_{e \in E_m} c_{j_e,k_e;D_e}\big\{G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}),\, G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e})\big\}. \tag{2.4}$$

A copula whose density can be represented this way is called a *simplified vine copula*. The pair-copulas $c_{j_e,k_e;D_e}$ no longer take the conditioning values as an argument. The conditioning values $\boldsymbol{u}_{D_e}$ only affect the joint density c through the arguments of the pair-copulas. Hence, the pair-copula densities $c_{j_e,k_e;D_e}$ encode *partial dependence* rather than conditional dependence. This reduction is similar to going from conditional correlations to partial correlations: the influence of the conditioning variable on the dependence is "averaged out".

For the multivariate Gaussian, the simplifying assumption holds for any vine. The corresponding vine copula model consists of only Gaussian pair-copulas whose parameters are the partial correlation coefficients. Hence, simplified vine copulas are an extension of the multivariate Gaussian copula that allows for non-Gaussian pair-copulas.

For more on partial dependence and copulas, see Bergsma (2011), Gijbels et al. (2015b), and Spanhel and Kurz (2016). More about the simplifying assumption can be found in Hobæk Haff et al. (2010), Stöber et al. (2013), Gijbels et al. (2015a), Spanhel and Kurz (2017), and Section 4.7.

Example 2.2. *The density of a simplified R-vine copula corresponding to the tree sequence in Figure 2.1 is*

$$\begin{aligned} c(u_1, \dots, u_5) = {} & c_{1,2}(u_1, u_2) \times c_{1,3}(u_1, u_3) \times c_{3,4}(u_3, u_4) \times c_{3,5}(u_3, u_5) \\ & \times c_{2,3;1}(u_{2|1}, u_{3|1}) \times c_{1,4;3}(u_{1|3}, u_{4|3}) \times c_{1,5;3}(u_{1|3}, u_{5|3}) \\ & \times c_{2,4;1,3}(u_{2|1,3}, u_{4|1,3}) \times c_{4,5;1,3}(u_{4|1,3}, u_{5|1,3}) \\ & \times c_{2,5;1,3,4}(u_{2|1,3,4}, u_{5|1,3,4}), \end{aligned}$$

where we used the abbreviation $u_{j_e|D_e} = G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e})$.

2.2.4 Recursive computation of conditional distributions

Vine copula densities involve conditional distributions $G_{j_e|D_e}$. We can express them in terms of conditional distributions corresponding to the pair-copulas in the model: Let $\mathcal{B} := \{c_{j_e,k_e;D_e} | e \in E_m, 1 \leq m \leq d-1\}$ be the set of copula densities associated with the edges in $\mathcal{V}$. Further, let $\ell_e \in D_e$ be another index such that $c_{j_e,\ell_e;D_e \setminus \ell_e} \in \mathcal{B}$ and define $D_e' := D_e \setminus \ell_e$. Then, we can write

$$G_{j_e|D_e}(u_{j_e}|\boldsymbol{u}_{D_e}) = h_{j_e|\ell_e;D_e'}\{G_{j_e|D_e'}(u_{j_e}|\boldsymbol{u}_{D_e'}) \,\big|\, G_{\ell_e|D_e'}(u_{\ell_e}|\boldsymbol{u}_{D_e'})\}, \tag{2.5}$$

where the *h-function* is defined as

$$h_{j_e|\ell_e;D_e'}(u|v) = \int_0^u c_{j_e,\ell_e;D_e'}(s, v)ds, \qquad \text{for } (u, v) \in [0, 1]^2. \tag{2.6}$$

The arguments $G_{j_e|D_e'}(u_{j_e}|\boldsymbol{u}_{D_e'})$ and $G_{\ell_e|D_e'}(u_{\ell_e}|\boldsymbol{u}_{D_e'})$ of the h-function in (2.5) can be rewritten in the same manner. In each step of this recursion the conditioning set D_e is reduced by one element. By construction, the copula density on the right hand side of (2.6) always belongs to the set $\mathcal{B}$. Eventually, this allows us to write any of the conditional distributions $G_{j_e|D_e}$ occurring in (2.4) as a recursion over h-functions that are directly linked to the pair-copula densities.

Example 2.3. *Consider a simplified vine copula corresponding to the R-vine tree sequence given in Figure 2.1. We have*

$$G_{3|1,2}(u_3|u_1,u_2) = h_{3|2;1}\big\{h_{3|1}(u_3|u_1)\big|h_{2|1}(u_2|u_1)\big\},$$

where

$$\begin{aligned}
h_{3|1}(u_3|u_1) &= \int_0^{u_3} c_{1,3}(u_1,s)ds,\\
h_{2|1}(u_2|u_1) &= \int_0^{u_2} c_{1,2}(u_1,s)ds,\\
h_{3|2;1}(u_{3|1}|u_{2|1}) &= \int_0^{u_{3|1}} c_{2,3;1}(u_{2|1},s)ds.
\end{aligned}$$

Altogether, we can express any simplified vine copula density in terms of bivariate copula densities and corresponding h-functions.

2.3 Stochastic convergence

Many results in the following chapters will make statements about the convergence of random sequences. In the following we shall recall the definitions of various modes of stochastic convergence, some basic properties, and introduce related notation. For further details and proofs, we refer to van der Vaart (1998, Chapters 2 and 18) on which much of the material in this section is based. In general, we shall not discuss measurability issues and simply assume that all quantities are sufficiently measurable.

2.3.1 Modes of stochastic convergence

Definition 2.1 (Convergence in distribution)**.** *A sequence of random vectors $\boldsymbol{X}_n \in \mathbb{R}^d$ is said to converge in distribution to some random vector $\boldsymbol{X}$ (denoted as $\boldsymbol{X}_n \to_d \boldsymbol{X}$), if*

$$\Pr(\boldsymbol{X}_n \leq \boldsymbol{x}) \to \Pr(\boldsymbol{X} \leq \boldsymbol{x}), \quad \textit{for all } \boldsymbol{x} \in \mathbb{R}^d, \quad \textit{as } n \to \infty.$$

If the limiting random variable $\boldsymbol{X}$ has distribution D, we also write $\boldsymbol{X}_n \to_d D$, e.g., $\boldsymbol{X}_n \to_d \mathcal{N}(0, I_d)$ for convergence to a standard normal limit. Convergence in distribution is sometimes called *weak convergence*, although we shall reserve that name for its generalization from Euclidean to metric spaces.

Definition 2.2 (Metric)**.** *A* metric *or* distance *defined on some set* $\mathcal{M}$ *is a map* $d\colon \mathcal{M} \times \mathcal{M} \mapsto (0, \infty)$ *satisfying the following properties:*

(i) $d(x, y) = d(y, x)$ *for all* $x, y \in \mathcal{M}$ *(symmetry),*

(ii) $d(x, y) \leq d(x, z) + d(z, y)$ *for all* $x, y, z \in \mathcal{M}$ *(triangle inequality),*

(iii) $d(x, y) = 0$ *if and only if* $x = y$ *(identity of indiscernibles).*

A function d that satisfied (i), (ii), and $d(x, y) = 0$ for $x = y$ is called *semi-metric* or *semi-distance*. A set $\mathcal{M}$ equipped with a (semi-)metric d is called (semi-) metric space. Convergence in a metric space is denoted by $x_n \to x$ and defined as $d(x_n, x) \to 0$. Generally, any norm $\|\cdot\|$ induces a metric via $d(x, y) = \|x - y\|$. Hence, any normed space is automatically a metric space.

An equivalent characterization of (Euclidean) convergence in distribution arises from Portmanteau's lemma (van der Vaart, 1998, Lemma 2.2): $\boldsymbol{X}_n \to_d \boldsymbol{X}$ if and only if $\mathrm{E}\{f(\boldsymbol{X}_n)\} \to \mathrm{E}\{f(\boldsymbol{X})\}$ for all bounded, Lipschitz functions $f\colon \mathbb{R}^d \mapsto \mathbb{R}$. A function f defined on a metric space $(\mathcal{M}, d)$ is called Lipschitz if there is a constant $L < \infty$ such that for all $x, y \in \mathcal{M}$,

$$|f(x) - f(y)| \leq Ld(x, y).$$

This alternative characterization extends naturally to general metric spaces.

Definition 2.3 (Weak convergence)**.** *Let* $(\mathcal{M}, d)$ *be a metric space. We say that a sequence of random elements* $X_n \in \mathcal{M}$ *converges weakly to some limit* X *as* $n \to \infty$ *(denoted as* $X_n \rightsquigarrow X$*), if*

$$\mathrm{E}\big\{f(X_n)\big\} \to \mathrm{E}\big\{f(X)\big\},$$

as $n \to \infty$ *for all bounded, Lipschitz functions* $f\colon \mathcal{M} \mapsto \mathbb{R}$.

Weak convergence and convergence in distribution play a central role in statistics. Statistical estimators are computed from random samples and, hence, are random quantities themselves. The exact finite sample distribution of an estimator is often unknown or overly complex. But if we can show that the estimator converges weakly, we can approximate its distribution by the distribution of the limit. The latter is usually much easier to derive and compute. For example, this technique is widely used to construct confidence intervals around an estimate.

Two further modes of stochastic convergence will come up in later chapters. We define them in the general setup of metric spaces, but emphasize again that Euclidean spaces are a special case.

Definition 2.4 (Convergence in probability)**.** *Let $(\mathcal{M}, d)$ be a metric space. We say that a sequence of random elements $X_n \in \mathcal{M}$ converges in probability to some limit X (denoted as $X_n \to_p X$) if for all $\epsilon > 0$,*

$$\Pr\big\{d(X_n, X) > \epsilon\big\} \to 0, \quad \textit{as } n \to \infty.$$

Definition 2.5 (Almost sure convergence)**.** *Let $(\mathcal{M}, d)$ be a metric space. We say that a sequence of random elements $X_n \in \mathcal{M}$ converges almost surely to some limit X (denoted as $X_n \to_{a.s.} X$) if*

$$\Pr\Big\{\lim_{n\to\infty} d(X_n, X) \to 0\Big\} = 1.$$

Another common name for almost sure convergence is *convergence with probability 1*.

2.3.2 Important properties

The modes of stochastic convergence have several interesting properties. The most important ones are summarized below and will be used without mention.

The first result establishes relationships between the different modes of convergence.

Lemma 2.1.

(i) $X_n \to_{a.s.} X$ implies $X_n \to_p X$.

(ii) $X_n \to_p X$ implies $X_n \rightsquigarrow X$.

(iii) $X_n \rightsquigarrow x$ for a constant x if and only if $X_n \to_p x$.

(iv) $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow y$ for a constant y implies $(X_n, Y_n) \rightsquigarrow (X, y)$.

Hence, almost sure convergence is stronger than convergence in probability which is stronger than convergence in distribution. The next result shows that convergence is preserved under continuous maps.

Lemma 2.2 (Continuous mappings)**.** *Let $\mathcal{M}$, $\mathcal{M}'$ are two metric spaces and $X_n, X \in \mathcal{M}$. Suppose that $g\colon \mathcal{M} \mapsto \mathcal{M}'$ is continuous.*

(i) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$.

(ii) If $X_n \to_p X$, then $g(X_n) \to_p g(X)$.

(iii) If $X_n \to_{a.s.} X$, then $g(X_n) \to_{a.s.} g(X)$.

We will use three further properties that are specific to Euclidean spaces. The first is often referred to as *Slutsky's lemma* and follows directly from Lemma 2.1 (iii) and (iv) and Lemma 2.2 (i).

Lemma 2.3 (Slutsky). *Let $\boldsymbol{X}_n, \boldsymbol{X}, \boldsymbol{Y}_n$ be (sequences) of random vectors. If $\boldsymbol{X}_n \to_d \boldsymbol{X}$ and $\boldsymbol{Y}_n \to \boldsymbol{c}$ for a constant $\boldsymbol{c}$, then*

(i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \to_d \boldsymbol{X} + \boldsymbol{c}$,

(ii) $\boldsymbol{X}_n \boldsymbol{Y}_n \to_d \boldsymbol{X} \cdot \boldsymbol{c}$, where $\cdot$ denotes the component-wise product.

In particular, $\boldsymbol{Y}_n \to_p \boldsymbol{0}$ implies $\boldsymbol{X}_n + \boldsymbol{Y}_n \to_d \boldsymbol{X}$. Hence, we can neglect all terms that vanish in probability when we study the asymptotic distribution of a random vector.

To establish convergence in probability, we need to consider limits of probabilities. It is often more convenient to work with expectations or variances. The next property connects convergence in probability to L_p-convergence, another mode of stochastic convergence. The p-norm is defined as $\|\boldsymbol{x}\|_p = (\sum_k |x_k|^p)^{1/p}$. The limiting case $p = \infty$ is also known as the sup-norm, i.e., $\|\boldsymbol{x}\|_\infty = \sup_k |x_k|$.

Lemma 2.4. *If $\lim_{n\to\infty} \mathrm{E}\{\|\boldsymbol{X}_n - \boldsymbol{X}\|_p^p\} \to 0$ for some $p \geq 1$, then $\boldsymbol{X}_n \to_p \boldsymbol{X}$.*

The proof is a simple application of Markov's inequality. An important special case is $p = 2$ for which Lemma 2.4 implies that any sequence with vanishing variance converges in probability.

The last property is called *delta method* and often helpful to derive asymptotic distributions.

Lemma 2.5 (Delta method). *Let $\phi\colon \mathbb{R}^d \mapsto \mathbb{R}^k$ be a map that is differentiable at $\boldsymbol{\theta} \in \mathbb{R}^d$ with derivative matrix $\phi'_{\boldsymbol{\theta}} \in \mathbb{R}^{k\times d}$. If for a sequence of random vectors $\boldsymbol{X}_n \in \mathbb{R}^d$ and some deterministic sequence $r_n \to \infty$ it holds $r_n(\boldsymbol{X}_n - \boldsymbol{\theta}) \to_d \boldsymbol{X}$, then $r_n\{\phi(\boldsymbol{X}_n) - \phi(\boldsymbol{\theta})\} \to_d \phi'_{\boldsymbol{\theta}}\boldsymbol{X}$.*

In particular, if $r_n(\boldsymbol{X}_n - \boldsymbol{\theta})$ is asymptotically normal, then so is $r_n\{\phi(\boldsymbol{X}_n) - \phi(\boldsymbol{\theta})\}$: for any $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$,

$$r_n(\boldsymbol{X}_n - \boldsymbol{\theta}) \to_d \mathcal{N}(\boldsymbol{\mu}, \Sigma) \quad \text{implies} \quad r_n\{\phi(\boldsymbol{X}_n) - \phi(\boldsymbol{\theta})\} \to_d \mathcal{N}\big(\phi'_{\boldsymbol{\theta}}\boldsymbol{\mu}, \phi'_{\boldsymbol{\theta}}\Sigma(\phi'_{\boldsymbol{\theta}})^\top\big).$$

2.3.3 Asymptotic notation

Statements about stochastic convergence and corresponding rates of convergence are essential in many proofs in this thesis. It is convenient to introduce additional notation to shorten such arguments. In particular, it is common to adapt the usual 'big-O/little-o' notation to stochastic modes of convergence. This will involve two additional properties of random variables.

Definition 2.6 (Tight)**.** *A random element X in a metric space $(\mathcal{M}, d)$ is called tight if for every $\epsilon > 0$, there is a compact set $K \subseteq \mathcal{M}$ such that* $\Pr(X \in K) \geq 1 - \epsilon$.

Definition 2.7 (Almost surely bounded)**.** *A random element X in a metric space $(\mathcal{M}, d)$ is called almost surely bounded if there is a compact set $K \subseteq \mathcal{M}$ such that* $\Pr(X \in K) = 1$.

Remark 2.1. *A Euclidean random vector $\boldsymbol{X}$*

(i) is tight if $\Pr(|\boldsymbol{X}| = \infty) = 0$,

(ii) is almost surely bounded if there is a constant $A < \infty$ such that $\Pr(\|X\|_\infty < A) = 1$.

Definition 2.8 (Stochastic big-O/little-o notation)**.** *Let $X, X_n \in \mathcal{M}$, $n \in \mathbb{N}$ be random elements in a metric space $(\mathcal{M}, d)$ and R_n be a sequence of random variables with* $\limsup_{n\to\infty} \Pr(R_n = 0) = 0$*. We write*

(i) $X_n = O_p(R_n)$ if there is a tight X such that $X_n/R_n \to_p X$,

(ii) $X_n = o_p(R_n)$ if $X_n/R_n \to_p 0$,

(iii) $X_n = O_{a.s}(R_n)$ if there is an almost surely bounded X such that $X_n/R_n \to_{a.s} X$,

(iv) $X_n = o_{a.s.}(R_n)$ if $X_n/R_n \to_{a.s.} 0$.

The 'convergence rate' R_n can (and mostly will) be deterministic. If X_n is deterministic as well, the stochastic big-O/little-o notation simplifies to its deterministic version.

For deterministic rates, we use an additional convention: we write $r_n \sim a_n$ if $|r_n/a_n| \to c \in (0, \infty)$.

2.4 Empirical processes

There is a joke about statisticians taking averages all day, and there's a grain of truth in every joke. Most statistical estimators can be represented (at least asymptotically) as sample averages. A general framework to analyze the behavior of such averages is the theory of *empirical processes* and will be used heavily in Chapter 5. This theory differs from classical asymptotics in that the random elements studied are random functions (called processes) as opposed to random vectors. The following sections introduces the core results and objects in the study of empirical processes, but leaves much more to say. For a more thorough treatment we refer to the excellent textbooks of van der Vaart and Wellner (1996) and Kosorok (2007).

2.4.1 Stochastic processes and the empirical measure

A *stochastic process* Z is a collection of random variables $\{Z_t, t \in T\}$ index by some set T. The term 'process' comes from the special case where the index t represents time. But the index set T does not need to be Euclidean. In the context of empirical processes, T is often a set of functions.

Coming back to sample averages, consider a random quantity of the form

$$\mathbb{P}_n g = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i),$$

where $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in \mathbb{R}^d$ is a sequence of random vectors (observations), and g is some function. The object $\mathbb{P}_n = n^{-1}\sum_{i=1}^{n} \delta_{\boldsymbol{X}_i}$ is the *empirical measure*, the discrete probability measure that assigns equal probabilities to all observations. In analogy, if P is the true probability measure, we write

$$Pg = \mathrm{E}\{g(\boldsymbol{X})\} = \int g(\boldsymbol{x})dP(\boldsymbol{x}),$$

for the expectation with respect to P. In the following, let $\mathcal{G}$ be class of functions. Then $\{\mathbb{P}_n g, g \in \mathcal{G}\}$ is a stochastic process indexed by $\mathcal{G}$.

2.4.2 Empirical distribution functions

The empirical distribution function is defined as

$$F_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\boldsymbol{X}_i \leq \boldsymbol{x}), \quad x \in \mathbb{R}^d.$$

Statements on the stochastic convergence of F_n are among the most classical results in statistics. The law of large numbers guarantees that $F_n(\boldsymbol{x}) \to_{a.s.} F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$, where F is the cumulative distribution function of $\boldsymbol{X}$. Glivenko (1933) and Cantelli (1933) showed that this can be strengthened to uniform convergence, i.e.,

$$\sup_{\boldsymbol{x}} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})| \to_{a.s.} 0.$$

Here, we view F_n as a stochastic process $\{\mathbb{P}_n \mathbb{1}(\cdot \leq \boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\}$ indexed by the Euclidean parameter $\boldsymbol{x} \in \mathbb{R}^d$. However, by defining $\mathcal{G} = \{\mathbb{1}(\cdot \leq \boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\}$, we can also see it as an empirical process $\{\mathbb{P}_n g, g \in \mathcal{G}\}$ indexed by a family of indicator functions. Then the last display becomes

$$\sup_{g \in \mathcal{G}} |\mathbb{P}_n g - Pg| \to_{a.s.} 0.$$

The latter viewpoint is more general and therefore more common in the modern theory of empirical processes.

From the (multivariate) central limit theorem, we know that

$$\sqrt{n}\big\{\big(F_n(\boldsymbol{x}^{(1)}),\dots,F_n(\boldsymbol{x}^{(k)})\big)-\big(F(\boldsymbol{x}^{(1)}),\dots,F(\boldsymbol{x}^{(k)})\big\}$$

is asymptotically normal for any finite collection $\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(k)}\in\mathbb{R}^d$. A different question is whether $\sqrt{n}(F_n-F)$ converges weakly *as a process* in $\boldsymbol{x}$. Such a 'uniform central limit theorem' was established much later by Donsker (1952).

2.4.3 Empirical processes indexed by functions

We return to the general case, where we study $\{\mathbb{P}_n g, g\in\mathcal{G}\}$ as a stochastic process indexed by a class of functions $\mathcal{G}$. A central question of empirical process theory is which classes of functions allow (uniform) almost sure and weak convergence. As a tribute to the authors of the fundamental results for empirical distributions, such classes are called *Glivenko-Cantelli* and *Donsker*, respectively.

Definition 2.9 (Glivenko-Cantelli classes)**.** *A class of functions $\mathcal{G}$ is called P-Glivenko-Cantelli if*

$$\sup_{g\in\mathcal{G}}|\mathbb{P}_n g-Pg|\to_{a.s.}0.$$

The reference to the probability measure P is necessary due to the fact that different probability measures induce different Glivenko-Cantelli classes.

This is also true for Donsker classes, for which we introduce additional notation. We call

$$\mathbb{G}_n=\sqrt{n}(\mathbb{P}_n-P)$$

the *empirical process* at g and define $\ell^\infty(\mathcal{G})$ as the space of bounded maps $f\colon\mathcal{G}\to\mathbb{R}$.

Definition 2.10 (Donsker classes)**.** *A class of functions $\mathcal{G}$ is called P-Donsker if $\mathbb{G}_n$ converges weakly to a tight element $\mathbb{G}$ in $\ell^\infty(\mathcal{G})$.*

If $\mathcal{G}$ is P-Donsker, the limiting process $\mathbb{G}$ is Gaussian with zero mean and covariance $\operatorname{Cov}(\mathbb{G}f,\mathbb{G}g)=Pgf-PgPf$.

2.4.4 Glivenko-Cantelli and Donsker theorems based on bracketing

If a class of functions is Glivenko-Cantelli or Donsker depends on its size or complexity. A convenient measure for the size of classes of real-valued functions is the bracketing number. A *bracket* $[f,g]$ is defined the set of all functions h such that $f\le h\le g$ (pointwise). It is called an *ϵ-bracket* if additionally $\|f-g\|<\epsilon$ for some norm $\|\cdot\|$.

Definition 2.11 (Bracketing numbers)**.** *The bracketing number $N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|)$ with respect to some norm $\|\cdot\|$ is defined as the minimum number of ϵ-brackets needed to cover $\mathcal{G}$.*

Theorem 2.1 (Glivenko-Cantelli theorem)**.** *Every class $\mathcal{G}$ such that*

$$N_{[]}\{\epsilon, \mathcal{G}, L_1(P)\} < \infty, \quad \textit{for every } \epsilon > 0,$$

is P-Glivenko-Cantelli.

The bracketing numbers in Theorem 2.1 are allowed to diverge as $\epsilon \to 0$. The speed of divergence turns out to be key for determining whether a class is also P-Donsker.

Theorem 2.2 (Donsker theorem)**.** *Every class $\mathcal{G}$ such that*

$$\int_0^1 \sqrt{\ln N_{[]}\{\epsilon, \mathcal{G}, L_2(P)\}} d\epsilon < \infty,$$

is P-Donsker.

Note that the bracketing integral in Theorem 2.2 implies the finiteness of bracketing numbers required by Theorem 2.1. In general, any class that is P-Donsker is also P-Glivenko-Cantelli.

For later reference, we also introduce an alternative measure of the size of $\mathcal{G}$.

Definition 2.12 (Covering numbers)**.** *The covering number $N(\epsilon, \mathcal{G}, \|\cdot\|)$ is defined as the minimum number of ϵ-balls needed to cover $\mathcal{G}$.*

We can state Glivenko-Cantelli and Donsker theorems similar to the ones above for the covering numbers. But we will only use the convenient property that $N(\epsilon/2, \mathcal{G}, \|\cdot\|) \leq N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|)$ for any class $\mathcal{G}$ and norm $\|\cdot\|$.

2.4.5 Weak convergence in the space of bounded functions

Definition 2.10 refers specifically to weak convergence in the space $\ell^\infty(\mathcal{G})$. Weak convergence in this space has equivalent characterizations that are often easier to deal with. One such characterization that can be used to prove Theorem 2.2 will be introduced in the following.

Definition 2.13 (Totally bounded)**.** *We say that a set T is totally bounded by a semi-metric ρ if for every $\epsilon > 0$ there exist $t_1, \ldots, t_k \in T$, $k < \infty$, such that for all $t \in T$ there is $i \in \{1, \ldots, k\}$ such that $\rho(t, t_i) < \epsilon$.*

A common semi-metric ρ is induced by the $L_2(P)$ norm, i.e.,

$$d(f,g) = \|f-g\|_{P,2} = \{P(f-g)^2\}^{1/2}.$$

Then a class of functions $\mathcal{G}$ is totally bounded by ρ if $\sup_{g\in\mathcal{G}} Pg^2 < \infty$.

Let $Z_n = \{Z_{n,t}, t \in T\}$, $n \geq 1$, be a sequence of stochastic processes indexed by T and recall that the limiting process in Definition 2.10 is required to be tight (see Definition 2.6). In that case we have the following result (Kosorok, 2007, Theorem 2.1):

Lemma 2.6. *Z_n converges weakly to a tight process Z in $\ell^\infty(T)$, if and only if:*

(i) For all $t_1, \dots, t_k \in T$, $k < \infty$,

$$(Z_{n,t_1}, \dots, Z_{n,t_k}) \rightarrow_d (Z_{t_1}, \dots, Z_{t_k}),$$

(ii) there exists a semi-metric ρ for which T is totally bounded and

$$\lim_{\delta\searrow 0} \limsup_{n\to\infty} \Pr\left\{ \sup_{t_1,t_2\in T\colon\, \rho(t_1,t_2)<\delta} |Z_{n,t_1} - Z_{n,t_2}| > \epsilon \right\} \to 0, \ \textit{for all } \epsilon > 0.$$

The first property corresponds to convergence of all finite dimensional distributions. The second condition is called *ρ-equicontinuity*. As the name suggests, it is a condition on the smoothness of the sample paths of Z_n: Z_{n,t_1} and Z_{n,t_2} are close to each other whenever t_1 and t_2 are sufficiently close. In summary, Z_n converges weakly in $\ell^\infty(T)$ if all finite-dimensional distributions converge and Z_n is ρ-equicontinuous.

2.4.6 The functional delta method

In Euclidean spaces, the delta method (Lemma 2.5) is a powerful to derive the limiting distribution of functions of random vectors. Its only condition is differentiability of the function. There is a generalization of the delta method to normed spaces, but this requires an extended definition of a derivative. Recall that a linear map ψ defined on elements in a normed space $\mathbb{D}$ is a map with the property $\psi(ax+by) = a\phi(x) + b\phi(y)$ for all $a, b \in \mathbb{R}$, $x, y \in \mathbb{D}$.

Definition 2.14 (Hadamard derivative)**.** *Suppose* $\mathbb{D}$, $\mathbb{E}$ *are normed spaces. A map* $\phi\colon \mathbb{D} \to \mathbb{E}$ *is called* Hadamard differentiable *at* $\theta \in \mathbb{D}$*, if there exists a linear map* $\phi'_\theta\colon \mathbb{D} \to \mathbb{E}$ *such that*

$$\frac{\phi(\theta + r_n h_n) - \phi(\theta)}{t_n} \to \phi'_\theta(h),$$

for all real-valued sequences $r_n \to 0$, $h_n \to h \in \mathbb{D}$ *with* $h_n \in \mathbb{D}$*, and* $\theta + r_n h_n \in \mathbb{D}$ *for* n *sufficiently large.*

The function h in the above definition is referred to as the the *direction* of the derivative. Hadamard differentiability turns out to be exactly what we need to extend the delta method to normed spaces.

Lemma 2.7 (Functional delta method)**.** *Let* $\phi\colon \mathbb{D} \to \mathbb{E}$ *be a map with Hadamard derivative* $\phi'_\theta\colon \mathbb{D} \to \mathbb{E}$ *for some* $\theta \in \mathbb{D}$*. If for a sequence of random elements* $X_n \in \mathbb{D}$*, a deterministic sequence* $r_n \to \infty$*, and* $X \in \mathbb{D}$*, it holds* $r_n(X_n - \theta) \rightsquigarrow X$*, then* $r_n\{\phi(X_n) - \phi(\theta)\} \rightsquigarrow \phi'_\theta(X)$*.*

Both Hadamard differentiability and the delta method can be refined for maps ϕ'_θ that do not exist on the whole $\mathbb{D}$, see van der Vaart (1998, Section 20.2) for details.

3 Nonparametric estimators of simplified vine copula densities

3.1 Introduction

Each pair-copula in a simplified vine copula model can be specified as a unique bivariate copula function. Thus, simplified vine copulas give rise to very flexible models which are often found to be superior to other multivariate copula models (Aas et al., 2009, Fischer et al., 2009). The models are also easily tractable because pair-copulas can be estimated sequentially. Parametric models for the pair-copulas are most common, but bear the risk of misspecification. In particular, most parametric families only allow for highly symmetric and monotone relationships between variables.

To remedy this issue, several nonparametric approaches have been proposed: penalized Bernstein polynomials and B-splines (Kauermann and Schellhase, 2014), kernel estimators (Nagler, 2014), and a non-penalized Bernstein estimator (Scheffer and Weiß, 2017). A related contribution of introduces the empirical pair-copula as an extension of the empirical copula (Hobæk Haff and Segers, 2015), but does not aim at estimation of the vine copula density which is the focus of this chapter.

From a practitioner's point of view, the question arises: which method should I choose for a given data set? This question is difficult to answer theoretically because asymptotic approximations of nonparametric vine copula density estimators are prohibitively unwieldy, see Propositions 4.1 and 4.5. In the following, we conduct an extensive simulation study to provide some guidance nevertheless. All estimation methods will be compared under several specifications of strength and type of dependence, sample size, and dimension, thereby covering a large range of practical scenarios.

Although our primary goal is to survey and compare existing methods, we extend the estimators proposed by Kauermann and Schellhase (2014), Scheffer and Weiß (2017), Nagler (2014) in several ways:

- The Bernstein and B-spline estimators of Kauermann and Schellhase (2014) and Scheffer and Weiß (2017) are extended to allow for general R-vine structures (opposed to just D- and/or C-vine structures).
- Besides linear B-splines as in Kauermann and Schellhase (2014), we also consider quadratic B-splines.

- All pair-copula estimators can be combined with structure selection algorithms using both Kendall's τ and a corrected AIC as target criterion.

The remainder of this chapter is organized as follows. Section 3.2 presents and extends several existing nonparametric methods for pair-copula estimation, introduces a step-wise estimation algorithm for the vine copula model, and discusses approaches for model selection. We describe the design of our simulation study in Section 3.3 and summarize the results in Section 3.4. In Section 3.5, a real data set is used to illustrate the estimators' behavior and demonstrate the necessity for nonparametric estimators. Section 3.6 offers conclusions.

3.2 Implementation

3.2.1 Nonparametric estimators for bivariate copula densities

We now give an overview of nonparametric estimators of bivariate copula densities. The classical approach to density estimation is to assume a parametric model and estimate its parameters by maximum likelihood. There is a large variety of bivariate parametric copula models. Special classes are the elliptical copulas (including the Gaussian and Student t families), and the Archimedean class (including the Clayton, Frank and Gumbel families); for more see (Joe, 2014b). However, parametric models notoriously lack flexibility and bear the risk of misspecification. Nonparametric density estimators are designed to remedy these issues. In the context of copula densities, these estimators have to take the bounded support into account.

In the following we summarize the state-of-the-art of the major strands of nonparametric copula density estimation. For simplicity, we only consider the bivariate case. We assume throughout that we are given n observations $(U_1^{(i)}, U_2^{(i)})$, $i = 1, \ldots, n$, from a copula density c that we want to estimate.

Empirical Bernstein copula

A classical tool in function approximation are Bernstein polynomials (Lorentz, 1953). The normalized Bernstein polynomial of degree K is defined as

$$B_{Kk}(u) = (K+1)\binom{K}{k} u^k (1-u)^{K-k}, \quad \text{for } k = 0, \ldots, K.$$

The collection of all Bernstein polynomials form a basis of the space of all square-integrable functions on $[0, 1]$. A natural idea is to approximate an arbitrary function by a linear combination of a finite number of basis functions. Based on this idea, Sancetta and Satchell (2004) defined the Bernstein copula density as an approximation of the true copula density. It can be expressed as

$$\widetilde{c}(u_1, u_2) = \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} B_{Kk_1}(u_1) B_{Kk_2}(u_2) v_{k_1,k_2},$$

where

$$v_{k_1,k_2} = \int_{k_1/\bar{K}}^{(k_1+1)/\bar{K}} \int_{k_2/\bar{K}}^{(k_2+1)/\bar{K}} c(u_1, u_2) du_1 du_2.$$

and $\bar{K} = (K+1)$. Note that the coefficient v_{k_1,k_2} describes the probability that $(U_1^{(i)}, U_2^{(i)})$ is contained in the cell $[k_1/\bar{K}, (k_1+1)/\bar{K}] \times [k_2/\bar{K}, (k_2+1)/\bar{K}]$. The empirical copula density estimator is defined by $\widetilde{c}(u_1, u_2)$, where the v_{k_1,k_2} are replaced by empirical frequencies obtained from a contingency table:

$$\widehat{c}(u_1, u_2) = \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} B_{Kk_1}(u_1) B_{Kk_2}(u_2) \widehat{v}_{k_1,k_2},$$

where

$$\widehat{v}_{k_1,k_2} = \frac{1}{n} \times \#\big\{(U_1^{(i)}, U_2^{(i)}) \in [k_1/\bar{K}, (k_1+1)/\bar{K}] \times [k_2/\bar{K}, (k_2+1)/\bar{K}]\big\},$$

which is the maximum-likelihood estimator for v_{k_1,k_2}.

The Bernstein copula density estimator was used in the context of vine copulas by Scheffer and Weiß (2017). As the marginal distributions of the Bernstein copula density do not need to be uniform, the authors calculate an approximation to the contingency table by solving a quadratic program, imposing constraints for uniform marginal distributions. The smoothing parameter for the Bernstein copula density estimator is K, the number of knots. Rose (2015) proposed selection rules for K that adapt to the sample size and strength of dependence. Our implementation is available in the `kdecopula` R package (Nagler, 2018c), and uses the rule

$$K^{opt} = \lfloor n^{1/3} \exp(|\widehat{\rho}|^{1/n})(|\widehat{\rho}| + 0.1) \rfloor,$$

where $\widehat{\rho}$ is the empirical Spearman's ρ.

Penalized Bernstein polynomials and B-splines

For fixed K, the Bernstein copula density estimator is a parametric model with $(K+1)^2$ parameters. As any parametric model with many parameters, it is prone to overfitting. To gain control of the smoothness of the fit, Kauermann and Schellhase (2014) proposed a penalized likelihood approach.

Viewing the Bernstein copula density as a parametric model with parameter vector $\boldsymbol{v} = (v_{00}, \dots, v_{0K}, \dots, v_{KK})$, i.e.,

$$\widetilde{c}(u_1, u_2; \boldsymbol{v}) = \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} B_{Kk_1}(u_1) B_{Kk_2}(u_2) v_{k_1,k_2}, \tag{3.1}$$

we can estimate the parameters by maximizing the log-likelihood,

$$\ell(\boldsymbol{v}) = \log \sum_{i=1}^{n} \widetilde{c}\big(U_1^{(i)}, U_2^{(i)}; \boldsymbol{v}\big). \tag{3.2}$$

Since each of the normalized Bernstein polynomials is a density, the weighted sum of normalized Bernstein polynomials is a density, if we ensure that

$$\begin{aligned} \sum_{k_1=0}^{K} \sum_{k_2=0}^{K} v_{k_1,k_2} &= 1, \\ v_{k_1,k_2} &\geq 0. \end{aligned} \tag{3.3}$$

However, we will need more stringent constraints to enforce uniform marginal distributions: for Bernstein polynomials, $\int \tilde{c}(u_1, u_2)\, \mathrm{d}u_1 \equiv 1$ holds if the marginal coefficients fulfill

$$v_{k_1 \cdot} = \sum_{k_2=0}^{K} v_{k_1,k_2} = 1/(K+1), \quad \text{for all } k_1 = 0, \dots, K,$$

and similarly for $\int \tilde{c}(u_1, u_2)\, \mathrm{d}u_2 \equiv 1$. These constraints can be reformulated in matrix notation yielding

$$A_K^T \boldsymbol{v} = \mathbf{1}/(K+1), \tag{3.4}$$

where A_K sums up the elements of v_{k_1,k_2} column-wise (i.e., over k_2) and row-wise (i.e. over k_1), i.e. $A_K^T = (I_K \otimes \mathbf{1}_K^T, \mathbf{1}_K^T \otimes I_K)$, where $\mathbf{1}_K$ is the column vector of dimension K with elements 1, I_K is the K dimensional identity matrix, and $\otimes$ denotes the tensor product.

The log-likelihood (3.2) can be maximized under the constraints (3.3) and (3.4), using quadratic programming (e.g., with the `quadprog` `R` package Weingessel, 2013). But since this is a parametric model with many parameters, the fitted copula density may be wiggly, see e.g., (Wahba, 1990). This issue can be resolved by imposing an appropriate penalty on the basis coefficients. We postulate that the integrated squared second order derivatives are small and formulate the penalty as

$$\int \left\{ \left(\frac{\partial^2 \widetilde{c}(u_1, u_2; \boldsymbol{v})}{(\partial u_1)^2} \right)^2 + \left(\frac{\partial^2 \widetilde{c}(u_1, u_2; \boldsymbol{v})}{(\partial u_2)^2} \right)^2 \right\} du_1 du_2,$$

(see also, Wood, 2006). This can be written as a quadratic form of a penalty matrix $\mathbf{P}$ (see, Kauermann and Schellhase, 2014). The corresponding penalized log-likelihood is defined as

$$\ell^p(\boldsymbol{v}, \lambda) = \ell(\boldsymbol{v}) - \frac{1}{2} \lambda \boldsymbol{v}^T \mathbf{P} \boldsymbol{v}, \tag{3.5}$$

which is again maximized with respect to the constraints (3.3) and (3.4).

The penalty parameter λ can be selected in a data-driven manner. In Section 2.5 of Kauermann and Schellhase (2014), the authors propose a method that formulates the penalized likelihood approach as linear mixed model and comprehend the penalty as normal prior imposed on the coefficient vector.

One can further use B-spline basis functions instead of Bernstein polynomials. Kauermann and Schellhase (2014) replace each B_{Kk} in (3.1) with a B-spline, located at equidistant knots $\kappa_k = k/K$ with $k = 0, \dots, K$, normalized so that it satisfies $\int_0^1 B_{Kk}(u)\, \mathrm{d}u = 1$ for $k = 0, \dots, K - 1 + q$. Kauermann and Schellhase (2014) only used normalized linear ($q = 1$) B-splines. To allow for more flexibility, we will also use normalized quadratic ($q = 2$) B-splines in our study.

In order to guarantee that $\widetilde{c}(u_1, u_2; \boldsymbol{v})$ is a bivariate copula density, we impose similar constraints as the ones for the Bernstein polynomials. The linear constraint (3.3) will be the same for B-splines, but the uniform margins condition (3.4) has to be adapted. The condition takes the form $A_K \boldsymbol{v} = \mathbf{1}$ with $A_K = \boldsymbol{B}_K(\kappa)$, choosing

$$\kappa = \begin{cases} \kappa_0, \dots, \kappa_K, & \text{for linear B-splines,} \\ 0, \frac{\kappa_1 - \kappa_0}{2} + \kappa_0, \frac{\kappa_2 - \kappa_1}{2} + \kappa_1, \dots, \frac{\kappa_{K+1} - \kappa_K}{2} + \kappa_K, 1, & \text{for quadratic B-splines.} \end{cases}$$

For the penalization, we work with a penalty on the m-th order differences of the spline coefficients $\boldsymbol{v}$, as suggested for B-spline smoothing by Eilers and Marx (1996), defining a penalty matrix $\mathbf{P^m}$, where we choose $m = q+1$. Further details of this smoothing concept can be found in Ruppert et al. (2003). In the following, we define the difference based penalty matrix $\mathbf{P^m}$ for the m-order differences through

$$\mathbf{P^m} := (\mathbf{1}_{K+q} \otimes L_m)^T (L_m \otimes \mathbf{1}_{K+q}). \tag{3.6}$$

Let $L_m \in \mathbb{R}^{K+q-m \times K+q}$ be a difference matrix of order m, e.g., for $q = 1$ we get $m = 2$ and

$$L_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{K-1 \times K+1}.$$

Then for B-splines, the penalized log-likelihood becomes

$$l^p(\boldsymbol{v}, \lambda) = l(\boldsymbol{v}) - \frac{1}{2} \lambda \boldsymbol{v}^T \mathbf{P^m} \boldsymbol{v}. \tag{3.7}$$

We recover the independence copula, if we set $\lambda = \infty$ in (3.5) or (3.7). The penalized Bernstein and B-splines estimators are implemented in the R package `penRvine` (Schellhase, 2016).

Kernel weighted local likelihood

Kernel estimators are well-established tools for nonparametric density estimation. Several kernel methods have been tailored to the problem of copula density estimation. Their main challenge is to avoid bias and consistency issues at the boundaries of the support. The earliest contribution is the mirror-reflection method (Gijbels and Mielniczuk, 1990). Later, the beta kernel density estimator of Chen (1999) was extended to the bivariate case by Charpentier et al. (2006).

The more recent contributions all focus on a transformation trick. Assume we want to estimate a copula density c given a random sample $\big(U_1^{(i)}, U_2^{(i)}\big), i = 1, \dots, n$. Let Φ be the standard normal distribution function and ϕ its density. Then the random vectors $(Z_1^{(i)}, Z_2^{(i)}) = \big(\Phi^{-1}(U_1^{(i)}), \Phi^{-1}(U_2^{(i)})\big)$ have normally distributed margins and are supported on $\mathbb{R}^2$. In this domain, kernel density estimators work very well and do not suffer from any boundary problems. By Sklar's Theorem for densities (2.2), the density g of $(Z_1^{(i)}, Z_2^{(i)})$ decomposes to

$$g(z_1, z_2) = c\big\{\Phi(z_1), \Phi(z_2)\big\}\phi(z_1)\phi(z_2), \qquad \text{for all } (z_1, z_2) \in \mathbb{R}. \tag{3.8}$$

By isolating c in (3.8) and the change of variables $u_j = \Phi(z_j), j = 1, 2$, we get

$$c(u_1, u_2) = \frac{g\big\{\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big\}}{\phi\big\{\Phi^{-1}(u_1)\big\}\phi\big\{\Phi^{-1}(u_2)\big\}}. \tag{3.9}$$

We can use any kernel estimator $\widehat{g}$ of g to define a kernel estimator of the copula density c :

$$\widehat{c}(u_1, u_2) = \frac{\widehat{g}\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big)}{\phi\big\{\Phi^{-1}(u_1)\big\}\phi\big\{\Phi^{-1}(u_2)\big\}}. \tag{3.10}$$

Estimators of this kind have an interesting feature. The denominator of (3.10) vanishes when u_1 or u_2 tend to zero or one. If the numerator vanishes at a slower rate, the estimated copula density explodes towards the corners of the unit square. This behavior is common for many popular parametric families, including the Gaussian, Student, Gumbel, and Clayton families. The transformation estimator (3.10) is well suited to resemble such shapes. However, its variance will also explode towards the corners and the estimator will be numerically unstable. To accommodate for this, we restrict the estimator to $[0.001, 0.999]^2$ and set estimates outside of this region to the closest properly defined estimate.

To estimate the density g, we consider the class of local polynomial likelihood estimators; see Loader (1999) for a general account and Geenens et al. (2017) in the context of bivariate copula estimation. Assume that the log-density $\log g(z_1, z_2)$ of the random vector $\boldsymbol{Z}^{(i)} = (Z_1^{(i)}, Z_2^{(i)})$ can be approximated locally by a polynomial of order q. For example, using a log-quadratic expansion, we get

$$\begin{aligned}\log g(z_1', z_2') &\approx P_{\boldsymbol{a}}(\boldsymbol{z} - \boldsymbol{z}') \qquad (3.11)\\ &= a_1 + a_2(z_1 - z_1') + a_3(z_2 - z_2') \qquad (3.12)\\ &\quad + a_4(z_1 - z_1')^2 + a_5(z_1 - z_1')(z_2 - z_2') + a_6(z_2 - z_2')^2 \qquad (3.13)\end{aligned}$$

for (z_1', z_2') in the neighborhood of $\boldsymbol{z} = (z_1, z_2)$. The polynomial coefficients $\boldsymbol{a}$ can be found by solving the weighted maximum likelihood problem

$$\begin{aligned}\widehat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}\in\mathbb{R}^6}\Bigg[&\sum_{i=1}^n \boldsymbol{K}\big\{B^{-1}(\boldsymbol{z} - \boldsymbol{Z}^{(i)})\big\}P_{\boldsymbol{a}}(\boldsymbol{z} - \boldsymbol{Z}^{(i)})\\ &- n\int_{\mathbb{R}^2}\boldsymbol{K}\big\{B^{-1}(\boldsymbol{z} - \boldsymbol{s})\big\}\exp\big\{P_{\boldsymbol{a}}(\boldsymbol{z} - \boldsymbol{s})\big\}d\boldsymbol{s}\Bigg],\end{aligned} \qquad (3.14)$$

where the kernel K is a symmetric probability density function, $\boldsymbol{K}(\boldsymbol{z}) = K(z_1)K(z_2)$ is the product kernel. The matrix $B \in \mathbb{R}^{2\times 2}$, $\det(B) > 0$, is called the bandwidth matrix and controls the degree of smoothing in each direction. The kernel K serves as a weight function that localizes the above optimization problem around $\boldsymbol{z}$.

The expression for the local likelihood in (3.14) can be motivated as follows (e.g., Loader, 1999, Chapter 5). We can formulate the non-local log-likelihood has

$$\sum_{i=1}^n \log\widehat{g}(\boldsymbol{Z}^{(i)}) - n\bigg(\int \widehat{g}(\boldsymbol{s})d\boldsymbol{s} - 1\bigg), \qquad (3.15)$$

where the second term is a penalty term that is zero whenever $\widehat{g}$ is a proper density. From there, we obtain the formula (3.14) in three steps. First, we we localize the sum and integral around $\boldsymbol{z}$ by inserting the kernel weights $\boldsymbol{K}\{B^{-1}(z - \boldsymbol{Z}^{(i)})\}$ (3.15). Second, the log-density estimate $\log\widehat{g}$ is approximated by a polynomial expansion (3.11). And finally, the term -1 is dropped, since it does not affect the solution of the maximization problem.

We obtain $\widehat{a}_1$ as an estimate for $\log g(z_1, z_2)$ and, consequently, $\exp(\widehat{a}_1)$ as an estimate for $g(z_1, z_2)$. An estimate of the copula density can be obtained by plugging this estimate in (3.9). For a detailed treatment of this estimator's asymptotic behavior we refer to Geenens et al. (2017). In general, the estimator does not yield a *bona fide* copula density because the margins may not be uniform. This issue can be resolved by normalizing the density estimate, for details see Nagler (2018c).

For application of the estimator, an appropriate choice of the bandwidth matrix is crucial. For the local constant approximation, a simple rule of thumb was shown to perform well in Nagler (2014). We use an extended version of this rule that also adjusts to the degree of the polynomial q:

$$B_{\text{rot}} = \nu_q n^{-1/(4q^*+2)}\widehat{\Sigma}_{\boldsymbol{Z}}^{1/2}, \quad q^* = 1 + \lfloor q/2\rfloor,$$

Algorithm 1 Sequential estimation of simplified vine copula densities

Input: (Pseudo-)Observations $(\widehat{U}_1^{(i)}, \dots, \widehat{U}_d^{(i)})$, $i = 1, \dots, n$, vine structure $(E_1, \dots, E_{d-1})$.

Output: Estimates of pair-copula densities and h-functions required to evaluate the vine copula density (3.16).

```
for $m = 1, \dots, d-1$:
    for all $e \in E_m$:
        (i) Based on $\big(\widehat{U}^{(i)}_{j_e|D_e}, \widehat{U}^{(i)}_{k_e|D_e}\big)_{i=1,\dots,n}$, obtain an estimate of the copula
            density $c_{j_e,k_e;D_e}$ which we denote as $\widehat{c}_{j_e,k_e;D_e}$.
        (ii) Derive corresponding estimates of the h-functions $\widehat{h}_{j_e|k_e;D_e}$, $\widehat{h}_{k_e|j_e;D_e}$
            by integration (eq. (2.6)).
        (iii) Set
            $\widehat{U}^{(i)}_{j_e|D_e \cup k_e} := \widehat{h}_{j_e|k_e;D_e}\big(\widehat{U}^{(i)}_{j_e|D_e} \big| \widehat{U}^{(i)}_{k_e|D_e}\big),$
            $\widehat{U}^{(i)}_{k_e|D_e \cup j_e} := \widehat{h}_{k_e|j_e;D_e}\big(\widehat{U}^{(i)}_{k_e|D_e} \big| \widehat{U}^{(i)}_{j_e|D_e}\big), \quad i = 1, \dots, n.$
    end for
end for
```

where $\widehat{\Sigma}_{\boldsymbol{Z}}$ is the empirical covariance matrix of $\boldsymbol{Z}^{(i)}$, $i = 1, \dots, n$, and $\nu_0 = 1.25$, $\nu_1 = \nu_2 = 5$. An implementation of the estimator is available in the R package `kdecopula` (Nagler, 2018c).

3.2.2 Step-wise estimation of simplified vine copula densities

We now turn to the question how a simplified vine copula density can be estimated. Most commonly, this is done in a sequential procedure introduced by Aas et al. (2009). The procedure is generic in the sense that it can be used with any consistent estimator for a bivariate copula density. It is summarized in Algorithm 1 and explained in more detail in the following.

From now on we use c to denote a d-dimensional vine copula density. Assume we have a random sample $\boldsymbol{U}^{(i)} = \big(U_1^{(i)}, \dots, U_d^{(i)}\big), i = 1, \dots, n$, from the target density c. Recall that this density can be written as

$$c(\boldsymbol{u}) = \prod_{m=1}^{d-1} \prod_{e \in E_m} c_{j_e,k_e;D_e}\big(C_{j_e|D_e}(u_{j_e} \mid \boldsymbol{u}_{D_e}), C_{k_e|D_e}(u_{k_e} \mid \boldsymbol{u}_{D_e})\big). \tag{3.16}$$

1. Based on two-dimensional subvectors of the observations $\boldsymbol{U}^{(i)}, i = 1, \dots, n$, we estimate all pair-copula densities and h-functions that correspond to edges of the first tree (the conditioning sets D_e are empty).

2. We use (2.6) to derive estimates of the h-functions, that is

$$\widehat{h}_{j_e|k_e}(u \mid v) := \int_0^u \widehat{c}_{j_e,k_e}(s,v)ds, \quad \text{for } (u,v) \in (0,1)^2 \text{ and } e = (j_e,k_e) \in E_1.$$

3. Any pair-copula density $c_{j_e,k_e;D_e}$ corresponding to an edge in the second tree is the density of a random vector $\big(G_{j_e|D_e}(X_{j_e}|X_{D_e}), G_{k_e|D_e}(X_{k_e}|X_{D_e})\big)$, $e \in E_2$. They are not observable, but we can use pseudo-observations such as

$$\widehat{U}^{(i)}_{j_e|D_e} := \widehat{G}_{j_e|D_e}\big(U^{(i)}_{j_e} \mid U^{(i)}_{D_e}\big) = \widehat{h}_{j_e|D_e}\big(U^{(i)}_{j_e} \mid U^{(i)}_{D_e}\big), \quad i = 1,\ldots,n, e \in E_2,$$

instead. This allows us to obtain estimates $\widehat{c}_{j_e,k_e;D_e}$, $\widehat{h}_{j_e|k_e;D_e}$, and $\widehat{h}_{k_e|j_e;D_e}$ for a conditioning set D_e with one element.

4. For estimation in the third tree, we need observations from random variables such as

$$U^{(i)}_{j_e|D_e} := G_{j_e|D_e}\big(X^{(i)}_{j_e} \mid \boldsymbol{X}^{(i)}_{D_e}\big), \quad i = 1,\ldots,n, e \in E_3. \tag{3.17}$$

Now the conditioning set D_e contains two elements. Recall from Chapter 2 that, by construction, we can find some edge $e' \in E_2$ such that $j_{e'} = j_e$ and $D_{e'} \cup k_{e'} = D_e$. Consequently, we can apply (2.5) and approximate (3.17) by the pseudo-observations

$$\begin{aligned}\widehat{U}^{(i)}_{j_e|D_e} = \widehat{U}^{(i)}_{j_{e'}|D_{e'} \cup k_{e'}} &:= \widehat{G}_{j_{e'}|D_{e'} \cup k_{e'}}\big(\widehat{U}^{(i)}_{j_{e'}} \mid \widehat{\boldsymbol{U}}^{(i)}_{D_{e'} \cup k_{e'}}\big) \\ &= \widehat{h}_{j_{e'}|k_{e'};D_{e'}}\big(\widehat{U}^{(i)}_{j_{e'}|D_{e'}} \mid \widehat{U}^{(i)}_{k_{e'}|D_{e'}}\big),\end{aligned}$$

where the last equality is again derived from (2.5).

5. For higher trees, proceed as in step 4.

3.2.3 Selection strategies for the vine structure

So far we assumed that the structure of the vine (i.e., the collection of edge sets $E_1,\ldots,E_{d-1}$) is known. In practice, however, the structure has to be chosen by the statistician. This choice is non-trivial, since there are $d!/2 \times d^{(d-2)(d-3)/2}$ possible vine structures (Morales-Nápoles et al., 2011), which grows excessively with d. When d is very small, it may still be practicable to estimate vine copula models for all possible structures and compare them by a suitable criterion (such as AIC). But already for a moderate number of dimensions one has to rely on heuristics.

A selection algorithm that seeks to capture most of the dependence in the first couple of trees was proposed by Dißmann et al. (2013). This is achieved by finding the maximum spanning tree using a dependence measure as edge weights, e.g., the absolute value of the empirical Kendall's τ. The resulting estimation and structure selection procedure is summarized in a general form in Algorithm 2.

Algorithm 2 Sequential estimation and structure selection for simplified vine copula models

Input: (Pseudo-)Observations $(\widehat{U}_1^{(i)}, \dots, \widehat{U}_d^{(i)})$, $i = 1, \dots, n$.
Output: Vine structure $(E_1, \dots, E_{d-1})$ and estimates of pair-copula densities and h-functions required to evaluate the vine copula density (3.16).

```
for $m = 1, \dots, d-1$:
    Calculate weights $w_e$ for all possible edges $e = \{j_e, k_e; D_e\}$ that satisfy
    the proximity condition (see Chapter 2) and select the edge set $E_m$ as

        $E_m = \arg\max_{E_m^*} \sum_{e \in E_m^*} w_e,$

    under the constraint that $E_m^*$ corresponds to a spanning tree.
    for all $e \in E_m$:
        (i) Based on $\big(\widehat{U}_{j_e|D_e}^{(i)}, \widehat{U}_{k_e|D_e}^{(i)}\big)_{i=1,\dots,n}$, obtain an estimate of the copula
            density $c_{j_e,k_e;D_e}$ which we denote as $\widehat{c}_{j_e,k_e;D_e}$.
        (ii) Derive corresponding estimates of the h-functions $\widehat{h}_{j_e|k_e;D_e}$, $\widehat{h}_{k_e|j_e;D_e}$
            by integration (2.6).
        (iii) Set

            $\widehat{U}_{j_e|D_e \cup k_e}^{(i)} := \widehat{h}_{j_e|k_e;D_e}\big(\widehat{U}_{j_e|D_e}^{(i)} \big| \widehat{U}_{k_e|D_e}^{(i)}\big),$
            $\widehat{U}_{k_e|D_e \cup j_e}^{(i)} := \widehat{h}_{k_e|j_e;D_e}\big(\widehat{U}_{k_e|D_e}^{(i)} \big| \widehat{U}_{j_e|D_e}^{(i)}\big), \quad i = 1, \dots, n.$
    end for
end for
```

Czado et al. (2013) investigated several specifications of the edge weight in a fully parametric context. The most common edge weight w_e is the absolute value of the empirical Kendall's τ. It was proposed by Dißmann et al. (2013) and used in a nonparametric context by Nagler (2014). On the other hand, Kauermann and Schellhase (2014) used a *corrected Akaike information criterion (cAIC)* (Hurvich et al., 1998) that accounts for higher-order terms in the asymptotic derivation of the AIC, making it more suitable for small sample sizes. When using the cAIC criterion in Algorithm 2, the weight w_e for edge e is

$$\mathrm{cAIC}_e = -2\ell_e + 2\mathrm{df}_e + \frac{2\mathrm{df}_e(\mathrm{df}_e + 1)}{n - \mathrm{df}_e - 1}, \tag{3.18}$$

where

$$\ell_e = \sum_{i=1}^{n} \ln \widehat{c}_{j_e,k_e;D_e}\big(\widehat{U}_{j_e|D_e}^{(i)}, \widehat{U}_{k_e|D_e}^{(i)}\big),$$

Dimension d	Sample Size n	Type of dependence	Strength of dependence
5	400	only tail dependence	weak
10	2 000	no tail dependence	strong
		both types	

Table 3.1: List of factors that determine the set of simulation scenarios.

is the log-likelihood and df_e is the *effective degrees of freedom (EDF)* of the estimator $\widehat{c}_e$. For explicit formulas for the EDF we refer to Kauermann and Schellhase (2014) for the spline approach and to Loader (1999) for the kernel estimators. For parametric copula models, the EDF typically equals the number of model parameters.

From a computational point of view, the cAIC has a big disadvantage: the cAIC for an edge can only be calculated *after* a model for this edge has been fitted. Hence, before a tree can be selected, the pair-copulas of all possible edges in this tree have to be estimated. Just for the first tree, this amounts to estimating $\binom{d}{2}$ bivariate copula densities, of which only $d1$ will be kept in the model. The empirical Kendall's τ on the other hand can be computed rapidly for all pairs. It allows to select the tree structure before any pair-copula has been estimated. Then, only $d-1$ pair-copulas have to be estimated in the first tree. The situation is similar for subsequent trees. Both approaches will be compared in our simulation study with regard to estimation accuracy and speed.

Other selection methods specialized on high-dimensional parametric vine copulas were proposed by Müller and Czado (2017a,b, 2018), and Nagler et al. (2018).

3.3 Description of the simulation study design

We compare the performance of the vine copula density estimators discussed in Section 3.2 over a wide range of scenarios. We consider several specifications of sample size, dimension, strength of dependence, and tail dependence. We randomize the simulation models and characterize the scenarios by probability distributions for the pair-copula families and dependence parameters. A detailed description of the study design procedure will be given in the following sections.

3.3.1 Simulation scenarios based on model randomization

To investigate how various factors influence the estimators' performance, we create a number of scenarios. Each of these scenarios is characterized by a combination of the factors shown in Table 3.1.

To make the results for a particular dependence scenario as general as possible, we randomly generate a model in the following steps:

Step 1. **Draw R-vine structure:**
We do this in the following steps:

(i) Draw n samples for d independent uniform random variables, $\tilde{U}_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, d$.

(ii) On these samples, run the structure selection algorithm of Dißmann et al. (2013) (only allowing for the independence family).

(iii) Set the model structure to the one selected by the algorithm.

Step 2. **Draw pair-copula families:**

- *only tail dependent copulas*: draw each of the $d(d-1)/2$ pair-copula families with equal probabilities from the Student t- ($df = 4$), Gumbel (with rotations) and Clayton (with rotations) copulas.
- *no tail dependence*: draw each of the $d(d-1)/2$ pair-copula families with equal probabilities from the Gaussian and Frank copulas.
- *both*: for each of $d(d-1)/2$ pair-copulas:

  (i) choose with equal probabilities whether the copula has tail dependence or not,

  (ii) proceed as above.

Step 3. **Draw pair-copula parameters:**
For each pair-copula:

(i) Randomly generate the absolute value of Kendall's τ from the following distributions:

- *weak dependence: Beta*$(1, 4)$-distribution ($E[|\tau|] = 0.2$),
- *strong dependence: Beta*$(5, 5)$-distribution ($E[|\tau|] = 0.5$).

The densities are shown in Figure 3.1.

(ii) Randomly choose the sign of Kendall's τ as *Bernoulli*(0.5) variable.

(iii) Usually, conditional dependence is weaker than direct pair-wise dependence. To mimic this behavior we decrease the simulated absolute Kendall's τ by a factor of 0.8^m, where m is the tree level of the pair-copula.

(iv) For all families under consideration there is a one-to-one relationship between the copula parameter and Kendall's τ, see e.g., Table 2 in Brechmann and Schepsmeier (2013). Hence, we set the copula parameter by inversion of the reduced value of Kendall's τ.

Step 4. **Draw observations from the final model:**
With the selected structure, copula families and their parameters, the vine copula model is fully specified. We can draw random samples from this vine copula model using the algorithm of Stöber and Czado (2012), as implemented in the `VineCopula` R library (Schepsmeier et al., 2018).

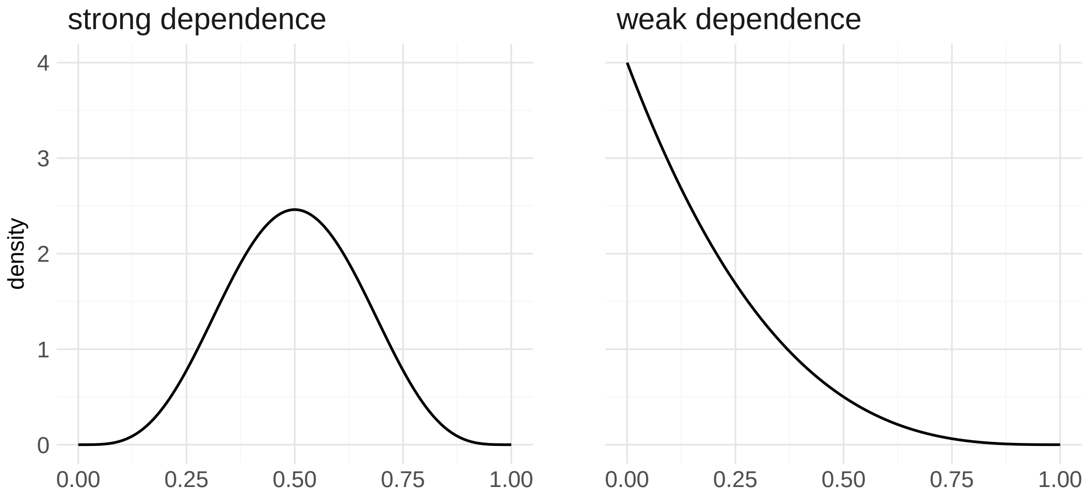


Figure 3.1: Densities for the simulation of absolute Kendall's τ in the scenarios with weak (left) and strong (right) dependence.

The stochastic model characterized by steps 1–4 can be interpreted as a whole. It is a mixture of vine copula models, mixed over its structure, families, and parameter. The mixing distribution for the families is uniform over sets determined by the 'type of dependence' hyper-parameter. The mixing distribution for the absolute Kendall's τ follows a Beta distribution with parameters characterized by the 'strength of dependence' hyper-parameter. Each scenario corresponds to a particular specification of the mixture's hyperparameters. The benefit of this construction is that it yields models that are representative for a wide range of scenarios encountered in practice. It also limits the degrees of freedom we would have when specifying all pair-copula families and parameters manually.

3.3.2 Estimation methods

We compare the following pair-copula estimators:

- `par`: parametric estimator as implemented in the function `BiCopSelect` of the R package `VineCopula` (Schepsmeier et al., 2018). It estimates the parameters for several parametric families and selects the best model based on AIC. The implemented families are: Independence, Gaussian, Student t, Clayton, Frank, Gumbel, Joe, BB1, BB6, BB7, BB8, Tawn types I and II,

- `bern`: non-penalized Bernstein estimator (see Section 3.2.1),

- `pbern`: penalized Bernstein estimator (see Section 3.2.1) with $K = 14$ knots,

- `pspl1`: penalized linear B-spline estimator (see Section 3.2.1) with $K = 14$,

- psp12: penalized quadratic B-spline estimator (see Section 3.2.1) with $K = 10$,
- tll0: transformation local likelihood kernel estimator of degree $q = 0$ (see Section 3.2.1),
- tll1: transformation local likelihood kernel estimator of degree $q = 1$ (see Section 3.2.1),
- tll2: transformation local likelihood kernel estimator of degree $q = 2$ (see Section 3.2.1).

We further implemented two structure selection methods for each pair-copula estimator (based on Kendall's τ and cAIC, see Section 3.2.3); additionally we computed each estimator under the true vine structure.

3.3.3 Performance measurement

As a performance measure, we choose the *integrated absolute error (IAE)*

$$\mathrm{IAE} = \int_{[0,1]^d} |\widehat{c}(\boldsymbol{u}) - c(\boldsymbol{u})| d\boldsymbol{u},$$

where $\widehat{c}$ is the estimated and c is the true copula density. The above expression requires us to calculate a d-dimensional integral, which can be difficult when d becomes large. To overcome this, we estimate this integral via importance sampling Monte Carlo, see e.g., Section 5.2 in (Ripley, 1987). That is,

$$\widehat{\mathrm{IAE}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{i=1,\ldots,N} \frac{|\widehat{c}(\boldsymbol{U}_i) - c(\boldsymbol{U}_i)|}{c(\boldsymbol{U}_i)},$$

where $\boldsymbol{U}_i \overset{iid}{\sim} c$ is a random vector drawn from the true copula density c. This results in an unbiased estimator of the IAE with relatively small variance: usually the numerator is large/small when the denominator is large/small. Hence, the variance of all terms in the sum is small and, hence, the variance of the sum is small. All results will be based on an importance sample of size $N = 1\,000$.

For each estimator and each possible simulation scenario emerging from Table 3.1, we record the $\widehat{\mathrm{IAE}}$ on $R = 100$ simulated data sets.

3.4 Results

Figure 3.2 and Figure 3.3 present the results of the simulation study described in Section 3.3. The analysis will be divided into several sections. The first takes a very broad view, whereas the remaining ones investigate the influence of individual factors. We acknowledge that the information density in the figures is extremely high. So we start with a detailed description of the figures' layout.

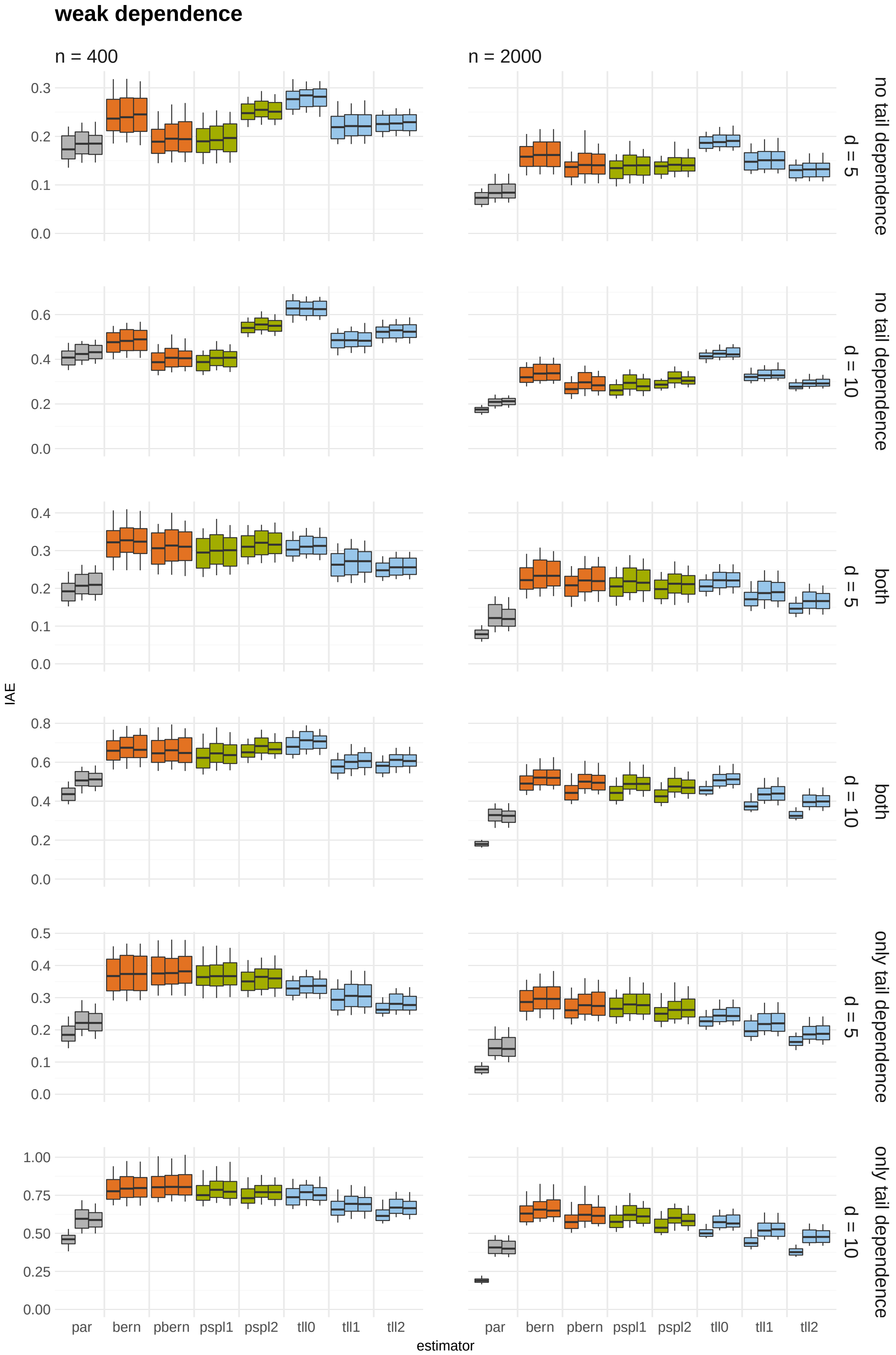


Figure 3.2: *weak dependence:* the box plots show the IAE achieved by each estimation method. Results are split by sample size, dimension, and type of dependence. Per estimator there are three boxes, corresponding to estimation under known structure, selection by Kendall's τ, and selection by cAIC (from left to right).

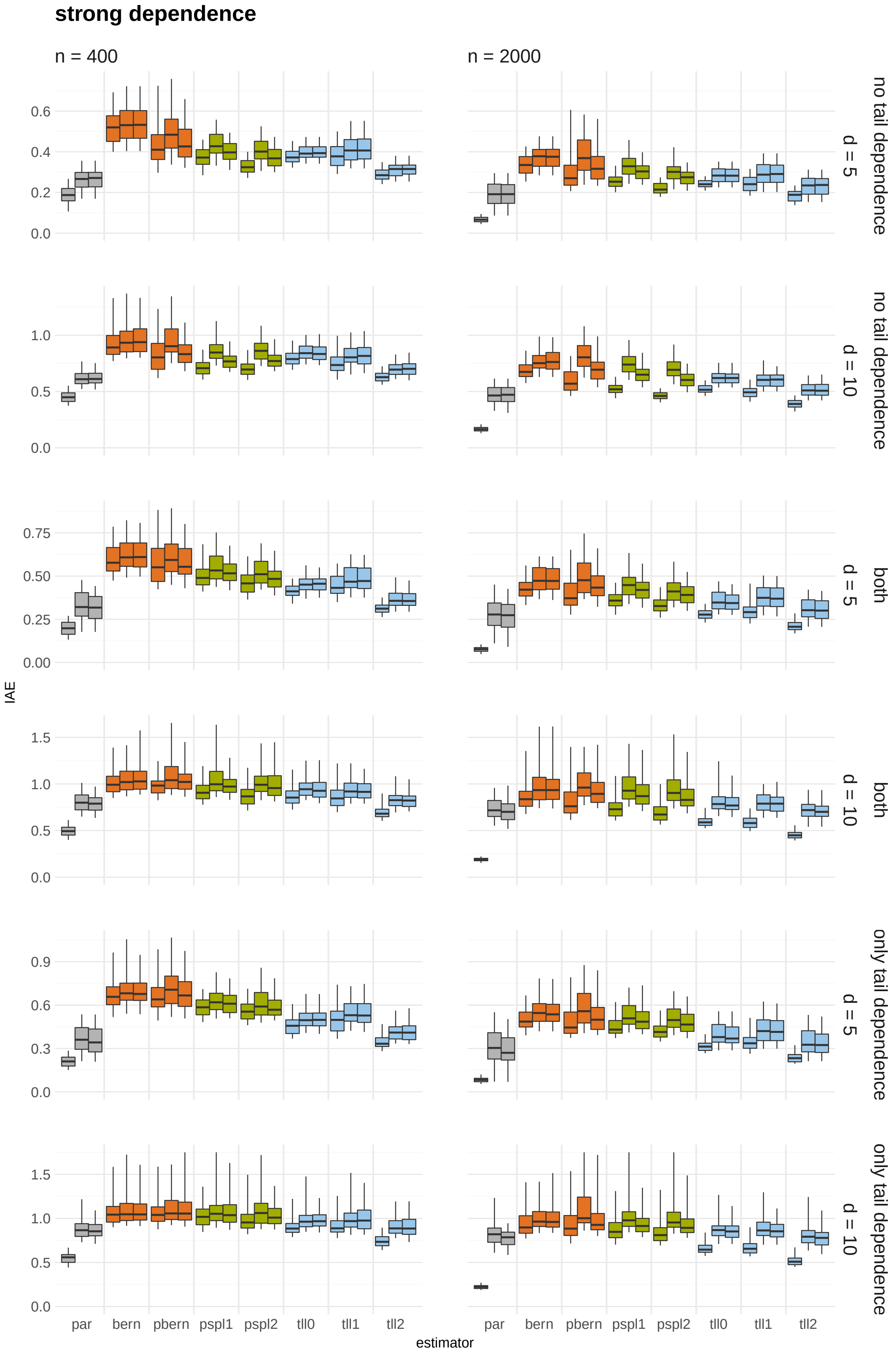


Figure 3.3: *strong dependence:* the box plots show the IAE achieved by each estimation method. Results are split by sample size, dimension, and type of dependence. Per estimator there are three boxes, corresponding to estimation under known structure, selection by Kendall's τ, and selection by cAIC (from left to right).

	`par`	`bern`	`pbern`	`pspl1`	`pspl2`	`tll0`	`tll1`	`tll2`
average rank	1.28	7.17	6.35	5.00	5.01	5.01	3.83	2.35

Table 3.2: The relative rank of estimators averaged over all scenarios.

Figure 3.2 contains the results for all scenarios with weak dependence; Figure 3.3 with strong dependence. The left columns correspond to the smaller sample size ($n = 400$) and the right columns to the larger sample size ($n = 2\,000$). The figures are also partitioned row-wise with an alternating pattern of the dimensions $d = 5$ and $d = 10$. Two subsequent rows correspond to the same type of dependence (no tail dependence, both, only tail dependence). In total there are 32 panels, each representing one of the 32 possible combinations of the factors listed in Table 3.1.

Each panel contains 24 boxes in 8 groups. Each group corresponds to one estimation method for the pair-copulas. The three boxes in each group represent the three different methods for structure selection: known structure, maximum spanning trees with Kendall's τ, maximum spanning trees with cAIC (from left to right). The box spans the interquartile range, the median is indicated by a horizontal line, the whiskers represent the 10% and 90% percentiles.

3.4.1 Overall ranking of methods for pair-copula estimation

We begin our analysis with a broad view on the relative performance of the pair-copula estimators. We want to assess the performance of the estimation methods, averaged over all scenarios and structure selection strategies. But just taking the average IAE could be misleading. It is evident from Figures 3.2 and 3.3 that the scale of the IAE varies between scenarios. Averaging the bare IAE leads to an unbalanced few, laying more weight on particular scenarios. As a more robust alternative, we take the following approach: in each scenario, average the IAE over replications and structure selection strategies. Then rank the estimation methods by their relative performance. Ranks are comparable across scenarios, so our final criterion will be the average rank across all scenarios. These numbers are listed in Table 3.2.

The parametric estimator performs best, which is no surprise since our simulation models consist of only parametric copula families. We included it in this study mainly to get a sense of what is possible in each scenario. Remarkably, it is outperformed in very few cases by a nonparametric estimator. This is due to the need for structure selection which will be discussed in more detail later on.

Among the nonparametric estimators, the kernel estimators (`tll2`, `tll1`, `tll0`) perform best, followed by the spline methods (`pspl1`, `pspl2`) which perform as well as the worst kernel estimator `tll0`. The Bernstein estimators (`pbern`, `bern`) perform worst. Within these three classes, the accuracy improves mostly by how complex the estimation method is: going from regular Bernstein copulas to penalized ones; and going from local constant, to local linear, to local quadratic likelihood. It is the other way around for the B-spline methods, but the difference in the average rank is minuscule.

We will find that this relative ranking is fairly robust across scenarios. In the following analysis, we treat it as the benchmark ranking and focus on deviations from it.

3.4.2 Strength and type of dependence

By looking at the scale in each panel, we see that the performance of all estimators gets worse for increasing strength of dependence and increasing proportion of tail dependent families. This is explained by the behavior of the true densities. Many copula densities (and their derivatives) explode at a corner of the unit square. From the pair-copula families in our simulation model, only the Frank copula is bounded. Within each family, the tails explode faster when the strength of dependence increases. And tail dependence means that the tails explode particularly fast. Exploding curves are difficult to estimate for nonparametric estimators because their asymptotic bias and variance are usually proportional to the true densities' derivatives. Our results give evidence that this effect transfers to finite samples.

The estimators' response to these difficulties is the main driver behind their relative performance. In most scenarios, the ranking of estimators is similar to the benchmark rankings. But there are deviations. Let us walk through the scenarios one by one.

- *weak, no tail dependence*: `pbern1` and `pspl1` perform better than `pspl2`, the kernel estimators, and even the parametric estimator for $n = 400$. For $n = 2\,000$, the parametric estimator gets ahead and the penalized methods are on par with `tll1` and `tll2`.

- *weak, both*: `pbern1` and `pspl1` perform better than `pspl2` and `tll0` for $n = 400$, and comparable for $n = 2\,000$.

- *weak, only tail dependent copulas*: similar to the benchmark ranking.

- *strong, no tail dependence*: `bspl2` beats `tll0` and `tll1` for $n = 400$ and is on par for $n = 2\,000$.

- *strong, both*: similar to benchmark ranking.

- *strong, only tail dependent copulas*: similar to benchmark ranking.

Overall, the penalized estimators tend to do better under weak dependence and only little tail dependence, whereas the kernel estimators do better in the other scenarios. The method `tll2` is the top performer in all but a few cases.

3.4.3 Sample size and dimension

When the sample size increases, the estimators become more accurate. Any reasonable estimator should satisfy this property. The kernel estimators and the non-penalized Bernstein estimator seem to benefit more from an increased sample size. The effect is most obvious in the weak dependence, no tail dependence case. This has an explanation: theoretically, the number of knots used by the penalized estimators should increase with the sample size. But our implementation uses a fixed number of knots, as the computational burden is already substantial compared with the other methods (see Section 3.4.5). All other methods adapt their smoothing parameterization to the sample size. It is very likely that the penalized methods improve when the number of knots is further increased.

Comparing a pair of panels corresponding to $d = 5$ and $d = 10$, we see only little differences. We conclude that the results are quite robust to changes in the dimensionality.

3.4.4 Structure selection

The first aspect we want to discuss is the loss in accuracy caused by the need to select the tree structure. Recall that the three subsequent boxes for each estimator correspond to: estimation under the true structure (in practice unknown), selection based on Kendall's τ, selection based on cAIC.

The IAEs for the two selection methods are always higher than the 'oracle' results with known structure. This makes sense: the true model is a simplified vine copula; if the true structure is known, the models are correctly specified and all estimators are consistent. In practice, the true structure is unknown, and a different structure will be selected most of the time. For the selected structure, there is no guarantee that the model is still simplified or that the estimators are consistent. For more details, we refer to Spanhel and Kurz (2017) and Section 4.7.

Overall, the average loss in accuracy when going from the true to a heuristically selected structure increases with strength of dependence and prevalence of tail dependence. But the extent of this effect varies between estimation methods. The parametric estimator suffers the most substantial losses. In fact, the parametric estimator's performance is often very close to that of the best nonparametric estimator when the structure is unknown. This is quite remarkable considering that we simulate from parametric models. Interestingly, the loss for the penalized Bernstein and B-spline methods (`pbern`, `pspl1`, `pspl2`) is negligible in most scenarios when cAIC is used—but not when Kendall's τ is used. This is a distinct property of these penalized methods. The non-penalized Bernstein and kernel methods perform similarly for the two structure selection criteria. In most scenarios, the relative performance ordering of the estimators is the same for each type of structure. But there are a few cases (strong dependence, n = 400) where the `bspl2` estimator is worse than `tll0` or `tll1` with Kendall's τ, but better with cAIC.

The results give evidence that the cAIC is the better criterion in terms of the estimators' accuracy. But it also makes the vine copula estimators more costly to fit (see Section 3.2.3). So there is a trade-off between speed and accuracy. It usually depends on the application which to prioritize. We will investigate this issue further in the next section.

d	n	criterion	par	bern	pbern	pspl1	pspl2	tll0	tll1	tll2
5	400	τ	7	3	788	758	517	3	4	6
		cAIC	19	10	1 000	1 175	786	10	11	13
	2 000	τ	34	19	1 578	1 455	1 394	7	12	16
		cAIC	91	31	2 163	2 336	2 243	25	32	35
10	400	τ	33	17	2 983	3 183	2 205	14	19	29
		cAIC	98	49	5 292	6 110	4 156	48	55	65
	2 000	τ	159	65	6 553	6 694	6 515	35	56	71
		cAIC	472	139	11 992	13 514	12 394	127	158	173

Table 3.3: Average computation time (in seconds) required for estimation and selection of one vine copula model.

3.4.5 Computation time

Table 3.3 lists the average computation time[1] required to fit a vine copula and evaluate its density on 1 000 importance Monte-Carlo samples. The results are divided into the combinations of dimension d and sample size n.

Let us first focus on the selection criterion. We clearly see that the computation time increases substantially for all estimators when cAIC is used instead of Kendall's τ. This effect size differs, but is usually a factor of around two or three.

The fastest two estimators are the simplest ones: `bern` and `tll0`. The other two kernel estimators are in the same ballpark, but the computation time increases slightly with the order of the polynomial. Only slightly slower is the parametric estimator. The reason is that the parametric estimator has to iterate through several different copula families before it can select the final model. The penalized estimators are roughly two orders of magnitude slower than their competitors. Take for example the case $d = 10$ and $n = 2\,000$, where most estimators take around one minute (using τ), but the penalized estimators take more than one and a half hours.

The large difference in computational demand is caused by the penalized estimation problem. One has to optimize over more than 100 parameters with more than 100 side constraints. Even worse, such a problem has to be solved multiple times until an optimal choice for the penalty parameter λ has been found. Reducing K (the number of knots) does significantly reduce this burden, but also limits the flexibility of the estimators.

3.4.6 Limitations

Two anonymous referees pointed out some limitations of our study which are addressed in the following.

[1] The time was recorded on a single thread of a 8-way Opteron (Dual-Core, 2.6 GHz) CPU with 64GB RAM.

Performance measure

All results focus on a single performance measure and therefore only provide a limited view on the estimators' performance. Although this is true, we considered several other measures in preliminary versions of this study (integrated squared error, Hellinger distance, and Kullback-Leibler divergence) and found the results to be quite robust with respect to the measure.

Estimation of marginal distributions

The study neglects the fact that observations from the copula are never observed and one has to rely on pseudo-observations that depend on estimated marginal distributions. An extensive simulation study in Kim et al. (2007) revealed that this can be a problem when severely misspecified parametric models are used for the margins. But the issue is largely resolved when the margins are estimated nonparametrically. In this case, maximum likelihood estimators are unbiased and only slightly less efficient (Genest et al., 1995).

In a purely nonparametric context, this is even less of an issue. In fact, many authors have found that errors stemming from estimating the marginal distributions are asymptotically negligible when estimating the copula density, see e.g., (Janssen et al., 2014, Geenens et al., 2017, Nagler and Czado, 2016). This is explained by the fact that distribution functions can be estimated consistently at the parametric $\sqrt{n}$-rate, whereas density estimators are necessarily slower (see, Stone, 1980). Accordingly, we can expect similar results to the ones presented even if margins were treated unknown.

Choice of smoothing parameters

It is a common theme in nonparametric estimation that the quality of estimators depends heavily on the choice of smoothing parameters. This is certainly also the case for the estimators considered in this study. However, we do not think it is feasible to assess the sensitivity of our results to this choice:

- Smoothing parameters are hardly comparable across estimation methods because they arrive at the density estimate in fundamentally different ways.

- There are too many smoothing parameters in a vine copula model: There are 10 ($d = 5$) resp. 45 ($d = 10$) pair-copulas, and for each pair-copula there are between one and three smoothing parameters (depending on the estimation method).

- Due to the sequential nature of the joint estimator, pair-copula estimators in later trees are affected by the estimates in earlier trees. This leads to significant interactions between smoothing parameters at different levels.

In our study, all smoothing parameters were selected by automatic procedures, which reflects statistical practice. But one should keep in mind that the per-

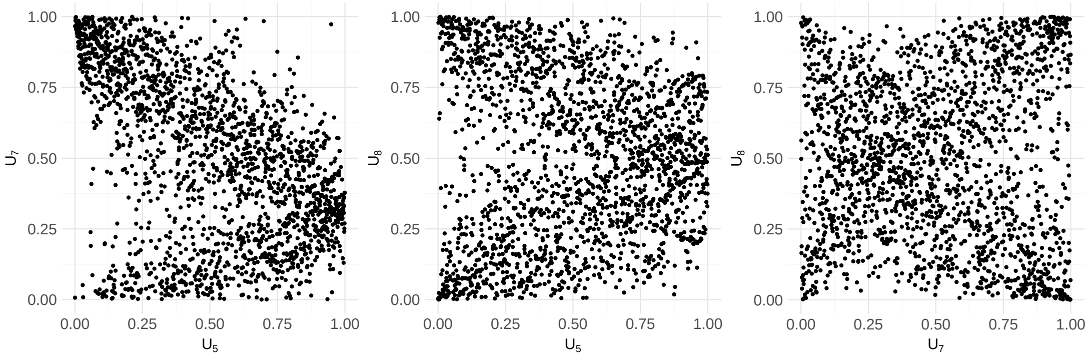


Figure 3.4: Scatterplots of pseudo observation ranks for pairs (left) fConc1 (U_5) and fM3Long (U_7), (middle) fConc1 (U_5) and fM3Trans (U_8) and (right) fM3Long (U_7) and fM3Trans (U_8) from the MAGIC data set ($n = 2\,000$).

formance of most estimators can likely be improved by advances in automatic selection methods.

3.5 Illustration with real data

In the simulation study, the parametric estimator performed best in virtually all scenarios. But this is simply a consequence of simulating from parametric models. Real data is not always that well-behaved and nonparametric methods are required to appropriately capture the dependence. Such a situation is illustrated in the following example.

We consider a data set representative of measurements taken on images from the MAGIC (Major Atmospheric Gamma-ray Imaging Cherenkov) Telescopes (https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope) with 19 020 observations for eleven different attributes, but focus only on gamma observations.

To show exemplary results of the different nonparametric copula density estimators, we select a random subset ($n = 2\,000$) from the MAGIC data with respect to to the three variables fConc1, fM3Long and fM3Trans. We compute pseudo-observations from the data by applying the marginal empirical distribution functions to each variable. Figure 3.4 shows scatter plots of the three pairs of the pseudo-observations. The shapes we see are different from what we know from popular parametric families. We fit several copula density estimators to each pair and show the results in Figure 3.5. The first column of Figure 3.5 shows the fitted pair-copula density between fConc1 (U_5) and fM3Long (U_7), the second column between fConc1 (U_5) and fM3Trans (U_8) and the third column contains the copula density between fM3Long (U_7) and fM3Trans (U_8).

The first pair of variables fConc1 (U_5) and fM3Long (U_7) a lot of pseudo-observations accumulate around the point $(0, 1)$, which is reflected as high density peaks in all fitted copula densities. But for the accumulation around the point

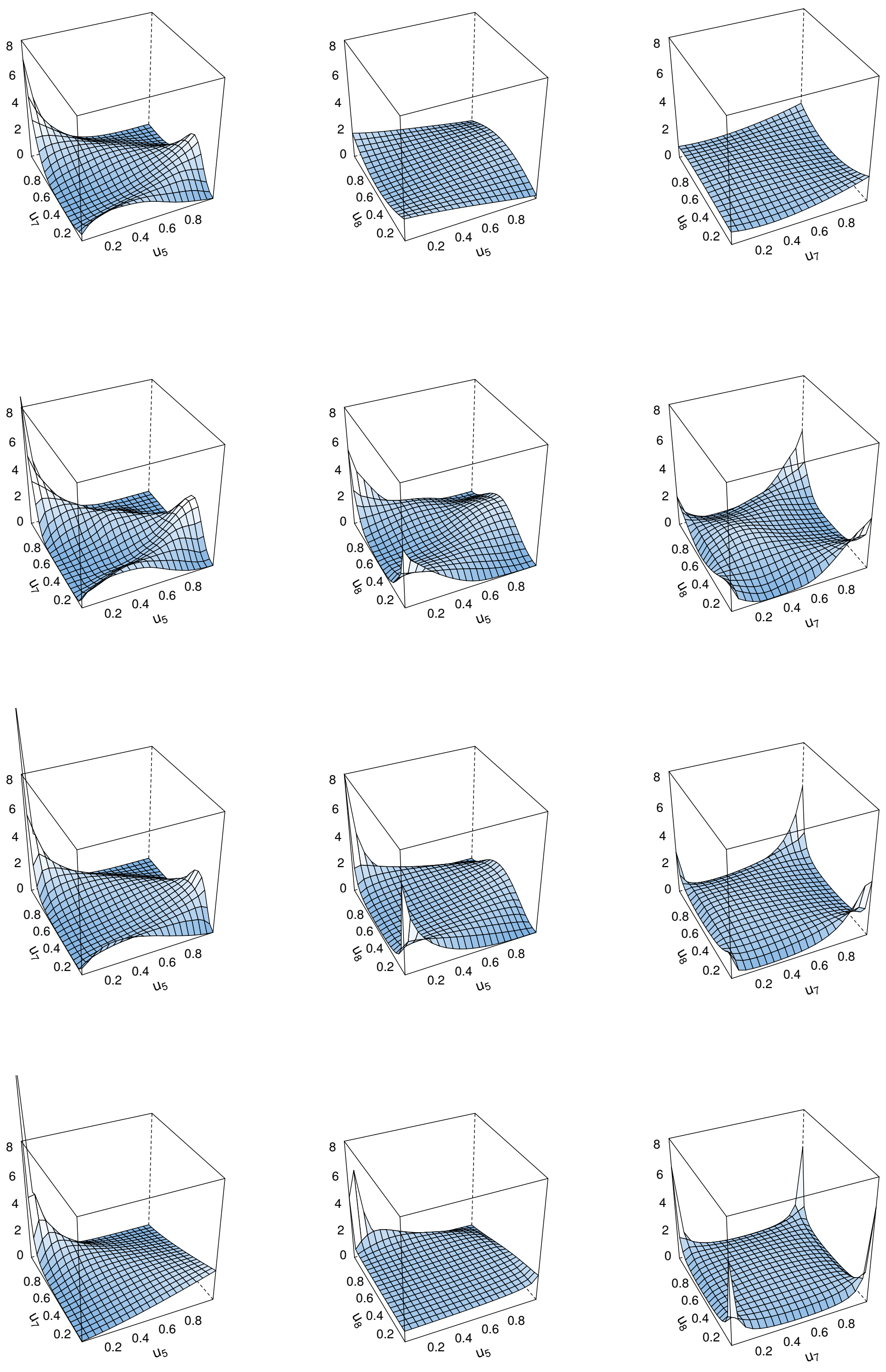


Figure 3.5: Exemplary density plots for MAGIC data ($n = 2\,000$). 1st row: Bernstein estimator `bern`, 2nd row: penalized quadratic B-splines estimator `pspl2`, 3rd row: kernel estimator `tll2`, 4th row: parametric estimator `par`.

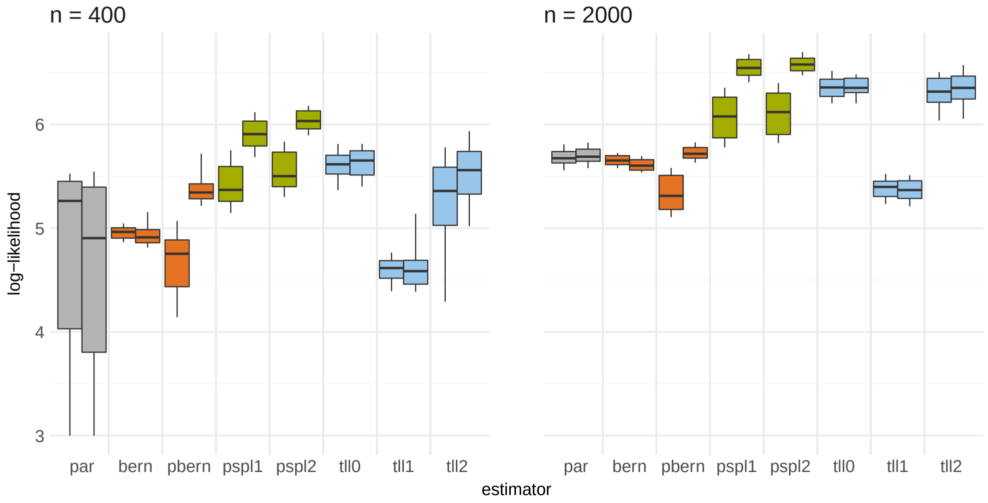


Figure 3.6: The box plots show the mean log-likelihood values attained by the different estimation methods. Each boxplot on the left hand side: structure selection based on Kendell's τ and each boxplot on the right hand side: structure selection based on cAIC.

$(1, 0.3)$, we observe a difference between the nonparametric estimators `bern`, `pspl2` and `ttl2` and the parametric copula density, which does not mirror this accumulation.

For the second pair, fConc1 (U_5) and fM3Trans (U_8), the estimated density varies considerably between methods. Estimates of `pspl2` and `ttl2` show peaks around the points $(0, 0)$ and $(0, 1)$, which reflects the large concentration of points in the scatter plot in Figure 3.4. The estimators `bern` and `par` do not contain these peaks. We observe similar differences for the estimated densities for the third data pair, presented in the right column of Figure 3.4. While `bern`, `pspl2` and `ttl2` show density peaks around the accumulation points $(1, 0)$ and $(1, 1)$, but the estimated parametric copula does not exhibit these structures of the data.

The previous examples have illustrated situations, in which the parametric estimator fails because of its lack of flexibility. In such situations, nonparametric methods are required to adequately capture the true dependence structure. However, for illustrations, we merely looked at three unconditional pairs of variables, not a full dependence model.

To analyze the the performance of the estimators in an application, we estimate nonparametric vine copulas for the complete MAGIC data set with eleven attributes focusing on gamma observations. Because the true density is unknown in applications, we assess the performance of the estimators via cross-validation. Similar to our simulation study, we randomly draw a subset $\boldsymbol{U}_{\text{train}}$ of the data of sizes $n = 400, 2\,000$, apply the estimators, and calculate the out-of-sample log-likelihood on $1\,000$ randomly selected remaining observations $\boldsymbol{U}_{\text{test}}$, i.e.,

$$\ell(\boldsymbol{U}_{\text{test}}) = \frac{1}{1\,000}\sum_{i=1}^{1\,000} \ln \hat{c}(\boldsymbol{U}_{\text{test}}^{(i)}),$$

where $\hat{c}$ is a vine copula density estimator based on $\boldsymbol{U}_{\text{train}}$. This is repeated $N = 100$ times for sample sizes $n = 400$ and $n = 2\,000$. The results are summarized as box plots in Figure 3.6 for all estimators and structure selection based on Kendall's τ (left box) and cAIC (right box).

The parametric estimator performs unsatisfactory for $n = 400$ since it varies enormously for both structure selection methods. But also for $n = 2\,000$, the parametric estimator is outperformed by most nonparametric alternatives. The performance of the nonparametric methods varies notably between methods. The methods `bern` and `tll1` do not perform well, but the other methods clearly outperform par. Furthermore, the performance differs significantly with respect to the structure selection criterion for `bern1` and `pspl1, pspl2`. For small sample size ($n = 400$) and using Kendall's τ as selection criterion, `tll0` results with highest out-of-sample likelihood, directly followed by `pspl2`. But choosing cAIC instead, the log-likelihood of `pspl2` increases, but not for `tll0`. The situation is similar for $n = 2\,000$. We conclude that the more sophisticated nonparametric methods adequately reflect the distribution of the data. In contrast, the dependence structure observed in Figure 3.4 can not be captured well with parametric models.

3.6 Conclusion

This chapter compared existing methods for nonparametric estimation of simplified vine copula densities. The estimators considered are the non-penalized Bernstein estimator, the penalized Bernstein estimator, penalized B-spline estimators (linear and quadratic), and kernel weighted local likelihood estimators (local constant, linear, and quadratic). We compared these methods by an extensive simulation study and a real data set.

The simulation study comprises several scenarios for sample size, dimension, strength of dependence, and tail dependence. The simulation models are set up as parametric vine copulas with randomized vine structure, pair-copula families, and parameters. Overall, the kernel methods were found to perform best (especially the local quadratic version), followed by the penalized B-spline estimators. The Bernstein estimators performed worst. An exception to this pattern was found in scenarios with small sample size, weak dependence, and no tail dependence. Here, the penalized B-spline and Bernstein estimators outperformed the kernel methods. Additionally, we demonstrated the need for nonparametric methods on real data whose dependence structure cannot be adequately captured by a parametric estimator.

Overall, we found that no estimator is uniformly better than the others; it depends on the data which is to be preferred. Our analysis highlighted which factors drive the performance of the various methods, and which methods should be preferred for certain scenarios. In applications, statisticians can determine the characteristics of their data by an exploratory analysis, and make a well-informed choice based on these results.

4

Evading the curse of dimensionality with simplified vine copulas

4.1 Introduction

Density estimation is one of the most classical 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 (Bock et al., 2004, Aitken and Lucy, 2004, Kie et al., 2010). 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. Scott (2008, Section 7.2) illustrates this phenomenon for a kernel density estimator when the standard Gaussian is the target density: to achieve an accuracy comparable to $n = 50$ observations in one dimension, more then $n = 10^6$ observations are required in ten dimensions.

In general, this issue cannot be solved: Stone (1980) 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, he showed that if an estimator satisfies

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

for all densities f of this class and some $r > 0$, then it must hold $r \leq p/(2p + d)$. 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 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 violated.

4.1.1 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 Section 2.2.3.

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 for densities (see Section 2.1) 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 from Section 2.2.2 is that any d-dimensional copula density can be decomposed into a product of $d(d-1)/2$ bivariate (conditional) copula densities. 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 $c_{1,3;2}\{F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)\,;\, x_2\}$ 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 Stöber et al. (2013). 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 (see, e.g., Kunihiro et al., 2004). 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 in Theorem 4.1.

Nonparametric estimation of simplified vine copula densities has been discussed earlier using kernels (Lopez-Paz et al., 2012) and smoothing splines (Kauermann and Schellhase, 2014). However, both contributions lack an analysis of the asymptotic behavior of the estimators. We treat the more general setting of densities with arbitrary support. Theorem 4.1 shows under high-level conditions that the convergence rate of a nonparametric estimator of a simplified vine density is independent of the dimension — an extremely powerful property that has been overlooked so far.

4.1.2 Organization

A generic estimator of simplified vine densities is described in detail in Section 4.2. In Section 4.3 we show under high-level assumptions that such an estimator is consistent and that the convergence rate is independent of the dimension. Hence, there is no curse of dimensionality. In Section 4.4 we discuss how the method can be implemented as a kernel estimator. For this particular implementation, we validate the high-level assumptions of Theorem 4.1 and establish asymptotic normality. We illustrate its advantages over the classical multivariate kernel density estimator via simulations in both simplified and non-simplified settings (Section 4.5). The method is applied to a classification problem from astrophysics in Section 4.6. We conclude with a discussion of our results and provide links to the existing literature on the simplifying assumption in Section 4.7.

4.2 Nonparametric density estimators based on simplifed vine copulas

Let $\boldsymbol{X} = (X_1, \dots, X_d)$ be a random vector with continuous joint distribution F and marginal distributions $F_1, \dots, F_d$. The support of $\boldsymbol{X}$ and $X_\ell, \ell = 1, \dots, d$, will be denoted as $\Omega_{\boldsymbol{X}}$ and Ω_{X_ℓ}, respectively. Let further $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \dots, X_d^{(i)})$, $i = 1, \dots, n$, be *iid* copies of $\boldsymbol{X}$ (acting as observations). Assume that F is a simplified vine distribution with structure $\mathcal{V} = (E_1, \dots, E_{d-1})$. Provided densities exist, we can use Sklar's theorem and (2.4) to write the joint density f for all $\boldsymbol{x} = (x_1, \dots, x_d) \in \Omega_{\boldsymbol{X}}$ as

$$\begin{aligned} f(\boldsymbol{x}) &= c\big\{F_1(x_1), \dots, F_d(x_d)\big\} \times \prod_{l=1}^{d} f_\ell(x_\ell) \\ &= \prod_{m=1}^{d-1} \prod_{e \in E_m} c_{j_e,k_e;D_e}\big\{F_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}),\, F_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\big\} \times \prod_{l=1}^{d} f_\ell(x_\ell). \end{aligned} \tag{4.1}$$

The conditional distribution functions $F_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})$ can equivalently be expressed as $G_{k_e|D_e}(u_{k_e}|\boldsymbol{u}_{D_e})$, where $\boldsymbol{u} = (u_1, \dots, u_d) = (F_1(x_1), \dots, F_d(x_d))$. This allows us to decompose $F_{k_e|D_e}$ recursively into marginal distributions and h-functions (see Section 2.2.4).

The idea is now to estimate all functions in the above expression separately. We proceed as follows:

1. Based on the observations $(X_1^{(i)}, \dots, X_d^{(i)})$, $i = 1, \dots, n$, we obtain estimates $\widehat{f}_1, \dots, \widehat{f}_d, \widehat{F}_1, \dots, \widehat{F}_d$ of the marginal densities $f_1, \dots, f_d$ and distribution functions $F_1, \dots, F_d$.

2. Recall that c is the density of the random vector $\boldsymbol{U} = \big(F_1(X_1), \dots, F_d(X_d)\big)$. We do not have access to observations from this vector. However, we can define pseudo-observations $\boldsymbol{U}^{(i)} = \big(\widehat{U}_1^{(i)}, \dots, \widehat{U}_d^{(i)}\big)$ by replacing $F_1, \dots, F_d$ with the estimators from the last step:
$$\big(\widehat{U}_1^{(i)}, \dots, \widehat{U}_d^{(i)}\big) = \big(\widehat{F}_1(X_1^{(i)}), \dots, \widehat{F}_d(X_d^{(i)})\big), \qquad i = 1, \dots, n. \tag{4.2}$$

3. Estimate all pair-copula densities and h-functions with Algorithm 1 from Section 3.2.2.

At the end of the procedure we have estimates for all marginal distributions/densities, bivariate copula densities, and all h-functions that are required to evaluate the R-vine density (4.1). For all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ we define an estimate of the simplified vine density f as

$$\widehat{f}_{\text{vine}}(\boldsymbol{x}) = \prod_{m=1}^{d-1} \prod_{e \in E_m} \widehat{c}_{j_e,k_e;D_e}\bigl\{\widehat{F}_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}),\, \widehat{F}_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\bigr\} \times \prod_{\ell=1}^{d} \widehat{f}_\ell(x_\ell). \tag{4.3}$$

4.3 Main results

We shall establish weak consistency of the simplified vine density estimator defined in Section 4.2. We furthermore show that its probabilistic convergence rate does not increase with dimension and, hence, there is no curse of dimensionality.

4.3.1 Consistency and rate of convergence

The sequential nature of the estimator complicates its analysis. Estimation errors will propagate from one tree to the next and affect the estimation in higher trees. We impose high-level assumptions on the uni- and bivariate estimators that allow us to establish our main result.

The first assumption considers the consistency of univariate density and distribution function estimators. Although estimators may converge at different rates, we will formulate all assumptions with respect to the to the same rate n^{-r}, $r > 0$. This rate then has to be the slowest among all estimators involved — typically the rate of the pair-copula density estimator.

Assumption 4.1. *For all $\ell = 1, \dots, d$, and all $x_\ell \in \Omega_{X_\ell}$, it holds*

$$(a) \quad \widehat{f}_\ell(x_\ell) - f_\ell(x_\ell) = O_p(n^{-r}), \qquad (b) \quad \sup_{x_\ell \in \Omega_{X_\ell}} \bigl|\widehat{F}_\ell(x_\ell) - F_\ell(x_\ell)\bigr| = o_{a.s.}(n^{-r}).$$

Next, assume we are in an ideal situation where, for each edge $e \in E_m, m = 1, \dots, d-1$, we have access to the true (but unobservable) pair-copula samples

$$U^{(i)}_{j_e|D_e} := F_{j_e|D_e}\bigl(X^{(i)}_{j_e}|\boldsymbol{X}^{(i)}_{D_e}\bigr), \qquad U^{(i)}_{k_e|D_e} := F_{k_e|D_e}\bigl(X^{(i)}_{k_e}|\boldsymbol{X}^{(i)}_{D_e}\bigr), \tag{4.4}$$

$i = 1, \dots, n,$. Recall that estimators are functions of the data, although this dependence is usually not made explicit in notation. Denote

$$\overline{c}_{j_e,k_e;D_e}(u, v) := \overline{c}_{j_e,k_e;D_e}\bigl(u, v, U^{(1)}_{j_e|D_e}, \dots, U^{(n)}_{k_e|D_e}\bigr) \tag{4.5}$$

as the oracle pair-copula density estimator that is based on the random samples (4.4). The h-function estimators corresponding to (4.5) are denoted $\overline{h}_{j_e|k_e;D_e}$ and $\overline{h}_{k_e|j_e;D_e}$. The second assumption requires the pair-copula density and h-function estimators to be consistent in this ideal world. For the h-functions we need strong uniform consistency on compact interior subsets of $[0, 1]^2$. We further assume that the errors from h-function estimation vanish faster than n^{-r}.

Assumption 4.2. *For all $e \in E_m, m = 1, \dots, d-1$, it holds:*

(*a*) *for all* $(u, v) \in (0,1)^2$,

$$\bar{c}_{j_e,k_e;D_e}(u, v) - c_{j_e,k_e;D_e}(u, v) = O_p(n^{-r}),$$

(*b*) *for every* $\delta \in (0, 0.5]$,

$$\sup_{(u,v)\in[\delta,1-\delta]^2} \left|\bar{h}_{j_e|k_e;D_e}(u|v) - h_{j_e|k_e;D_e}(u|v)\right| = o_{a.s.}(n^{-r}),$$
$$\sup_{(u,v)\in[\delta,1-\delta]^2} \left|\bar{h}_{k_e|j_e;D_e}(u|v) - h_{k_e|j_e;D_e}(u|v)\right| = o_{a.s.}(n^{-r}).$$

In practice, one has to replace (4.4) by pseudo-observations which have to be estimated. Thus, we only have access to perturbed versions of the random variables (4.4). Similar to a Lipschitz condition, the last assumption ensures that the pair-copula and h-function estimators are not overly sensitive to such perturbations. Denote

$$\widehat{c}_{j_e,k_e;D_e}(u, v) := \bar{c}_{j_e,k_e;D_e}\left(u, v, \widehat{U}^{(1)}_{j_e|D_e}, \dots, \widehat{U}^{(n)}_{k_e|D_e}\right) \tag{4.6}$$

as the estimator based on pseudo-observations $\widehat{U}^{(i)}_{j_e|D_e}, \widehat{U}^{(i)}_{k_e|D_e}$ (as defined in Algorithm 1). The h-function estimators corresponding to (4.6) are denoted $\widehat{h}_{j_e|k_e;D_e}$ and $\widehat{h}_{k_e|j_e;D_e}$, respectively.

Assumption 4.3. *For all $e \in E_m, m = 1, \dots, d-1$, it holds:*

(*a*) *for all* $(u, v) \in (0,1)^2$,

$$\widehat{c}_{j_e,k_e;D_e}(u, v) - \bar{c}_{j_e,k_e;D_e}(u, v) = O_p(a_{e,n}),$$

(*b*) *for every* $\delta \in (0, 0.5]$,

$$\sup_{(u,v)\in[\delta,1-\delta]^2} \left|\widehat{h}_{j_e|k_e;D_e}(u|v) - \bar{h}_{j_e|k_e;D_e}(u|v)\right| = O_{a.s.}(a_{e,n}),$$
$$\sup_{(u,v)\in[\delta,1-\delta]^2} \left|\widehat{h}_{k_e|j_e;D_e}(u|v) - \bar{h}_{k_e|j_e;D_e}(u|v)\right| = O_{a.s.}(a_{e,n}),$$

where

$$a_{e,n} := \sup_{i=1,\dots,n} \left\{\left|\widehat{U}^{(i)}_{j_e|D_e} - U^{(i)}_{j_e|D_e}\right| + \left|\widehat{U}^{(i)}_{k_e|D_e} - U^{(i)}_{k_e|D_e}\right|\right\}.$$

Finally, we require the true pair-copula densities to be smooth. Note that smoothness of pair-copula densities already guarantees smoothness of related h-functions by (2.6).

Assumption 4.4. *For all $e \in E_m$, $m = 1, \dots, d-1$, the pair-copula densities $c_{j_e,k_e;D_e}$ are continuously differentiable on $(0,1)^2$.*

Now we can state our theorem. The proof is deferred to Section 4.8.

Theorem 4.1. *Let f be a d-dimensional density corresponding to a simplified vine distribution with structure $\mathcal{V} = (T_1, \dots, T_{d-1})$ and let $(X_1^{(i)}, \dots, X_d^{(i)}), i = 1, \dots, n$, be iid observations from this density. Denote further $\widehat{f}_{\text{vine}}$ as the estimator from Section 4.2. Under Assumptions 4.1–4.4, it holds for all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$,*

$$\widehat{f}_{\text{vine}}(\boldsymbol{x}) - f(\boldsymbol{x}) = O_p(n^{-r}).$$

Usually, convergence of nonparametric density estimators slows down as dimension increases. This phenomenon is widely known as the curse of dimensionality and restricts the practical application of the estimators to very low-dimensional problems. By Theorem 4.1, the vine copula based density estimator inherits the convergence rate of the bivariate copula density estimator. It does not depend on the dimension d and, therefore, suffers no curse of dimensionality. This is a direct consequence of the simplifying assumption allowing us to subdivide the d-dimensional estimation problem into several one- and two-dimensional tasks.

Assuming that the pair-copula densities are p times continuously differentiable, we can achieve convergence with $r = p/(2p+2)$. Recalling from Stone (1980) that a general nonparametric density estimator has optimal rate $p/(2p+d)$, we see that the vine copula based estimator converges at a rate that is equivalent to the rate of a two-dimensional classical estimator. As this property is independent of dimension, we can expect large benefits of the vine copula approach especially in higher dimensions. We emphasize that a necessary condition for Theorem 4.1 to hold with $r = p/(2p+2)$ is that the density f belongs to the class of simplified vine densities. If this is not the case, the estimator described in Section 4.2 is not consistent, but converges towards a simplified vine density that is merely an approximation of the true density. More specifically, its limit is the *partial vine copula approximation*, first defined in Spanhel and Kurz (2017). In Section 4.5 we will illustrate that even in this situation an estimator based on simplified vine copulas can outperform the classical approach on finite samples.

Remark 4.1. *Theorem 4.1 allows for densities f with arbitrary support. Their support, $\Omega_{\boldsymbol{X}}$, only relates to the marginal distributions; copulas are always supported on a subset of $[0,1]^d$ (zero-density subsets are detected automatically by nonparametric estimators as $n \to \infty$). If some of the X_ℓ have bounded support, we just have to use estimators for $\widehat{f}_\ell$ that takes this into account. This underlines how flexible the vine copula based approach is.*

Remark 4.2. *It is straightforward to extend Theorem 4.1 to non-simplified vine densities by extending the pair-copula densities to functions of more than two* *variables. Besides that, the proof given in Section 4.8 does not make use of the simplifying assumption at all. The simplifying assumption is necessary for $r = p/(2p+2)$ to be feasible. More generally, if we assume that the pair-copulas depend on at most d' conditioning variables, and walk through the steps of the proof, we find that the optimal rate is $p/(2p+2+d')$.*

Remark 4.3. *Theorem 4.1 can be extended to*

$$\sup_{\boldsymbol{x}\in\Omega_{\boldsymbol{X}}} \bigl|\widehat{f}_{\mathrm{vine}}(\boldsymbol{x}) - f(\boldsymbol{x})\bigr| = O_p\bigl\{(\ln n/n)^r\bigr\},$$

provided that the rate n^{-r} in our assumptions is replaced by $(\ln n/n)^r$ and holds uniformly on Ω_{X_ℓ} and $[0,1]^2$ respectively. But this requires that the pair-copula densities are bounded which is unusual. For example, it does not hold when f is a multivariate Gaussian density with non-diagonal covariance matrix. If the assumptions are met, $\widehat{f}_{\mathrm{vine}}$ is able to achieve the optimal uniform rate of a two-dimensional nonparametric density estimator which is attained at $r = p/(2p+2)$ (see, Stone, 1983).

Assumptions 4.1–4.3 are very general and hold for a large class of estimators under mild regularity conditions. In Section 4.4 we validate them for a particular implementation which will be used in the simulations (Section 4.5).

4.3.2 A note on the asymptotic distribution

We also want to give a brief and general account of the asymptotic distribution of the estimator. Let $d^* = d + d(d-1)/2$ and $\widehat{\boldsymbol{f}}^*(\boldsymbol{x}) \in \mathbb{R}^{d^*}$ be the stacked vector of all components of the product $\widehat{f}_{\mathrm{vine}}(\boldsymbol{x})$ in Eq. (4.3), i.e.,

$$\widehat{\boldsymbol{f}}^*(\boldsymbol{x}) := \Bigl(\widehat{f}_1(x_1), \widehat{f}_2(x_2), \dots, \widehat{c}_{j_e,k_e|D_e}\bigl\{\widehat{F}_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}), \widehat{F}_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\bigr\}, \dots\Bigr),$$

and similarly $\boldsymbol{f}^*(\boldsymbol{x})$. Then $\prod_{k=1}^{d^*} \widehat{f}_k^* = \widehat{f}_{\mathrm{vine}}(\boldsymbol{x})$ and $\prod_{k=1}^{d^*} f_k^* = f(\boldsymbol{x})$. The following result is a simple application of the multivariate delta method.

Proposition 4.1. *If for some $\boldsymbol{\mu}_{\boldsymbol{x}} \in \mathbb{R}^{d^*}$, $\Sigma_{\boldsymbol{x}} \in \mathbb{R}^{d^* \times d^*}$,*

$$n^r\bigl\{\widehat{\boldsymbol{f}}^*(\boldsymbol{x}) - \boldsymbol{f}^*(\boldsymbol{x})\bigr\} \xrightarrow{d} \mathcal{N}_{d^*}\bigl(\boldsymbol{\mu}_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}}\bigr), \tag{4.7}$$

then for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$n^r\bigl\{\widehat{f}_{\mathrm{vine}}(\boldsymbol{x}) - f(\boldsymbol{x})\bigr\} \xrightarrow{d} \mathcal{N}_d\bigl(\boldsymbol{\theta}^\top \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\theta}^\top \Sigma_{\boldsymbol{x}} \boldsymbol{\theta}\bigr),$$

where $\theta_k = \prod_{j\neq k} f_j^*(\boldsymbol{x})$, $k = 1, \dots, d^*$.

The standard way to establish the joint normality assumption (4.7) is to check the conditions of the multivariate Lindeberg-Feller central limit theorem (see, van der Vaart, 1998, Proposition 2.27). We will do this for a particular implementation in Section 4.4 (see Proposition 4.5).

4.4 On an implementation as kernel estimator

So far we did not specify how the marginal densities, pair-copula densities, and h-functions should be estimated. In general, we can do this parametrically, semi-parametrically, or nonparametrically and tap into the full potential of existing methods. In this section, we discuss a particular implementation as a kernel estimator. We give low-level conditions under which the assumptions of Theorem 4.1 can be verified. We present corresponding consistency results and establish asymptotic normality of $\widehat{f}_{\mathrm{vine}}$. Similar results could be obtained for other implementations. Another issue is that we assumed the structure of the vine to be known. Some heuristics to select an appropriate vine structure are discussed at the end of this section.

4.4.1 Estimation of marginal densities and distribution functions

Univariate kernel density and distribution function estimators have been extensively studied in the literature. To this day, they are most popular in their original form (Rosenblatt, 1956, Parzen, 1962): for all $x \in \mathbb{R}$,

$$\widehat{f}_\ell(x) = \frac{1}{nb_n}\sum_{i=1}^n K\biggl(\frac{X_\ell^{(i)} - x}{b_n}\biggr), \quad \widehat{F}_\ell(x) = \frac{1}{n}\sum_{i=1}^n J\biggl(\frac{X_\ell^{(i)} - x}{b_n}\biggr), \tag{4.8}$$

where $b_n > 0$ is the bandwidth parameter, K is a kernel function and $J(x) = \int_{-\infty}^x K(s)ds$ the integrated kernel. We impose the following assumptions on the kernel function, bandwidth sequence, and marginal distributions.

Assumption 4.5.

K1 The kernel function K is a symmetric probability density function supported on $[-1, 1]$ and has continuous first-order derivative.

K2 The bandwidth sequence satisfies $b_n \to 0$ and $nb_n^4/\ln n \to \infty$.

Assumption 4.6.

M1 For all $\ell = 1\,\ldots, d$, f_ℓ is strictly positive on $\mathbb{R}$ and has uniformly continuous second-order derivative.

K1 is a common technical condition satisfied by many popular kernels (e.g., Epanechnikov and cosine kernels), but it excludes the Gaussian kernel. The restriction to kernels with bounded support can usually be relaxed, but brings additional difficulties in the proofs.

The following result gives the rate of strong uniform consistency for $\widehat{f}_\ell$.

Proposition 4.2. *Under Assumptions 4.5 and 4.6, the estimator* (4.8) *satisfies*

$$\sup_{x\in\mathbb{R}}\bigl|\widehat{f}_\ell(x) - f_\ell(x)\bigr| = O_{a.s.}\bigl(b_n^2 + \sqrt{\ln n/(nb_n)}\bigr).$$

for all $\ell = 1\,\ldots, d$.

Proof. A standard result for kernel density estimation (see, e.g., Scott, 2008, Section 6.2.1) is

$$\mathrm{E}\bigl\{\widehat{f}_\ell(x)\bigr\} - f_\ell(x) = \frac{1}{2}b_n^2\sigma_K^2\frac{\partial^2}{\partial x^2}f_\ell(x) + o(b_n^2),$$

where $\sigma_K^2 = \int_{[-1,1]} x^2K(x)dx < \infty$ by K1 and $\partial^2/\partial x^2 f_\ell(x)$ is bounded by M1. The claim then follows from Giné and Guillou (2002, Theorem 2.3) which states

$$\sup_{x\in\mathbb{R}}\bigl|\widehat{f}_\ell(x) - \mathrm{E}\bigl\{\widehat{f}_\ell(x)\bigr\}\bigr| = O_{a.s.}\bigl(\sqrt{\ln n/(nb_n)}\bigr).$$ □

Proposition 4.2 implies pointwise weak consistency of $\widehat{f}_\ell$ as well as strong uniform consistency of $\widehat{F}_\ell$ with the same rate. In both cases the rate could be improved, but the result will be sufficient for our purposes. The mean-square optimal bandwidth for $\widehat{f}_\ell$ is $b_n = O(n^{-1/5})$ for which Proposition 4.2 holds with rate $O_{a.s.}(n^{-2/5}\sqrt{\ln n})$.

Extensions of the above estimator comprise variable bandwidth methods (Sain and Scott, 1996), transformation techniques for heavy-tailed distributions (Bolancé et al., 2008), and boundary kernel estimators that avoid bias and consistency issues on bounded support (Bouezmarni and Rombouts, 2010).

4.4.2 Estimation of pair-copula densities

Nonparametric estimation of copula densities requires caution because they are supported on the unit hypercube. An estimator that takes no account of this property will suffer from bias issues at the boundaries of the support. A few kernel estimators particularly suited for bivariate copula densities were proposed in the literature (Gijbels and Mielniczuk, 1990, Charpentier et al., 2006, Geenens et al., 2017). Other nonparametric estimators can be constructed based on Bernstein polynomials (Sancetta and Satchell, 2004), B-splines (Kauermann et al., 2013), or wavelets (Genest et al., 2009).

We shall focus on the local-constant transformation estimator of Geenens et al. (2017), explained in more detail in Section 3.2.1. Using the same notation, the estimator can be written as

$$\overline{c}_{j_e,k_e;D_e}(u,v) = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{K}\left\{B_n^{-1}\begin{pmatrix}\Phi^{-1}(u)-\Phi^{-1}(U^{(i)}_{j_e|D_e}) \\ \Phi^{-1}(v)-\Phi^{-1}(V^{(i)}_{k_e|D_e})\end{pmatrix}\right\}}{\phi\{\Phi^{-1}(u)\}\phi\{\Phi^{-1}(v)\}}. \tag{4.9}$$

In order to verify the high-level assumptions 4.2a and 4.3a, we need the following conditions to hold for all $e \in E_1, \ldots, E_{d-1}$:

Assumption 4.7.

C1 The true pair-copula densities $c_{j_e,k_e;D_e}$ are twice continuously differentiable on $(0,1)^2$.

C2 The transformed densities $\psi_{j_e,k_e;D_e}(x,y) = c_{j_e,k_e;D_e}\{\Phi(x),\Phi(v)\}\phi(x)\phi(y)$ have continuous and bounded first- and second-order derivatives on $\mathbb{R}^2$.

C1 is a smoothness condition that is very common in nonparametric estimation. C2 is less standard as it relates to the transformed density. Sufficient conditions for C2 are given in Lemma A.1 of (Geenens et al., 2017) and can be verified for many parametric families, including the ones used in our simulation study.

To avoid unnecessary technicality, we will assume here that the bandwidth matrix is a multiple of the identity matrix: $B_n = b_n \times I_2$.

Proposition 4.3. *Under Assumptions 4.5 and 4.7, the estimator (4.9) satisfies for all $(u,v) \in (0,1)^2$, $e \in E_1, \ldots, E_m$,*

$$\begin{aligned}\overline{c}_{j_e,k_e;D_e}(u,v) - c_{j_e,k_e;D_e}(u,v) &= O_p\bigl(b_n^2 + \sqrt{1/(nb_n^2)}\bigr), \\ \widehat{c}_{j_e,k_e;D_e}(u,v) - \overline{c}_{j_e,k_e;D_e}(u,v) &= O_p(a_{e,n}).\end{aligned}$$

Proof. For the first equality, see Corollary 3.4 of Nagler (2014). For the second, see Lemma 4.1 in Section 4.8.2. □

When the mean-square optimal bandwidth $b_n = O(n^{-1/6})$ is used, the right hand side of the first equality is $O_p\bigl(n^{-1/3}\bigr)$.

4.4.3 Estimation of h-functions

Recall that h-functions are actually conditional distribution functions:

$$h_{j_e|k_e;D_e}(u|v) = \Pr(U_{j_e|D_e} \leq u \mid U_{k_e|D_e} = v) = \mathrm{E}\bigl\{\mathbb{1}(U_{j_e|D_e} \leq u) \mid U_{k_e|D_e} = v\bigr\}.$$

The second equality relates the conditional *cdf* to a regression problem. Hence, any nonparametric regression estimator is suitable for estimation of the h-functions.

In our case, it is even simpler to integrate the density estimate to obtain an estimate of the corresponding h-function: for the oracle estimators,

$$\begin{aligned} \overline{h}_{j_e|k_e;D_e}(u|v) &:= \int_0^u \overline{c}_{k_e,j_e;D_e}(s,v)ds, \\ \overline{h}_{k_e|j_e;D_e}(v|u) &:= \int_0^v \overline{c}_{j_e,k_e;D_e}(u,s)ds, \end{aligned} \tag{4.10}$$

and the feasible estimators $\widehat{h}_{j_e|k_e;D_e}$ and $\widehat{h}_{k_e|j_e;D_e}$ are defined similarly. Such estimators are closely related to the smoothed Nadaraya-Watson conditional estimator of the conditional distribution (Hansen, 2004). In fact, they coincide when we choose diagonal B_n in (4.9). For an explicit formula, see (4.18) in Section 4.8.2. The following result puts this estimator in the context of 4.2b and 4.3b.

Proposition 4.4. *Under Assumptions 4.5 and 4.7, the estimator defined by (4.10) and (4.9) satisfies for all $\delta \in (0, 0.5]$, and $e \in E_1, \dots, E_m$,*

$$\begin{aligned} \sup_{(u,v)\in[\delta,1-\delta]^2} \left|\overline{h}_{j_e|k_e;D_e}(u|v) - h_{j_e|k_e;D_e}(u|v)\right| &= O_{a.s.}\big(b_n^2 + \sqrt{\ln n/(nb_n)}\big), \\ \sup_{(u,v)\in[\delta,1-\delta]^2} \left|\overline{h}_{k_e|j_e;D_e}(u|v) - h_{k_e|j_e;D_e}(u|v)\right| &= O_{a.s.}\big(b_n^2 + \sqrt{\ln n/(nb_n)}\big), \\ \sup_{(u,v)\in[\delta,1-\delta]^2} \left|\widehat{h}_{j_e|k_e;D_e}(u|v) - \overline{h}_{j_e|k_e;D_e}(u|v)\right| &= O_{a.s.}\big(a_{e,n}\big), \\ \sup_{(u,v)\in[\delta,1-\delta]^2} \left|\widehat{h}_{k_e|j_e;D_e}(u|v) - \overline{h}_{k_e|j_e;D_e}(u|v)\right| &= O_{a.s.}\big(a_{e,n}\big). \end{aligned}$$

Proof. See Lemmas 4.2 and 4.3 in Section 4.8.2. □

The optimal rate of convergence in the first two equalities is $O_{a.s.}\big\{(\ln n/n)^{2/5}\big\}$ and attained for $b_n = O\big\{(\ln n/n)^{1/5}\big\}$.

Assumption 4.2b requires that the error of estimating the h-function vanishes faster than the error of pair-copula density estimation. Because conditional distributions can be estimated at a faster rate than joint densities, this is readily achieved by using mean-square optimal bandwidths in each component. However, it may be more convenient to use the same bandwidth for pair-copula density as well as h-function estimation. It seems natural to use the mean-square optimal rate for pair-copula density estimation, $b_n \sim n^{-1/6}$. But this violates Assumption 4.2, because both estimators converge with the same rate: $n^{-1/3}$. To overcome this, we have to increase the speed of b_n by a small amount, i.e., to undersmooth the pair-copula density estimate. When $b_n = \alpha_n n^{-1/6}$, $\alpha_n = o(1)$, the pair-copula density estimators converges with rate $\alpha_n^{-1}n^{-1/3}$ and the h-function estimator with (almost sure) rate $\alpha_n^2 n^{-1/3} + \sqrt{\ln n}\alpha_n^{-1/2}n^{-5/12} = o(\alpha_n^{-1}n^{-1/3})$. But the sequence α_n can converge arbitrarily slow. So we should not expect any problems with using the mean-square optimal rate $b_n = n^{-1/6}$ in practice. This was confirmed by preliminary numerical experiments.

4.4.4 Asymptotic normality

We now put all pieces together and show that the estimator $\widehat{f}_{\text{vine}}$ composed of (4.8), (4.9), and (4.10) is asymptotically normal. We start by establishing the joint asymptotic normality of all components. The proof is deferred to Section 4.8.3.

Proposition 4.5. *Assume that*

(i) Assumptions 4.5, 4.6, and 4.7 hold,

(ii) $\widehat{f}_\ell$ and $\widehat{F}_\ell$ are defined by (4.8) with (marginal) bandwidth parameter $b_{n,m}$,

(iii) $\widehat{c}_{j_e,k_e;D_e}$ are defined by (4.9) with (copula) bandwidth parameter $b_{n,c}$,

(iv) $\widehat{h}_{j_e|k_e;D_e}$ and $\widehat{h}_{j_e|k_e;D_e}$ are defined by (4.10) and (4.9) with (h-function) bandwidth parameter $b_{n,h}$,

(v) it holds $b_{n,c} = O(n^{-1/6})$, and for sufficiently large n,

$$b_{n,c}^2 < b_{n,m} \leq b_{n,h} \leq \min\{b_{n,c}, n^{-1/6}/\log n\}.$$

Recall the definition of $\widehat{\boldsymbol{f}}^*(\boldsymbol{x})$, $\boldsymbol{f}^*(\boldsymbol{x})$, and d^* from Section 4.3.2. It holds for all $\boldsymbol{x} \in R^d$,

$$(nb_{n,c}^2)^{1/2}\big\{\widehat{\boldsymbol{f}}^*(\boldsymbol{x}) - b_{n,c}^2\boldsymbol{\mu}_{\boldsymbol{x}} - \boldsymbol{f}^*(\boldsymbol{x})\big\} \xrightarrow{d} \mathcal{N}_{d^*}\big(0, \Sigma_{\boldsymbol{x}}\big), \tag{4.11}$$

where $\boldsymbol{\mu}_{\boldsymbol{x}} = (\boldsymbol{0}_d^\top, \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}^\top)^\top$, $\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} = (\tilde{\mu}_{\boldsymbol{x},e})_{e \in E_1,\dots,E_{d-1}}$, and $\Sigma_{\boldsymbol{x}}$ is diagonal with first d diagonal entries equal to 0 and remaining diagonal entries $(\tilde{\sigma}_{\boldsymbol{x},e})_{e \in E_1,\dots,E_{d-1}}$. Explicit expressions for $\tilde{\mu}_{\boldsymbol{x},e}$ and $\tilde{\sigma}_{\boldsymbol{x},e}$ are given in (4.26) and (4.28) in Section 4.8.3.

The asymptotic normality of $\widehat{f}_{\text{vine}}$ follows from an application of the delta method.

Corollary 4.1. *Under the assumptions of Proposition 4.5 it holds for all $\boldsymbol{x} \in \mathbb{R}^d$,*

$$(nb_{n,c}^2)^{1/2}\big\{\widehat{f}_{\text{vine}}(\boldsymbol{x}) - b_{n,c}^2\boldsymbol{\theta}^\top\boldsymbol{\mu}_{\boldsymbol{x}} - f(\boldsymbol{x})\big\} \xrightarrow{d} \mathcal{N}\big(0, \boldsymbol{\theta}^\top\Sigma_{\boldsymbol{x}}\boldsymbol{\theta}\big),$$

where $\theta_k = \prod_{j \neq k} f_j^*(\boldsymbol{x})$, $k = 1, \dots, d^*$, and $\boldsymbol{\mu}_{\boldsymbol{x}}$, $\Sigma_{\boldsymbol{x}}$ are as in Proposition 4.5.

4.5 Simulations

In this section, we study the finite sample behavior of the vine copula based kernel density estimator. We illustrate its advantages compared with the classical kernel density estimator in three scenarios that comprise one simplified and two non-simplified target densities.

4.5.1 Implementation of estimators

The study was carried out in the statistical computing environment R (R Core Team, 2014). We use the implementation of $\widehat{f}_{\mathrm{vine}}$ introduced in the previous section:

Marginal densities are estimated by the standard kernel density estimator (4.8) with the Epanechnikov kernel. Bandwidths are selected by the plug-in method of Chacón and Duong (2010), as implemented in the function `hpi` of the `ks` package (Duong, 2014).

Marginal distributions are estimated by integrating the estimates of the marginal densities.

Pair-copula densities are estimated by the transformation estimator (4.9) with bandwidth matrix selected by the normal reference rule (see, Nagler, 2014, Section 3.4.4).

The vine structure is considered unknown and selected by the method of Dißmann et al. (2013) using empirical estimates of τ_e as weight function (see Section 3.2.3).

The estimator $\widehat{f}_{\mathrm{vine}}$ is implemented in the R package `kdevine` (Nagler, 2017). The package also includes estimators for marginals with bounded support as well as more sophisticated pair-copula estimators which further improve the performance. For the classical multivariate kernel density estimator ($\widehat{f}_{\mathrm{mvkde}}$ from here on) we use the function `kde` provided by the `ks` package (Duong, 2014). It selects the bandwidths by the plug-in method of Chacón and Duong (2010).

4.5.2 Performance measurement

To assess the convergence behavior of the estimators under increasing dimension, we consider five different sample sizes $n = 200, 500, 1\,000, 2\,500, 5\,000$, and three different dimensions $d = 3, 5, 10$. For any fixed target density, sample size, and dimension, we measure the performance as follows:

1. Simulate $n_{sim} = 250$ samples of size n, from a d-dimensional target density f.
2. On each sample, estimate the density with estimators $\widehat{f}_{\mathrm{vine}}$ and $\widehat{f}_{\mathrm{mvkde}}$.
3. For each estimator $\widehat{f} \in \{\widehat{f}_{\mathrm{vine}}, \widehat{f}_{\mathrm{mvkde}}\}$ and sample, calculate the *integrated absolute error (IAE)* as a performance measure:

$$\mathrm{IAE}\big(\widehat{f}\big) := \int_{\mathbb{R}^d} \big|\widehat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\big| d\boldsymbol{x}.$$

   The integral is estimated by importance sampling Monte Carlo (see Section 3.3.3), where we take the true density f as the sampling distribution. The number of Monte Carlo samples was set to 1 000.

In the following section we will present the median IAE attained over 250 simulations.

4.5.3 Results

In the following, we illustrate the main insights of our numerical experiments in three scenarios — one where the simplifying assumption holds, and two where it does not. Since the simplifying assumption is a property of the copula, we focus on this part and set the marginal densities to standard Gaussian in all scenarios. For these margins, the two estimators $\widehat{f}_{\text{vine}}$ and $\widehat{f}_{\text{mvkde}}$ are asymptotically equivalent when $d = 2$. But they become different as soon as the simplifying assumption becomes relevant, i.e., when $d > 2$. Hence, differences in the performance of the two estimators can be directly related to the fact that $\widehat{f}_{\text{vine}}$ assumes a simplified model.

Scenario 1: simplified models

The first scenario concerns the estimation of a multivariate Gaussian density. For simplicity, we choose the model parameters such that all pair-wise Kendall's τ are equal. Recall that the simplifying assumption is a property of the dependence, i.e. the copula. The copula underlying a multivariate Gaussian density is the Gaussian copula which belongs to the class of simplified vine distributions (Stöber et al., 2013). Consequently, the vine copula based estimator is consistent in this situation.

Figure 4.1 shows the median IAE of $\widehat{f}_{\text{vine}}$ (circles) and $\widehat{f}_{\text{mvkde}}$ (triangles) for varying sample size n and dimension d. The vine copula based estimator strictly outperforms the classical estimator by a considerable margin. As predicted by Theorem 4.1, we observe that — in contrast to the classical kernel density estimator — the vine copula based estimator converges at the same rate independent of dimension. Thus, the gap widens as dimension or sample size increase. For $d = 5$, $\widehat{f}_{\text{vine}}$ is almost two times as accurate; for $d = 10$ almost three times as accurate. These numbers are remarkable considering how slowly $\widehat{f}_{\text{mvkde}}$ can improve its accuracy when increasing sample size.

The same conclusions can be drawn when the data originate from a Clayton copula (see Figure 4.2), which also satisfies the simplifying assumption (Stöber et al., 2013).

Scenario 2: mild violations of the simplifying assumption

Our second scenario (Gumbel or Frank copulas), violates the simplifying assumption; see Stöber et al. (2013, Theorem 3.1). Again, we choose the parameters such that all pair-wise Kendall's τ are equal. In this case, $\widehat{f}_{\text{mvkde}}$ is guaranteed to outperform $\widehat{f}_{\text{vine}}$ as $n \to \infty$, because the latter is not consistent. On finite samples, however, the picture seems to be different.

The performance of the two estimators in this scenario is displayed in Figure 4.3 and Figure 4.4. For $d = 3$, $\widehat{f}_{\text{vine}}$ is slightly worse than its competitor, but the

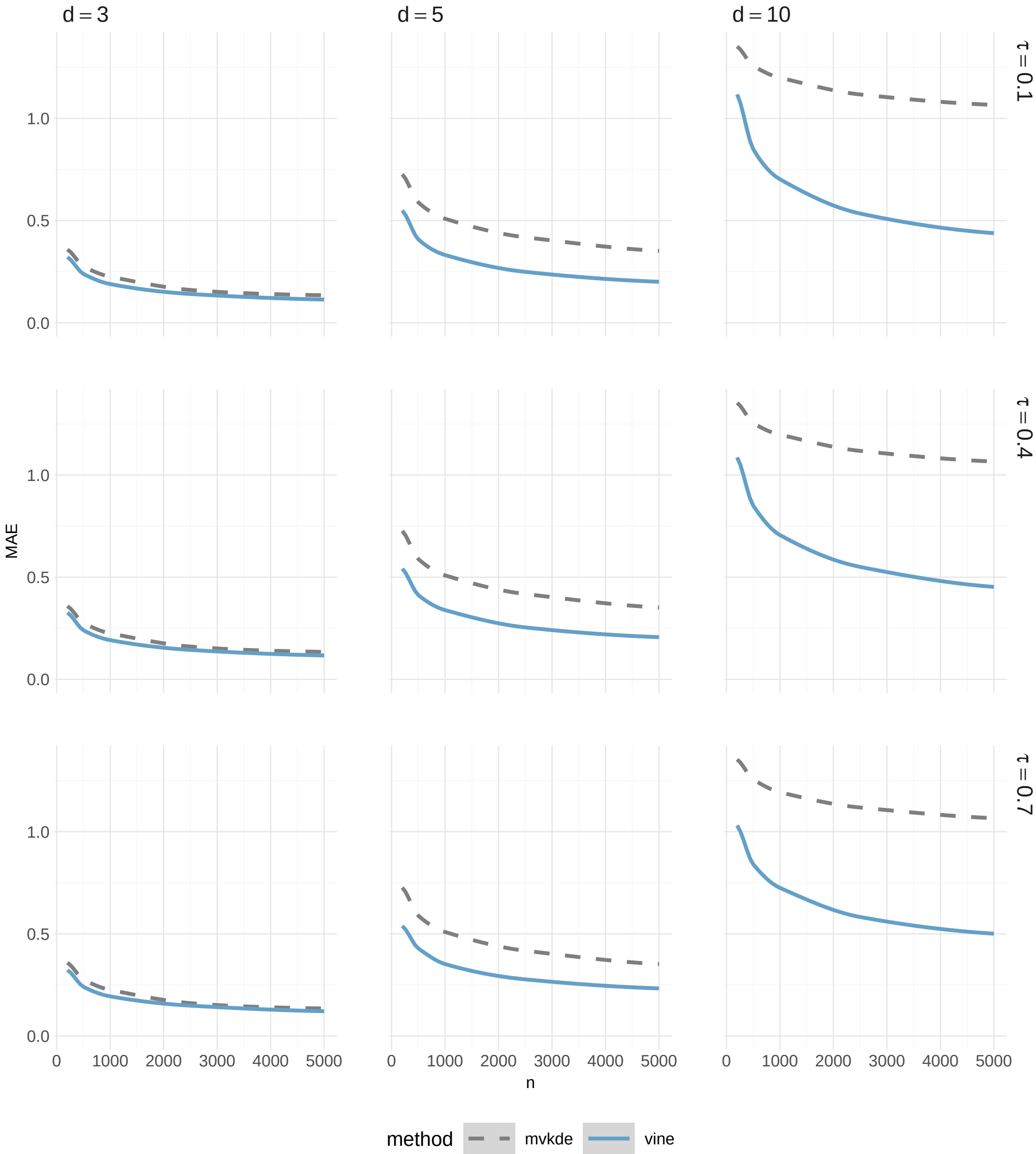


Figure 4.1: Gaussian copula: Median integrated absolute error achieved by the two estimators for varying sample size n, dimension d, and Kendall's τ.

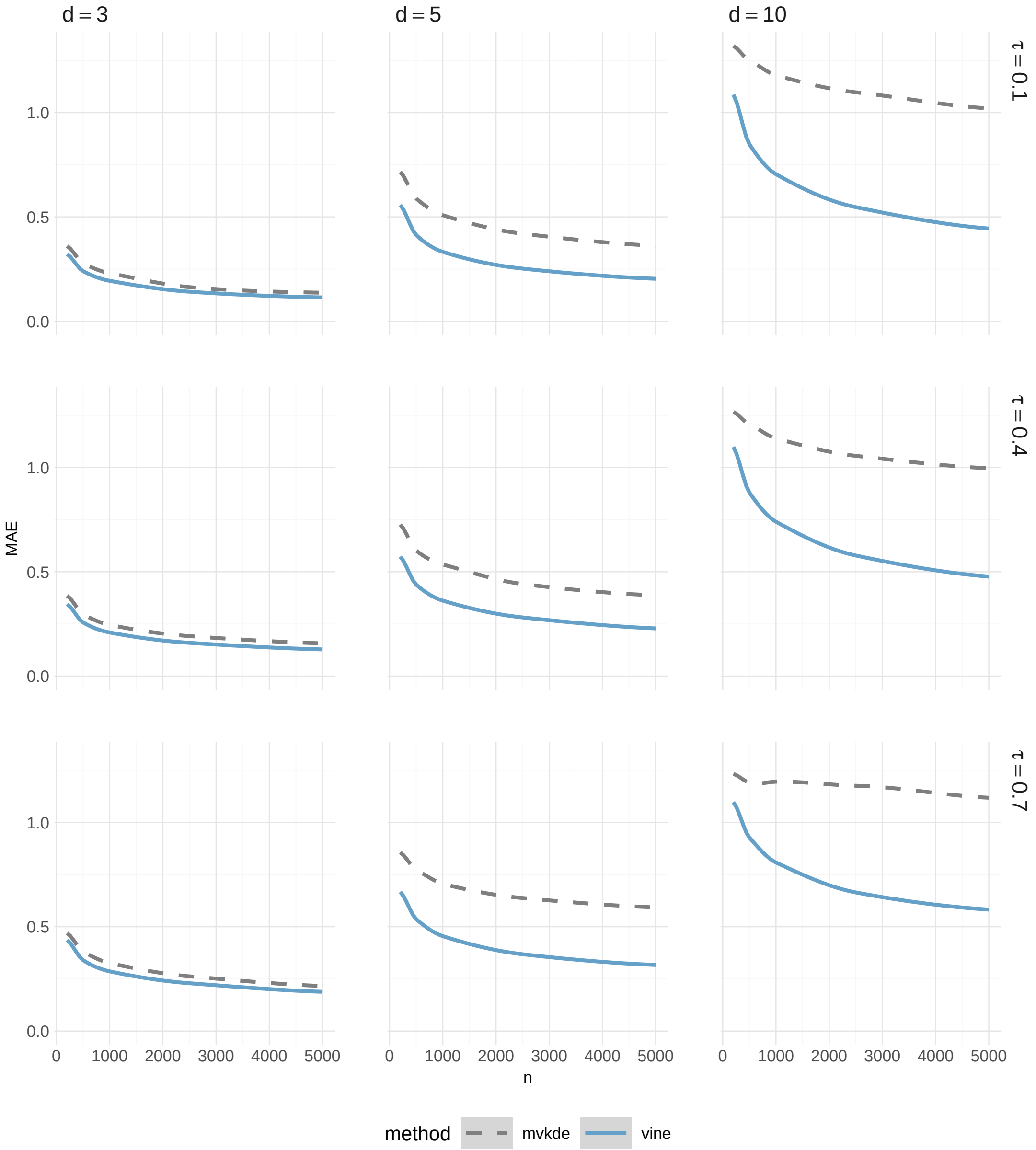


Figure 4.2: Clayton copula: Median integrated absolute error achieved by the two estimators for varying sample size n, dimension d, and Kendall's τ.

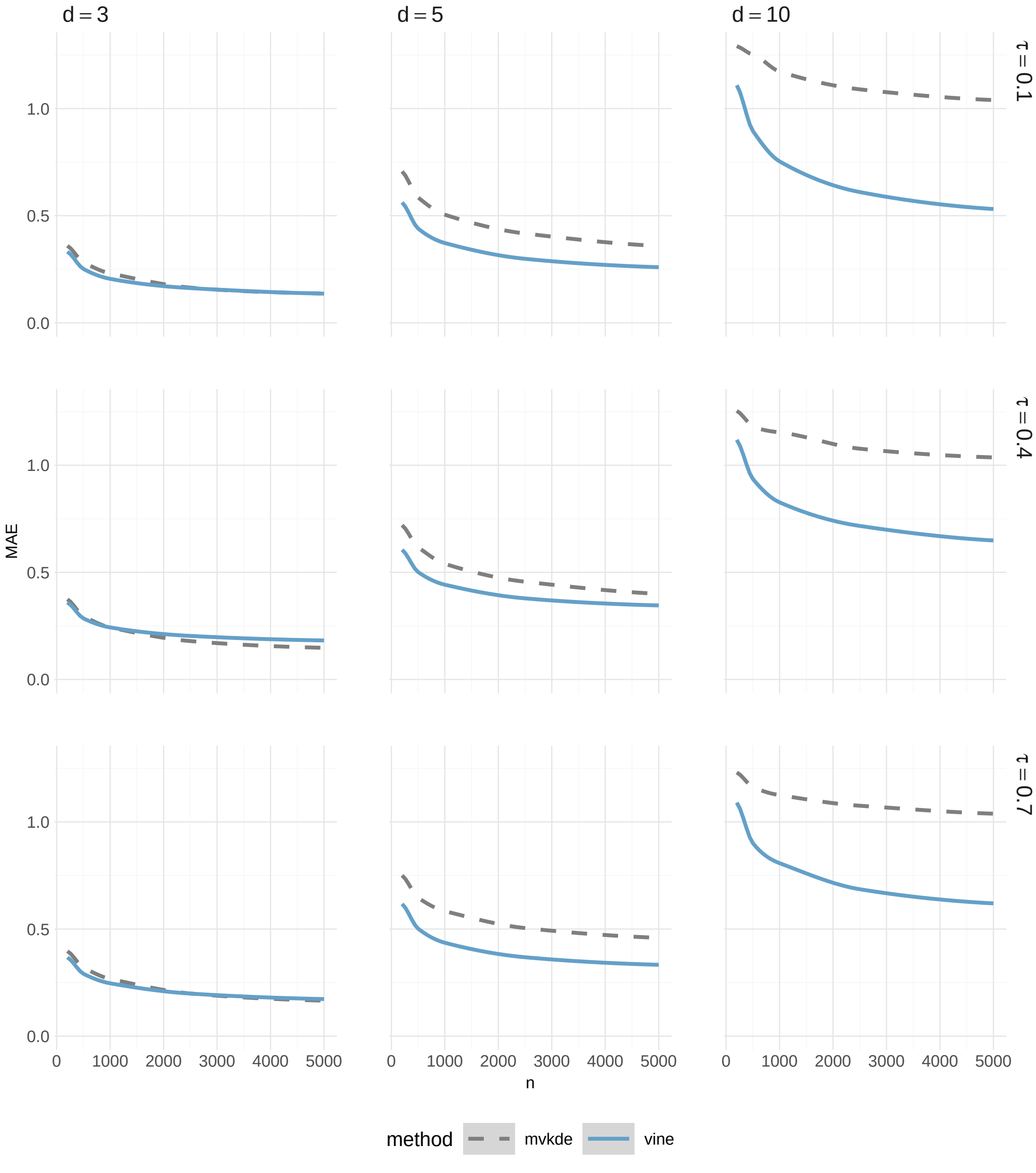


Figure 4.3: Gumbel copula: Median integrated absolute error achieved by the two estimators for varying sample size n, dimension d, and Kendall's τ.

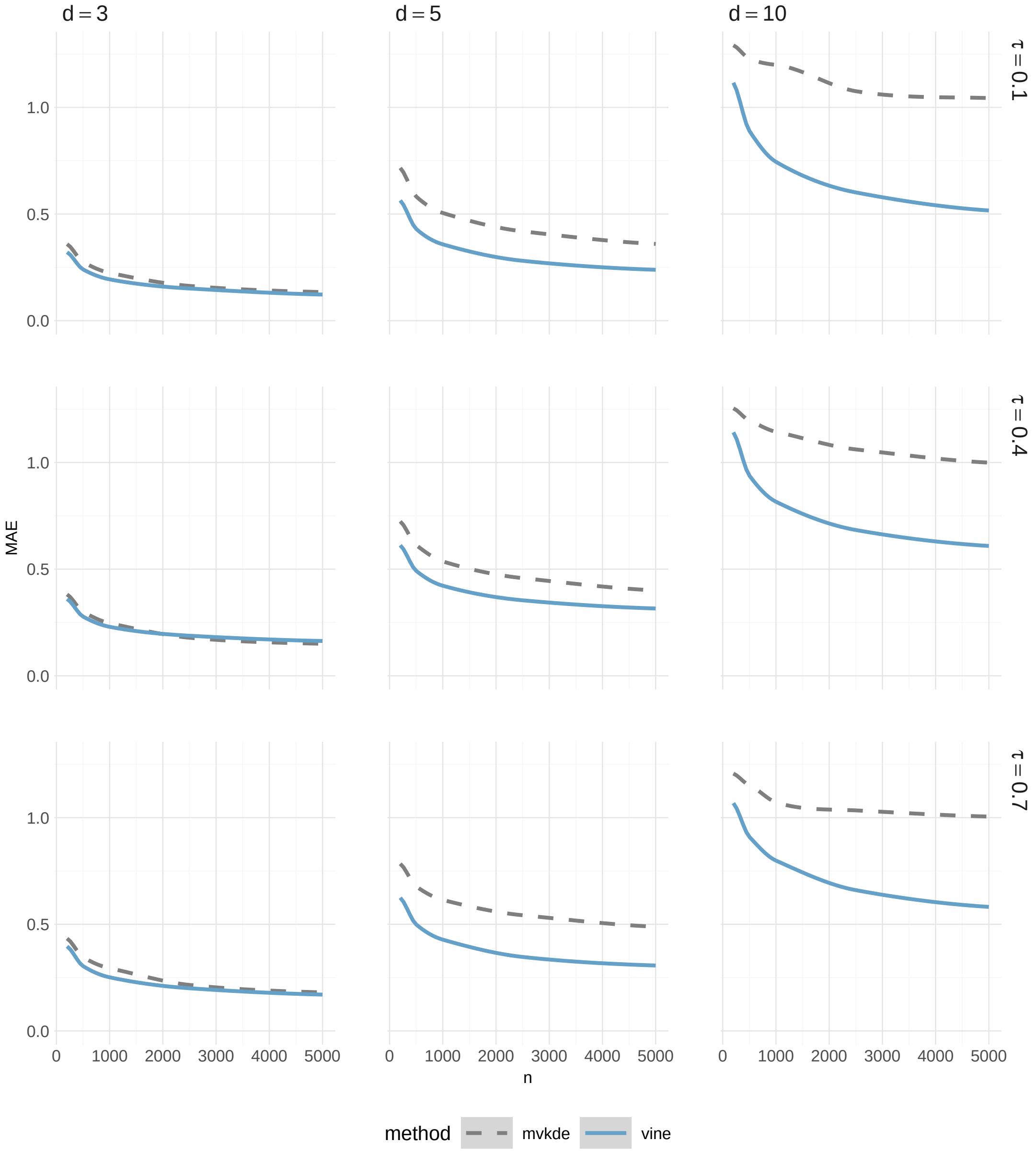


Figure 4.4: Frank copula: Median integrated absolute error achieved by the two estimators for varying sample size n, dimension d, and Kendall's τ.

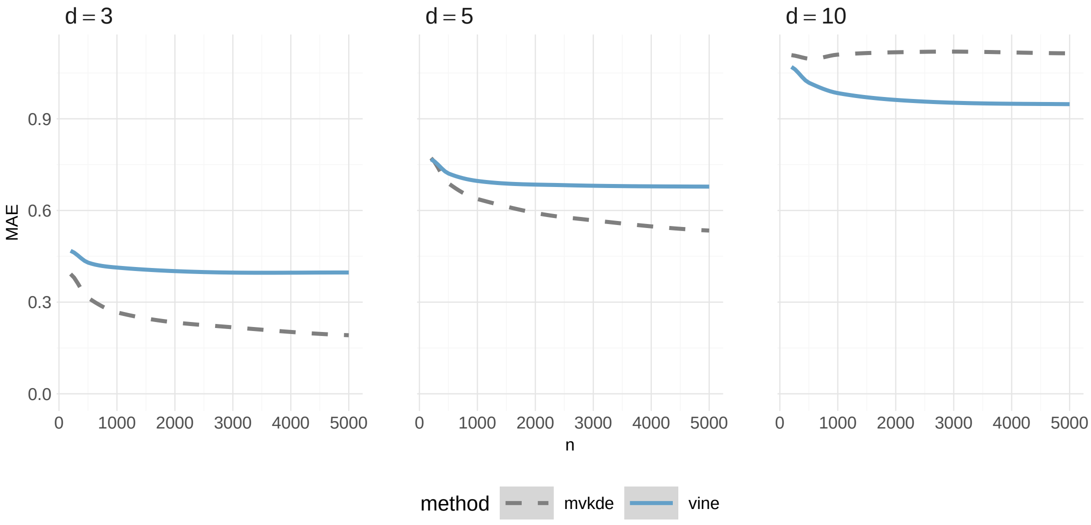


Figure 4.5: Non-simplified Gaussian copula: Median integrated absolute error achieved by the two estimators for varying sample size n, dimension d, and Kendall's τ.

difference is only significant for large sample sizes. For increasing dimension, the gap widens in favor of $\widehat{f}_{\text{vine}}$ which performs significantly better for $d = 5$ and $d = 10$. For $d = 10$ and $n = 5\,000$, the vine copula based estimator is almost two times as accurate — although it is not consistent. Since $\widehat{f}_{\text{mvkde}}$ converges so slowly, an extremely large number of observations would be required until it becomes the better choice. But for commonly available sample sizes and $d > 3$, the vine copula based estimator seems preferable.

Scenario 3: severe violation of the simplifying assumption

Lastly, we want to investigate how the vine copula based estimator behaves in a sort of 'worst case scenario'. We set up a non-simplified vine copula with Gaussian pair-copulas and formulate their parameters as a regression on the conditioning variables implied by the vine. For each conditional pair-copula, the correlation parameter function $\rho_e\colon [0,1]^{|D_e|} \to [-1,1]$ describes a linear hyperplane ranging from -1 to 1:

$$\rho_e(\boldsymbol{u}_{D_e}) = 1 - \frac{2}{|D_e|}\sum_{j \in D_e} u_j, \qquad \text{for } e \in E_m,\ m \geq 2.$$

Since $\int \rho_e(\boldsymbol{u}_{D_e}) d\boldsymbol{u}_{D_e} = 0$ for all $e \in E_2, \dots, E_{d-1}$, we also set $\rho_e \equiv 0$ for $e \in E_1$. This model is severely violating the simplifying assumption for each conditional pair in the vine.

The results for this scenario are shown in Figure 4.5. The vine copula based estimator performs significantly worse for $d = 3, 5$. Remarkably, $\widehat{f}_{\text{vine}}$ manages to significantly outperform the classical estimator for $d = 10$. The severely non-simplified dependence structure appears to be too difficult to identify even for a nonparametric estimator that does not rely on the simplifying assumption. Extrapolating the curves, we can expect that to hold for sample sizes much larger than those considered in our study. Also, we can expect the advantage of $\widehat{f}_{\text{vine}}$ to become even bigger in higher dimensions. We can conclude that even in this extremely unfavorable example, the estimator $\widehat{f}_{\text{vine}}$ proves useful when more than a few variables are involved.

4.6 Application: filtering noise in telescope imaging

We revisit a classification problem from astrophysics which has previously been investigated in Bock et al. (2004). In their study, the authors consider synthetic data imitating measurements taken on images from the *MAGIC (Major Atmospheric Gamma-ray Imaging Cherenkov) Telescopes* located on the Canary islands. The goal is to identify primary gamma rays (the signal) amongst a large amount of hadron showers (background noise). The authors of the study evaluate the performance of several classification methods and judge the kernel density based Bayes classifier as one of the most convincing. We aim to augment their results and investigate how the vine copula based kernel density estimator performs on this problem.

The data set is available from the *UCI Machine Learning Repository* web page (url: `https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope`) and consists of $n = 19\,020$ observations on $d = 10$ variables. $n_G = 12\,332$ of the observations are classified as gamma (signal) and $n_H = 6\,688$ as hadron (background). A subset of the data is illustrated in Figure 4.6. There is no clear separation between the two classes, although both margins and dependence appear to be slightly different. We also observe that some pairs (e.g., fConc – FM3Long) have non-monotonic dependence patterns. Such patterns are hard to capture by classical parametric models. Mixtures of parametric models could solve this problem, but require more careful modeling. For more information on the astrophysical background and a more thorough description of the data we refer the reader to Bock et al. (2004) and the UCI web page.

Bayes classifiers follow the idea of maximizing the posterior probability of a class given the data. Let G (for gamma) and H (for hadron) be the two classes and $\widehat{f}_G$ and $\widehat{f}_H$ be two estimates fitted separately in each class. Assume further we have knowledge of the class prior probabilities π_G, π_H. With a straightforward application of Bayes' theorem, we can estimate the posterior probability that the class is G as

$$\widehat{\mathrm{P}}(\text{Class} = G \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_G \widehat{f}_G(\boldsymbol{x})}{\pi_G \widehat{f}_G(\boldsymbol{x}) + \pi_H \widehat{f}_H(\boldsymbol{x})}, \tag{4.12}$$

where $\boldsymbol{x}$ is a realization of the random vector $\boldsymbol{X}$. In the most general case, we classify an observation as G whenever the estimated posterior probability is

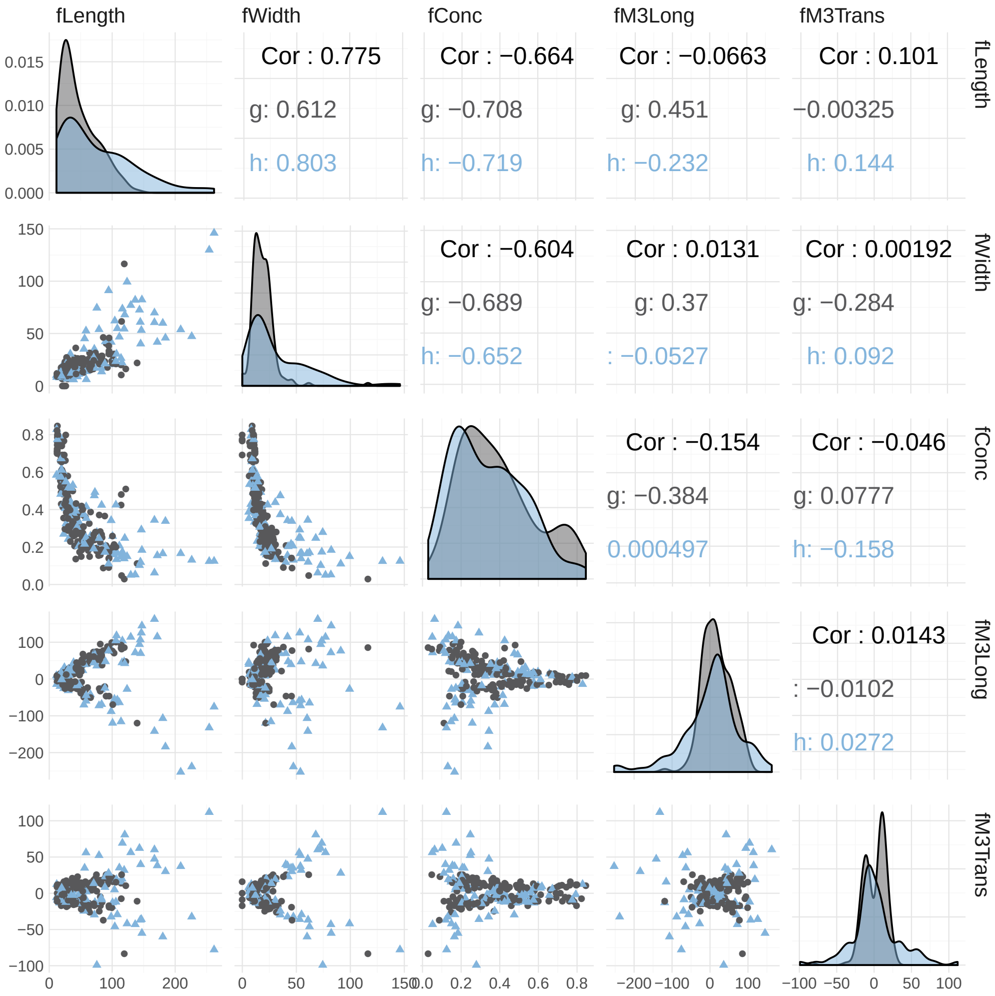


Figure 4.6: Illustration of five variables in the MAGIC data set. Lower triangle: pair-wise scatter plots (gamma: dots, hadron: triangles); diagonal: kernel density estimates per class; upper triangle: overall and class-wise correlation.

greater than $\alpha = 0.5$. However, by changing the threshold α we can furthermore control how many observations get classified as G, and thereby influence key quantities such as the *false positive rate (FPR)* or *true positive rate (TPR)*. The FPR is defined as the ratio of the number of false positives (here: hadron events that were misclassified as gamma) and the number of all negative (hadron) events. The TPR is defined as the ratio of the number of correctly classified positive (gamma) events and the number of all positive events. In general, it is desirable to have a low FPR and a high TPR. But usually, there is a tradeoff between the two quantities: If we increase the threshold level α, a higher posterior probability is required for an observation to get classified as gamma event. As a result, less observations will be classified as gamma event, which in turn reduces *both* FPR and TPR.

We repeat the experiment of Bock et al. (2004) with the vine copula based and

FPR	0.01	0.02	0.05	0.1	0.2
vine	0.335	0.428	0.652	0.780	0.918
mvkde	0.335	0.408	0.567	0.730	0.868

Table 4.1: True positive rates for the two estimators (second and third row) for given target levels of the false positive rate (first row).

classical kernel estimators. The implementations are similar to our simulation study (see Section 4.5.1). As is common in applications, we induce sparsity of the estimated vine copula model by adding an independence test to the structure selection algorithm(see, Dißmann et al., 2013, Section 4). We also found it necessary to multiply the marginal bandwidth parameters of $\widehat{f}_{\text{vine}}$ by 2 to stabilize the classification boundary in low-density regions. The experiment's setup is the following: First, the densities for each class are estimated on the first 2/3 of the data which is used as training set. These estimates are used in combination with (4.12) to obtain class predictions for the remaining 1/3. For simplicity, the prior probabilities are set to $\pi_G = \pi_H = 0.5$. The predictions are then compared to the actual class of the observations which allows to assess the quality of the predictions.

Bock et al. (2004) stressed that in this application the focus is on low FPR levels; in particular the 0.01, 0.02, 0.05, 0.1 and 0.2 levels. The corresponding TPR values at these levels are displayed in Table 4.1. The were found by obtaining predictions for a fine grid of values for the threshold α and then interpolating the corresponding TPR values.

The TPR values of the vine copula based estimator are uniformly larger than the ones of the classical multivariate kernel density estimator. This means that for a target FPR level, the vine copula based classifier is able to identify more observations correctly as signal events than the classical multivariate kernel density estimator. The results confirm what we could expect from our simulation study where, for $d = 10$ and several thousand observations, the vine copula based approach delivered much more accurate estimates.

But also in comparison with other classification algorithms, the classifier based on $\widehat{f}_{\text{vine}}$ performs extraordinary well. A total of 14 algorithms were surveyed by Bock et al. (2004), including variants of classification trees, neural networks, support vector machines, and nearest-neighbor methods. Two of the main performance measures used in their study are the average of the TPR at the 0.01, 0.02 and 0.05 FPR levels (termed *loacc*), and the average of the TPR at the 0.1 and 0.2 FPR levels (termed *highacc*). From Table 4.1 we calculate $loacc = 0.472$ and $highacc = 0.849$. None of the 14 algorithms was able to produce a better *loacc* value than our approach. And only one method, random forests, delivered a slightly higher *highacc* of 0.852. This is particularly remarkable when we consider that the parameterization of our estimator was not tuned with respect to classification accuracy (unlike other classification algorithms). It might well be that the performance can be further improved by bandwidth and structure selection strategies that aim for classification rather than estimation accuracy.

4.7 Discussion

In this chapter, we discuss a vine copula approach to nonparametric density estimation. By assuming that the target density belongs to the class of simplified vine densities, we can divide the estimation of a d-dimensional density into several one- and two-dimensional tasks. This allows us to achieve faster convergence rates than classical nonparameteric estimators when $d > 3$. In particular, the speed of convergence is independent of dimension. The advantages of this approach become more and more striking as dimension increases. It shows that a simplified vine model for the dependence between variables is an appealing structure for nonparametric problems. For example, we can expect that similar results can be obtained for copula-based regression models (Noh et al., 2013, Kraus and Czado, 2017).

The crunchpoint in our approach is the simplifying assumption. If the simplifying assumption is not satisfied, the proposed estimator is not consistent — but can nevertheless outperform its competitor in most practicable situations. However, the latter finding may not be true if the simplifying assumption is violated in an extreme fashion and dimension is small. We guess that this is a rather unlikely situation to encounter in real data. However, appropriate tests for a formal empirical assessment have yet to be developed. From a theoretical point of view, this answer is highly unsatisfying and several urging questions arise:

- How dense does the set of simplified densities lie in the set of all densities? Put differently: how far off will we be by assuming a simplified model?
- How can we interpret the components of an estimated simplified model when the assumption does not hold?

Owing to the infancy of vine copula models, these questions remain open to this day. But several recent works have advanced the understanding of the simplifying assumption. A discussion of its appropriateness can be found in Hobæk Haff et al. (2010). Copula classes where the simplifying assumption is satisfied are given in Stöber et al. (2013). Gijbels et al. (2015a) propose a general estimator of the copula for the case where a covariate affects only the marginal distributions (i.e., when the simplifying assumption does hold). Semiparametric estimation of three-dimensional non-simplified PCCs was tackled by Acar et al. (2012) and Vatter and Nagler (2018); a test for the simplifying assumption was proposed by Acar et al. (2013) under a semiparametric model. The empirical pair-copula, an extension of the empirical copula to simplified vine copulas, was analyzed in Hobæk Haff and Segers (2015). The authors conjecture that this estimator converges at the parametric rate — even when pseudo-observations are used. The situation is different from ours since empirical copulas do not suffer the curse of dimensionality.

Spanhel and Kurz (2017) introduced the notion of *partial vine copula approximations (PVCA)*, i.e., the limit of a step-wise estimator under a simplified model. The authors show that the PVCA is not necessarily the best simplified approximation to the true density. They further illustrate in an example that spurious dependence patterns can appear in trees $T_m, m \geq 3$, when the simplifying assumptions has falsely been assumed in previous trees $T_{m'}, 2 \leq m' \leq m$. This property may not matter much in terms of estimation accuracy, but can corrupt the interpretability of an estimated PVCA. The estimator proposed in this paper is in fact an estimator of the PVCA. Our results suggest that the PVCA is a useful inferential object in any case:

- Any d-dimensional PVCA can be consistently estimated at a rate that is equivalent to a two-dimensional problem.
- If the simplifying assumption does hold, the PVCA coincides with the true density.
- If the simplifying assumption does not hold, inference of the PVCA is still less difficult than inference of the actual density. This led to the following observation: On finite samples, a consistent estimate of the PVCA can be much closer to the true density than a consistent estimate of the actual density (see Scenario 2 in Section 4.5).

A related perspective on the phenomenon is that the simplifying assumption allows us to achieve more accurate estimates by model shrinkage. We incorporate the additional 'information' that the simplifying assumption is at least approximately true. This allows us to reduce the set of possible solutions and thereby make the estimation problem 'less difficult'. The most well known example of a shrinkage estimator is the sample variance. When dividing by n instead of $n-1$ we give up unbiasedness of the estimator in order to achieve a smaller error. The same holds true for the vine copula based density estimator: if we make the simplifying assumption although it is not satisfied, we introduce additional bias. In fact, we even give up consistency of the estimator in order to achieve better finite-sample accuracy.

The main advantage of the vine copula based approach is striking: Classical multivariate nonparametric density estimators converge very slowly to the true density when more than a few variables enter the model. Hence, one was unable to benefit from the increasing number of observations in modern data. A vine copula based estimator, on the other hand, converges at a high speed, no matter how many variables are involved. This makes it particularly appealing in the age of big data.

4.8 Proofs

4.8.1 Proof of Theorem 4.1

The proof consists of three steps. In the first step, we show by induction that all pseudo-observations converge sufficiently fast to the true observations. In the second step, we establish pointwise consistency of the feasible pair-copula density estimators $\widehat{c}_{j_e,k_e;D_e}$ and conditional distribution function estimators $\widehat{F}_{j_e|D_e}$ and

$\widehat{F}_{k_e|D_e}$. In the last step, we combine these results to establish the consistency of $\widehat{f}_{\text{vine}}$.

Step 1: Convergence of pseudo-observations

We will show by induction that for all $e \in E_1, \ldots, E_{d-1}$, $i = 1, \ldots, n$,

$$\widehat{U}^{(i)}_{j_e|D_e} - U^{(i)}_{j_e|D_e} = o_{a.s.}(n^{-r}), \quad \widehat{U}^{(i)}_{k_e|D_e} - U^{(i)}_{k_e|D_e} = o_{a.s.}(n^{-r}). \tag{4.13}$$

Let $e \in E_1$ (the conditioning set D_e is empty). Because of 4.1b we have,

$$\big|\widehat{U}^{(i)}_{j_e} - U^{(i)}_{j_e}\big| = \big|\widehat{F}\big(X_{j_e}\big) - F\big(X_{j_e}\big)\big| \leq \sup_{x_{j_e} \in \Omega_{X_{j_e}}} \big|\widehat{F}(x_{j_e}) - F(x_{j_e})\big| = o_{a.s.}(n^{-r}),$$

and the same argument applies to the second equality of (4.13). Now consider $e \in E_m$, $1 \leq m \leq d-2$, and assume that (4.13) holds for all $e \in E_m$. Recall that all pseudo-observations for $e' \in E_{m+1}$ can be written as $\widehat{U}^{(i)}_{j_e|D_e \cup k_e}$ or $\widehat{U}^{(i)}_{k_e|D_e \cup j_e}$ for some $e \in E_m$. Recall the definition of the oracle estimators $\overline{c}$ and $\overline{h}$ (4.5). By the definition of the pseudo-observations and the triangle inequality,

$$\begin{aligned}
\big|\widehat{U}^{(i)}_{j_e|D_e \cup k_e} - U^{(i)}_{j_e|D_e \cup k_e}\big| &= \big|\widehat{h}_{j_e|k_e;D_e}\big\{\widehat{U}^{(i)}_{j_e|D_e}\big|\widehat{U}^{(i)}_{k_e|D_e}\big\} - h_{j_e|k_e;D_e}\big\{U^{(i)}_{j_e|D_e}\big|U^{(i)}_{k_e|D_e}\big)\big\}\big| \\
&\leq \quad \big|\widehat{h}_{j_e|k_e;D_e}\big\{\widehat{U}^{(i)}_{j_e|D_e}\big|\widehat{U}^{(i)}_{k_e|D_e}\big\} - \overline{h}_{j_e|k_e;D_e}\big\{\widehat{U}^{(i)}_{j_e|D_e}\big|\widehat{U}^{(i)}_{k_e|D_e}\big\}\big| \\
&\quad + \big|\overline{h}_{j_e|k_e;D_e}\big\{\widehat{U}^{(i)}_{j_e|D_e}\big|\widehat{U}^{(i)}_{k_e|D_e}\big\} - h_{j_e|k_e;D_e}\big\{\widehat{U}^{(i)}_{j_e|D_e}\big|\widehat{U}^{(i)}_{k_e|D_e}\big\}\big| \\
&\quad + \big|h_{j_e|k_e;D_e}\big\{\widehat{U}^{(i)}_{j_e|D_e}\big|\widehat{U}^{(i)}_{k_e|D_e}\big\} - h_{j_e|k_e;D_e}\big\{U^{(i)}_{j_e|D_e}\big|U^{(i)}_{k_e|D_e}\big\}\big| \\
&= H_{1,n} + H_{2,n} + H_{3,n}
\end{aligned}$$

Note that, almost surely, each realization of $(U^{(i)}_{j_e|D_e}, U^{(i)}_{k_e|D_e})$ is contained in $[\delta_i, 1-\delta_i]^2$ for $\delta_i := \min\big\{U^{(i)}_{j_e|D_e}, U^{(i)}_{k_e|D_e}, 1 - U^{(i)}_{j_e|D_e}, 1 - U^{(i)}_{k_e|D_e}\big\} > 0$. And by invoking (4.13) we see that for sufficiently large n, also each realization of $(\widehat{U}^{(i)}_{j_e|D_e}, \widehat{U}^{(i)}_{k_e|D_e})$ is contained in $[\delta_i/2, 1-\delta_i/2]^2$. Together with 4.2b and 4.3b this yields for large n,

$$\begin{aligned}
H_{1,n} &\leq \sup_{(u,v) \in [\delta_i/2, 1-\delta_i/2]^2} \big|\widehat{h}_{j_e|k_e;D_e}(u|v) - \overline{h}_{j_e|k_e;D_e}(u|v)\big| = O_{a.s.}(a_{e,n}), \\
H_{2,n} &\leq \sup_{(u,v) \in [\delta_i/2, 1-\delta_i/2]^2} \big|\overline{h}_{j_e|k_e;D_e}(u|v) - h_{j_e|k_e;D_e}(u|v)\big| = o_{a.s.}(n^{-r}),
\end{aligned}$$

and invoking (4.13),

$$a_{e,n} = \sup_{i=1,\ldots,n} \bigg\{\big|\widehat{U}^{(i)}_{j_e|D_e} - U^{(i)}_{j_e|D_e}\big| + \big|\widehat{U}^{(i)}_{k_e|D_e} - U^{(i)}_{k_e|D_e}\big|\bigg\} = o_{a.s.}(n^{-r}),$$

which gives $H_{1,n} = o_{a.s.}(n^{-r})$. It remains to show that $H_{3,n} = o_{a.s.}(n^{-r})$. Let $\nabla h_{j_e|k_e;D_e}$ denote the gradient of $h_{j_e|k_e;D_e}$. A first-order Taylor approximation of

$h_{j_e|k_e;D_e}\big(\widehat{U}^{(i)}_{j_e|D_e}|\widehat{U}^{(i)}_{k_e|D_e}\big)$ around $\big(U^{(i)}_{j_e|D_e}, U^{(i)}_{k_e|D_e}\big)$ yields

$$H_{3,n} \leq \left|\nabla^\top h_{j_e|k_e;D_e}\big(U^{(i)}_{j_e|D_e}|U^{(i)}_{k_e|D_e}\big)\begin{pmatrix}\widehat{U}^{(i)}_{j_e|D_e} - U^{(i)}_{j_e|D_e}\\ \widehat{U}^{(i)}_{k_e|D_e} - U^{(i)}_{k_e|D_e}\end{pmatrix}\right| + o_{a.s.}\begin{pmatrix}\widehat{U}^{(i)}_{j_e|D_e} - U^{(i)}_{j_e|D_e}\\ \widehat{U}^{(i)}_{k_e|D_e} - U^{(i)}_{k_e|D_e}\end{pmatrix}.$$

Invoking (4.13), we get $H_{3,n} = o_{a.s.}(n^{-r})$. This establishes the first equality of (4.13) for all $e \in E_{m+1}$. The second equality follows by symmetric arguments and the induction is complete.

Step 2: Consistency of conditional cdf and pair-copula density estimators

In the first step, we have already shown that the estimators of the h-functions converge at rate $o(n^{-r})$. The required conditional distributions $F_{j_e|D_e}$ and $F_{k_e|D_e}$ can be expressed as a chain of h-functions. This implies that for all $e \in E_1, \dots, E_{d-1}$, and all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$,

$$\begin{aligned}\widehat{F}_{j_e|D_e}\big(x_{j_e}|\boldsymbol{x}_{D_e}\big) - F_{j_e|D_e}\big(x_{j_e}|\boldsymbol{x}_{D_e}\big) &= o_p(n^{-r}),\\ \widehat{F}_{k_e|D_e}\big(x_{k_e}|\boldsymbol{x}_{D_e}\big) - F_{k_e|D_e}\big(x_{k_e}|\boldsymbol{x}_{D_e}\big) &= o_p(n^{-r}).\end{aligned} \tag{4.14}$$

Next, we establish that for all $e \in E_1, \dots, E_{d-1}$, and all $(u,v) \in (0,1)^2$,

$$\widehat{c}_{j_e,k_e;D_e}\big(u,v\big) - c_{j_e,k_e;D_e}\big(u,v\big) = O_p(n^{-r}). \tag{4.15}$$

The triangle inequality gives

$$\begin{aligned} & \big|\widehat{c}_{j_e,k_e;D_e}\big(u,v\big) - c_{j_e,k_e;D_e}\big(u,v\big)\big| \\ \leq\ & \big|\widehat{c}_{j_e,k_e;D_e}\big(u,v\big) - \overline{c}_{j_e,k_e;D_e}\big(u,v\big)\big| + \big|\overline{c}_{j_e,k_e;D_e}\big(u,v\big) - c_{j_e,k_e;D_e}\big(u,v\big)\big| \\ =\ & R_{n,1} + R_{n,2}. \end{aligned} \tag{4.16}$$

We have $R_{n,1} = o_{a.s.}(n^{-r})$ by Assumption 4.3a and (4.13), whereas $R_{n,2} = O_p(n^{-r})$ by Assumption 4.2a.

Step 3: Consistency of the vine copula based density estimator

The consistency of $\widehat{f}_{\text{vine}}$ now follows from (4.15) and 4.1a (second equality) together with (4.14) and the fact that $c_{j_e,k_e;D_e}$ is continuously differentiable (third equality):

$$\begin{aligned}
\widehat{f}_{\mathrm{vine}}(\boldsymbol{x}) &= \prod_{k=1}^{d-1}\prod_{e\in E_k} \widehat{c}_{j_e,k_e;D_e}\big\{\widehat{F}_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}),\ \widehat{F}_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\big\} \times \prod_{j=1}^{d}\widehat{f}_j(x_j)\\
&= \prod_{k=1}^{d-1}\prod_{e\in E_k}\Big[c_{j_e,k_e;D_e}\big\{\widehat{F}_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}),\ \widehat{F}_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\big\} + O_p(n^{-r})\Big]\\
&\quad\times\prod_{j=1}^{d}\big\{f_j(x_j)+O_p(n^{-r})\big\}\\
&= \prod_{k=1}^{d-1}\prod_{e\in E_k}\Big[c_{j_e,k_e;D_e}\big\{F_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}),\ F_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\big\} + O_p(n^{-r}) + o_p(n^{-r})\Big]\\
&\quad\times\prod_{j=1}^{d}\big\{f_j(x_j)+O_p(n^{-r})\big\}\\
&= f(\boldsymbol{x}) + O_p(n^{-r}). \qquad \square
\end{aligned}$$

4.8.2 Lemmas

In what follows, we present three lemmas: Lemma 4.1 is for the proof of Proposition 4.3, and Lemma 4.2 and Lemma 4.3 are for the proof of Proposition 4.4.

To ease notation, we write $(u,v) = \big(\Phi(z_1),\Phi(z_2)\big)$,

$$W_1^{(i)} := U_{j_e|D_e}^{(i)},\ W_2^{(i)} := U_{k_e|D_e}^{(i)},\ Z_1^{(i)} := \Phi^{-1}\big(U_{j_e|D_e}^{(i)}\big),\ Z_2^{(i)} := \Phi^{-1}\big(U_{k_e|D_e}^{(i)}\big), \tag{4.17}$$

and $K_{b_n}(\cdot) = b_n^{-1}K(b_n^{-1}\times\cdot)$. In this notation, the (oracle) transformation pair-copula density estimator is

$$\overline{c}(u,v) = \overline{c}\big\{\Phi(z_1),\Phi(z_2)\big\} = \frac{1}{n}\sum_{i=1}^{n}\frac{K_{b_n}\big(z_1 - Z_1^{(i)}\big)K_{b_n}\big(z_2 - Z_2^{(i)}\big)}{\phi(z_1)\phi(z_2)}.$$

The corresponding (oracle) h-function estimator $\overline{h}$ is obtained by integration of $\overline{c}$:

$$\overline{h}(u|v) = \overline{h}\big\{\Phi(z_1)|\Phi(z_2)\big\} = \frac{1}{n}\sum_{i=1}^{n}\frac{J_{b_n}\big(z_1 - Z_1^{(i)}\big)K_{b_n}\big(z_2 - Z_2^{(i)}\big)}{\phi(z_2)}, \tag{4.18}$$

where $J_{b_n}(\cdot) = \int_{-\infty}^{\cdot} K_{b_n}(s)ds$. The feasible estimators $\widehat{c}$ and $\widehat{h}$ are obtained by replacing $W_j^{(i)}$ and $Z_j^{(i)}$ with pseudo-observations $\widehat{W}_j^{(i)}$ and $\widehat{Z}_j^{(i)} := \Phi^{-1}(\widehat{W}_j^{(i)})$. Finally, we define

$$a_n = \sup_{i\in\{1,\dots,n\}}\big|\widehat{W}_1^{(i)} - W_1^{(i)}\big| + \sup_{i\in\{1,\dots,n\}}\big|\widehat{W}_2^{(i)} - W_2^{(i)}\big|.$$

Lemma 4.1. *Under Assumptions 4.5 and 4.7 it holds for all* $(u, v) \in (0,1)^2$,

$$\widehat{c}(u,v) = \overline{c}(u,v) + O_{a.s}(a_n).$$

Proof. By a first-order Taylor approximation of Φ^{-1}, $j = 1, 2$,

$$\begin{aligned}\widehat{Z}_j^{(i)} - Z_j^{(i)} &= (\widehat{W}_j^{(i)} - W_j^{(i)})/\phi(Z_j^{(i)}) + o_{a.s.}(\widehat{W}_j^{(i)} - W_j^{(i)}) \\ &= 1/\phi(Z_j^{(i)}) \times O_{a.s.}(a_n),\end{aligned} \tag{4.19}$$

where the $O_{a.s.}(a_n)$ term does not depend on the index i since the supremum was taken. Denote $\nabla_{\boldsymbol{z}} = (\partial/\partial z_1, \partial/\partial z_2)^\top$. A first-order Taylor approximation of K yields

$$\begin{aligned}&\phi(z_1)\phi(z_2)\big|\widehat{c}\big\{\Phi(z_1),\Phi(z_2)\big\} - \overline{c}\big\{\Phi(z_1),\Phi(z_2)\big\}\big| \\ &= \left|\frac{1}{n}\sum_{i=1}^n K_{b_n}\big(z_1 - \widehat{Z}_1^{(i)}\big)K_{b_n}\big(z_2 - \widehat{Z}_2^{(i)}\big) - \frac{1}{n}\sum_{i=1}^n K_{b_n}\big(z_1 - Z_1^{(i)}\big)K_{b_n}\big(z_2 - Z_2^{(i)}\big)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{z}}\big\{K_{b_n}(z_1 - Z_1^{(i)})K_{b_n}(z_2 - Z_2^{(i)})\big\}\begin{pmatrix}\widehat{Z}_1^{(i)} - Z_1^{(i)} \\ \widehat{Z}_2^{(i)} - Z_2^{(i)}\end{pmatrix} + o_{a.s.}\left\{\begin{pmatrix}\widehat{Z}_1^{(i)} - Z_1^{(i)} \\ \widehat{Z}_2^{(i)} - Z_2^{(i)}\end{pmatrix}\right\}\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{z}}\big\{K_{b_n}(z_1 - Z_1^{(i)})K_{b_n}(z_2 - Z_2^{(i)})\big\}\begin{pmatrix}1/\phi(Z_1^{(i)}) \\ 1/\phi(Z_2^{(i)})\end{pmatrix}\right| \times O_{a.s.}(a_n),\end{aligned}$$

where the last inequality is due to (4.19). Since K_{b_n} is zero outside of $[-b_n, b_n]$, we can bound this further by

$$\eta_n(\boldsymbol{z}) \times \left|\nabla_{\boldsymbol{z}}\left\{\frac{1}{n}\sum_{i=1}^n K_{b_n}(z_1 - Z_1^{(i)})K_{b_n}(z_2 - Z_2^{(i)})\right\}\right| \times O_{a.s.}(a_n), \tag{4.20}$$

where $\eta_n(\boldsymbol{z}) := \sup_{y \in [\min\{z_1,z_2\} - b_n, \max\{z_1,z_2\} + b_n]} 1/\phi(y) = O(1)$ for all $\boldsymbol{z} \in \mathbb{R}^2$. The second term is the absolute value of the gradient of a classical kernel density estimator. Since the derivatives of ψ are continuous and bounded by C2, it holds for $\psi(z_1, z_2) = c\{\Phi(z_1), \Phi(z_2)\}\phi(z_1)\phi(z_2)$,

$$\left|\nabla_{\boldsymbol{z}}\left\{\frac{1}{n}\sum_{i=1}^n K_{b_n}(z_1 - Z_1^{(i)})K_{b_n}(z_2 - Z_2^{(i)})\right\}\right| = \big|\nabla_{\boldsymbol{z}}\psi(z_1, z_2)\big| + o_{a.s.}(1),$$

see Theorem 9 of Hansen (2008). Plugging this into (4.20) proves our claim. □

Lemma 4.2. *Under Assumptions 4.5 and 4.7 it holds for all* $(u, v) \in (0,1)^2$, $\delta \in (0, 0.5]$,

$$\sup_{(u,v)\in[\delta,1-\delta]^2} \big|\overline{h}(u|v) - h(u|v)\big| = O_{a.s.}\big(b_n^2 + \sqrt{\ln n/(nb_n)}\big).$$

Proof. Equations 40 and 41 of Hansen (2004) yield

$$\mathrm{E}\big\{\overline{h}(u|v)\big\} - h\big(u|v\big) = b_n^2\beta(u,v) + o(b_n^2),$$

for some bias term $\beta(u,v)$ involving h and ϕ as well as their first- and second order derivatives. Since all parts are continuous on $[\delta, 1-\delta]^2$ by C1 for all $\delta \in (0, 0.5]$, it holds

$$\sup_{(u,v)\in[\delta,1-\delta]^2} \big|\mathrm{E}\big\{\overline{h}(u|v)\big\} - h\big(u|v\big)\big| = O_{a.s.}\big(b_n^2\big).$$

On the other hand, Lemma 2.2 of Härdle et al. (1988) ensures that

$$\sup_{(u,v)\in[\delta,1-\delta]^2} \big|\overline{h}(u|v) - \mathrm{E}\big\{\overline{h}(u|v)\big\}\big| = O_{a.s.}\big(\sqrt{\ln n/(nb_n)}\big).$$

Combining the previous two equations concludes the proof. □

Lemma 4.3. *Under Assumptions 4.5 and 4.7 it holds for all* $(u,v) \in (0,1)^2$, $\delta \in (0, 0.5]$,

$$\sup_{(u,v)\in[\delta,1-\delta]^2} \big|\widehat{h}(u|v) - \overline{h}(u|v)\big| = O_{a.s.}\big(a_n\big).$$

Proof. With arguments similar to the proof of Lemma 4.1, we can show

$$\begin{aligned} &\sup_{(u,v)\in[\delta,1-\delta]^2} \big|\widehat{h}(u|v) - \overline{h}(u|v)\big| \\ = &\sup_{\boldsymbol{z}\in[\Phi^{-1}(\delta),\Phi^{-1}(1-\delta)]^2} \big|\widehat{h}\big\{\Phi(z_1)|\Phi(z_2)\big\} - \overline{h}\big\{\Phi(z_1)|\Phi(z_2)\big\}\big| \\ \leq &\sup_{\boldsymbol{z}\in[\Phi^{-1}(\delta),\Phi^{-1}(1-\delta)]^2} \left|\frac{\eta_n(\boldsymbol{z})}{\phi(z_2)} \times \nabla_{\boldsymbol{z}} h\big\{\Phi(z_1)|\Phi(z_2)\big\}\right| \times O_{a.s.}(a_n), \end{aligned}$$

where $\eta_n(\boldsymbol{z}) = \sup_{y\in[\min\{z_1,z_2\}-b_n,\max\{z_1,z_2\}+b_n]} 1/\phi(y)$ and the $O_{a.s}$ term is independent of $\boldsymbol{z}$. The supremum on the right hand side is $O(1)$ because all functions are continuous in $\boldsymbol{z}$ on every compact subset of $\mathbb{R}^2$. As a result, the right can be bounded by a constant times the $O_{a.s.}(a_n)$ term. This establishes our claim. □

4.8.3 Proof of Proposition 4.5

Step 1: Reduce problem to oracle estimators

From Proposition 4.2 we get for all $\ell = 1, \dots, d$, and $x \in \mathbb{R}$, that

$$\widehat{f}_\ell(x) = f_\ell(x) + O_p\{b_{n,m}^2 + (nb_{n,m}/\ln n)^{-1/2}\} = f_\ell(x) + o_p\{b_{n,c}^2 + (nb_{n,c}^2)^{-1/2}\},$$

where the second equality follows from condition (v) in Proposition 4.5. This implies $(nb_{n,c}^2)^{1/2}\{\widehat{f}_\ell(x) - f_\ell(x)\} = o_p(1)$ and we have established that the first

d components of (4.11) converge to zero in probability. Hence, the first d components of $\boldsymbol{\mu}_{\boldsymbol{x}}$ as well as the first d rows and columns of $\Sigma_{\boldsymbol{x}}$ will be zero and we only have to deal with the remaining components in (4.11).

Recall that the rate n^{-r} in Section 4.8.1 corresponds to $b_{n,c}^2 + (nb_{n,c}^2)^{-1/2}$ in the implementation used for Proposition 4.5. From (4.14) and the bound for $R_{n,1}$ in (4.16) we thus know that

$$\widehat{F}_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}) = F_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}) + o_p\{b_{n,c}^2 + (nb_{n,c}^2)^{-1/2}\},$$

and

$$\widehat{c}_{j_e,k_e;D_e}(u,v) = \overline{c}_{j_e,k_e;D_e}(u,v) + o_p\{b_{n,c}^2 + (nb_{n,c}^2)^{-1/2}\}.$$

Similar to Lemma 4.3, we can now show that

$$\begin{aligned}&\overline{c}_{j_e,k_e;D_e}\big\{\widehat{F}_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}), \widehat{F}_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\big\}\\ &= \overline{c}_{j_e,k_e;D_e}\big\{F_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}), F_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\big\} + o_p\{b_{n,c}^2 + (nb_{n,c}^2)^{-1/2}\}.\end{aligned}$$

Hence, for (4.11) to hold it suffices to show that

$$(nb_{n,c}^2)^{1/2}\{\overline{\boldsymbol{c}}^*(\boldsymbol{x}) - b_{n,c}^2\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} - \boldsymbol{c}^*(\boldsymbol{x})\} \xrightarrow{d} \mathcal{N}\big(0, \tilde{\Sigma}_{\boldsymbol{x}}\big), \tag{4.21}$$

where

$$\overline{\boldsymbol{c}}^*(\boldsymbol{x}) = \big(\overline{c}_{j_e,k_e;D_e}\{F_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e}), F_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\}\big)_{e\in E_1,\dots,E_{d-1}},$$

and $\boldsymbol{c}^*(\boldsymbol{x})$ is defined similarly, but replacing $\overline{c}_{j_e,k_e;D_e}$ with $c_{j_e,k_e;D_e}$.

Step 2: Reformulate problem according to Lindeberg-Feller CLT

Define $Z_{j_e|D_e}^{(i)} := \Phi^{-1}(U_{j_e|D_e}^{(i)})$, $Z_{k_e|D_e}^{(i)} := \Phi^{-1}(U_{k_e|D_e}^{(i)})$, $z_{j_e|D_e} := \Phi^{-1}\big\{F_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e})\big\}$, $z_{k_e|D_e} := \Phi^{-1}\big\{F_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})\big\}$. Let $\boldsymbol{Y}_{n,i} := (Y_{n,i,e})_{e\in E_1,\dots,E_{d-1}}$, be a vector with entries

$$Y_{n,i,e} := (nb_{n,c}^2)^{-1/2}\frac{K\left(\frac{Z_{j_e|D_e}^{(i)} - z_{j_e|D_e}}{b_n}\right)K\left(\frac{Z_{k_e|D_e}^{(i)} - z_{k_e|D_e}}{b_n}\right)}{\phi(z_{j_e|D_e})\phi(z_{k_e|D_e})}.$$

Then, $\sum_{i=1}^n \boldsymbol{Y}_{n,i} = (nb_{n,c}^2)^{1/2}\overline{\boldsymbol{c}}^*(\boldsymbol{x})$. By the multivariate Lindeberg-Feller central limit theorem (van der Vaart, 1998, Proposition 2.27), (4.21) holds when

$$\sum_{i=1}^n \mathrm{E}\big(\boldsymbol{Y}_{n,i}\big) = (nb_{n,c}^2)^{1/2}\big\{\boldsymbol{c}^*(\boldsymbol{x}) + b_{n,c}^2\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}} + o(b_{n,c}^2)\big\}, \tag{4.22}$$

$$\sum_{i=1}^n \mathrm{Cov}(\boldsymbol{Y}_{n,i}) \to \tilde{\Sigma}_{\boldsymbol{x}}, \tag{4.23}$$

$$\sum_{i=1}^n \mathrm{E}\big\{\|\boldsymbol{Y}_{n,i}\|^2 \mathbb{1}\big(\|\boldsymbol{Y}_{n,i}\| > \varepsilon\big)\big\} \to 0, \quad \text{for all } \varepsilon > 0. \tag{4.24}$$

Since $\boldsymbol{Y}_{n,i}$ are independent for $i = 1, \dots, n$, it holds

$$\sum_{i=1}^n \mathrm{E}\big(\boldsymbol{Y}_{n,i}\big) = n\mathrm{E}\big(\boldsymbol{Y}_{n,i}\big), \qquad \sum_{i=1}^n \mathrm{Cov}(\boldsymbol{Y}_{n,i}) = n\mathrm{Cov}(\boldsymbol{Y}_{n,i}).$$

Step 3: Check conditions of the Lindeberg-Feller CLT

Step 3.1: Check Equation 4.22

Denote further $u_{j_e|D_e} := F_{j_e|D_e}(x_{j_e}|\boldsymbol{x}_{D_e})$, $u_{k_e|D_e} := F_{k_e|D_e}(x_{k_e}|\boldsymbol{x}_{D_e})$. Corollary 3.4 of Nagler (2014) states

$$n\mathrm{E}\big(Y_{n,i,e}\big) = (nb_{n,c}^2)^{1/2}\big\{c_{j_e,k_e;D_e}(u_{j_e|D_e}, u_{k_e|D_e}) + b_{n,c}^2\tilde{\mu}_{\boldsymbol{x},e} + o(b_{n,c}^2)\big\}, \tag{4.25}$$

where

$$\begin{aligned}\tilde{\mu}_{\boldsymbol{x},e} := \Bigg\{ & \frac{\partial^2 c_{j_e,k_e;D_e}\big(u_{j_e|D_e}, u_{k_e|D_e}\big)}{\partial u_{j_e|D_e}^2}\phi^2\big(z_{j_e|D_e}\big) + \frac{\partial^2 c_{j_e,k_e;D_e}\big(u_{j_e|D_e}, u_{k_e|D_e}\big)}{\partial u_{k_e|D_e}^2}\phi^2\big(z_{k_e|D_e}\big) \\ & - \frac{3\partial c_{j_e,k_e;D_e}\big(u_{j_e|D_e}, u_{k_e|D_e}\big)}{\partial u_{j_e|D_e}}\phi\big(z_{j_e|D_e}\big)z_{j_e|D_e} \\ & - \frac{3\partial c_{j_e,k_e;D_e}\big(u_{j_e|D_e}, u_{k_e|D_e}\big)}{\partial u_{k_e|D_e}}\phi\big(z_{k_e|D_e}\big)z_{k_e|D_e} \\ & + c_{j_e,k_e;D_e}\big(u_{j_e|D_e}, u_{k_e|D_e}\big) \times \big(z_{j_e|D_e}^2 + z_{k_e|D_e}^2 - 2\big)\Bigg\}\frac{\sigma_K^2}{2},\end{aligned} \tag{4.26}$$

and $\sigma_K^2 := \int_{[-1,1]} x^2 K(x)dx$. This validates (4.22).

Step 3.2: Check Equation 4.23

We first consider the diagonal elements of $\mathrm{Cov}(Y_{n,i})$, i.e., $\mathrm{Var}(Y_{n,i,e})$, $e \in E_m, m = 1, \dots, d-1$. By the change of variable $s_1 = (z_1 - z_{j_e|D_e})/b_{n,c}$, $s_2 = (z_2 - z_{k_e|D_e})/b_{n,c}$, and a Taylor approximation of $\psi_{j_e,k_e;D_e}$ (as defined in Assumption 4.7), we get

$$
\begin{aligned}
&nE\big(Y_{n,i,e}^2\big)\phi^2(z_{j_e|D_e})\phi^2(z_{k_e|D_e})\\
&= nE\bigg\{\frac{1}{nb_{n,c}^2}K^2\bigg(\frac{Z_{j_e|D_e}^{(i)}-z_{j_e|D_e}}{b_{n,c}}\bigg)K^2\bigg(\frac{Z_{k_e|D_e}^{(i)}-z_{k_e|D_e}}{b_{n,c}}\bigg)\bigg\}\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} K^2(s_1)K^2(s_2)\psi_{j_e,k_e;D_e}(z_{j_e|D_e}-b_{n,c}s_1, z_{k_e|D_e}-b_{n,c}s_2)ds_1ds_2\\
&= \nu_K^2\psi_{j_e,k_e;D_e}(z_{j_e|D_e}, z_{k_e|D_e}) + o(1), \qquad (4.27)
\end{aligned}
$$

where $\nu_K := \int_{\mathbb{R}} K^2(s)ds$. Using (4.27) and (4.25), we obtain

$$
n\mathrm{Var}(Y_{n,i,e}) \to \nu_K^2\frac{c_{j_e,k_e;D_e}\big(u_{j_e|D_e}, u_{k_e|D_e}\big)}{\phi(z_{j_e|D_e})\phi(z_{k_e|D_e})} =: \tilde{\sigma}_{\boldsymbol{x},e}. \qquad (4.28)
$$

Now consider the off-diagonal elements of $\mathrm{Cov}(\boldsymbol{Y}_{n,i})$, i.e., $\mathrm{Cov}(Y_{n,i,e}, Y_{n,i,e'})$ with $e \neq e'$. We get

$$
\begin{aligned}
&nE(Y_{n,i,e}Y_{n,i,e'})\phi(z_{j_e|D_e})\phi(z_{k_e|D_e})\phi(z_{j_{e'}|D_{e'}})\phi(z_{k_{e'}|D_{e'}})\\
&= \frac{1}{b_{n,c}^2}E\bigg\{K\bigg(\frac{Z_{j_e|D_e}^{(i)}-z_{j_e|D_e}}{b_{n,c}}\bigg)K\bigg(\frac{Z_{k_e|D_e}^{(i)}-z_{k_e|D_e}}{b_{n,c}}\bigg)\\
&\qquad\times K\bigg(\frac{Z_{j_{e'}|D_{e'}}^{(i)}-z_{j_{e'}|D_{e'}}}{b_{n,c}}\bigg)K\bigg(\frac{Z_{k_{e'}|D_{e'}}^{(i)}-z_{k_{e'}|D_{e'}}}{b_{n,c}}\bigg)\bigg\}. \qquad (4.29)
\end{aligned}
$$

Since $e \neq e'$, the elements of the set $\{(j_e, D_e), (k_e, D_e), (j_{e'}, D_{e'}), (k_{e'}, D_{e\prime})\}$ are either all distinct or there is a match between exactly two of them.

We first consider the case where all are distinct. Define $\psi_{e,e'}$ the joint density of $(Z_{j_e|D_e}^{(i)}, Z_{k_e|D_e}^{(i)}, Z_{j_{e'}|D_{e'}}^{(i)}, Z_{k_{e'}|D_{e'}}^{(i)})$ and $\boldsymbol{z} = (z_{j_e|D_e}, z_{k_e|D_e}, z_{j_{e'}|D_{e'}}, z_{k_{e'}|D_{e'}})$. Using a change of variables and Taylor expansion similar to (4.27), we get

$$
\begin{aligned}
(4.29) &= b_{n,c}^2\int_{\mathbb{R}^4}\{K(s_1)K(s_2)K(s_3)K(s_4)\psi_{e,e'}(\boldsymbol{z}-b_{n,c}\boldsymbol{s})d\boldsymbol{s}\\
&= b_{n,c}^2\bigg\{\int_{\mathbb{R}^4}\{K(s_1)K(s_2)K(s_3)K(s_4)d\boldsymbol{s}\psi_{e,e'}(\boldsymbol{z}) + o(b_{n,c})\bigg\}\\
&= O(b_{n,c}^2) = o(1).
\end{aligned}
$$

Now consider the case where there is a match between two pseudo observations. Without loss of generality, we assume that $(j_e, D_e) = (j_{e'}, D_{e'})$ and define $\psi_{e,e'}$ the joint density of $(Z_{j_e|D_e}^{(i)}, Z_{k_e|D_e}^{(i)}, Z_{k_{e'}|D_{e'}}^{(i)})$ and $\boldsymbol{z} = (z_{j_e|D_e}, z_{k_e|D_e}, z_{k_{e'}|D_{e'}})$. This yields

$$\begin{aligned}
(4.29) &= b_{n,c} \int_{\mathbb{R}^3} \{K^2(s_1)K(s_2)\psi_{e,e'}(\boldsymbol{z} - b_{n,c}\boldsymbol{s})d\boldsymbol{s} \\
&= b_{n,c}\Big\{ \int_{\mathbb{R}^3} \{K^2(s_1)K(s_2)K(s_3))d\boldsymbol{s}\psi_{e,e'}(\boldsymbol{z}) + o(b_{n,c}) \Big\} \\
&= O(b_{n,c}) = o(1).
\end{aligned}$$

We have shown that (4.23) holds with $\tilde{\Sigma}_{\boldsymbol{x}}$ being a diagonal matrix with diagonal entries $\tilde{\sigma}_{\boldsymbol{x},e}$ given in (4.28).

Step 3.3: Check Equation 4.24

Instead of checking the remaining condition (4.24) directly, we will verify the stronger Lyapunov-type condition $\sum_{i=1}^n \mathrm{E}(\|\boldsymbol{Y}_{n,i}\|^3) \to 0$. By Jensen's inequality we get

$$n\mathrm{E}\big(\|\boldsymbol{Y}_{n,1}\|^3\big) = nE\bigg\{\bigg(\sum_{m=1}^{d-1}\sum_{e\in E_m} Y_{n,1,e}^2\bigg)^{3/2}\bigg\} \le n\sqrt{d(d-1)/2}\sum_{m=1}^{d-1}\sum_{e\in E_m} \mathrm{E}\big(Y_{n,1,k}^3\big),$$

where $d(d-1)/2$ is the number of terms in the double sum. Hence, it suffices to show $n\mathrm{E}(Y_{n,1,e}^3) \to 0$ for any $e \in E_1, \dots, E_{d-1}$. Similar to (4.27), we get

$$\begin{aligned}
& n\mathrm{E}(Y_{n,1,e}^3)\phi^3(z_{j_e|D_e})\phi^3(z_{k_e|D_e}) \\
&= n\mathrm{E}\bigg\{\frac{1}{(nb_{n,c}^2)^{3/2}} K^3\bigg(\frac{Z_{j_e|D_e}^{(i)} - z_{j_e|D_e}}{b_{n,c}}\bigg) K^3\bigg(\frac{Z_{k_e|D_e}^{(i)} - z_{k_e|D_e}}{b_{n,c}}\bigg)\bigg\} \\
&= \frac{1}{n^{1/2}b_{n,c}^3}\mathrm{E}\bigg\{K^3\bigg(\frac{Z_{j_e|D_e}^{(i)} - z_{j_e|D_e}}{b_{n,c}}\bigg) K^3\bigg(\frac{Z_{k_e|D_e}^{(i)} - z_{k_e|D_e}}{b_{n,c}}\bigg)\bigg\} \\
&= \frac{1}{n^{1/2}b_{n,c}} \int_{\mathbb{R}}\int_{\mathbb{R}} K^3(s_1)K^3(s_2)\psi_{j_e,k_e;D_e}(z_{j_e|D_e} - b_{n,c}s_1, z_{k_e|D_e} - b_{n,c}s_2)ds_1 ds_2 \\
&= O\big\{(nb_{n,c}^2)^{-1/2}\big\},
\end{aligned}$$

which is $o(1)$ by Assumption 4.5 . □

5 Solving estimating equations with copulas

5.1 Introduction

While copulas are often used as convenience tools to glue together arbitrary marginal distributions, they have also been applied to solve regression problems (Song, 2000, Oakes and Ritz, 2000, Pitt et al., 2006, Kolev and Paiva, 2009, Song et al., 2009, Yin and Yuan, 2009, Noh et al., 2013, Cooke et al., 2015). Another recent development is in quantile regression (Koenker, 2005), with contributions from the econometrics and statistics literatures, respectively in the context of univariate quantile auto-regression (Bouyé and Salmon, 2009, Chen et al., 2009) and conditional quantile estimation (Noh et al., 2015, Kraus and Czado, 2017, Rémillard et al., 2017). A problem with those approaches is that they use parametric copula families, which only allow for monotonic regression functions (Dette et al., 2014). In De Backer et al. (2017), the authors alleviate this issue by considering a semiparametric estimator. Schallhorn et al. (2017) suggests a fully nonparametric estimator, but lacks theoretical results.

Using the fact that both the mean and quantiles can be regarded as roots of *estimating functions* (Godambe, 1991), we propose a unified framework to solve a wide range of statistical learning problems using copulas. The proposed framework is very general and covers essentially all regression problems, e.g., mean, quantile, expectile, exponential family, and instrumental variable regression.

Assume that one is interested in finding the zeros of estimating functions of a set of response variables conditionally on covariates. The main idea explored in this chapter is that conditional expectations can be replaced by weighted unconditional ones, with the weight being a ratio of copula densities. In other words, copula-based Z-estimators are built by estimating the weight function first, and then solving for the zeros of an estimating function using this weight (Z-estimators are estimators that solve for zeros of an estimating equation, see, van der Vaart, 1998, Section 5.1, for a general introduction).

We justify this approach by a rigorous asymptotic theory based on empirical process, along with verifiable assumptions. In particular, we prove consistency, weak convergence, and validity of the bootstrap of the corresponding Z-estimators. This complements and generalizes several known results for mean and quantile regression. Note that our theory also applies to consistent M-estimators (defined as maximizers of a criterion function) when their estimating functions are differentiable (see, Kosorok, 2007, Section 2.2.6).

The remainder of this chapter is organized as follows. In Section 5.2, we formalize the problem and give motivating examples. The main results are presented in Section 5.3. We also discuss our assumptions, both in general and in the context of the motivating examples. We showcase our method in various simulations setups in Section 5.4. Finally, we discuss the results in a larger context and propose possible extensions in Section 5.5.

5.2 Copula-based solutions to estimating equations

5.2.1 Representing estimating equations in terms of copulas

For arbitrary random vectors $\boldsymbol{Y}$ and $\boldsymbol{X}$, we denote by $F_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x}) = \mathrm{P}(\boldsymbol{Y} \leq \boldsymbol{y}, \boldsymbol{X} \leq \boldsymbol{x})$ their joint distribution, $F_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x}) = \mathrm{P}(\boldsymbol{Y} \leq \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})$ the distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$, $F_{Y_j}(y_j) = \mathrm{P}(Y_j \leq y_j)$ and $F_{X_j}(x_j) = \mathrm{P}(X_j \leq x_j)$ the marginal distributions. Assuming that all random variables are absolutely continuous, we write the corresponding densities as $f_{\boldsymbol{Y},\boldsymbol{X}}$, $f_{\boldsymbol{Y}|\boldsymbol{X}}$, f_{Y_j}, and f_{X_j}, respectively.

Let $\boldsymbol{Y} \in \mathcal{Y} \subseteq \mathbb{R}^q$ the subject of interest (called *response*), and $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ a vector of auxiliary variables (called *predictors* or *covariates*). The response is often a univariate random variable in the context of regression or classification, but can be enriched to encompass an exogenous treatment effect or instrumental variables (see Example 5.5 below). Fix $\boldsymbol{x} \in \mathcal{X}$ and let $\theta = \theta(\boldsymbol{x})$ be a parameter of interest. It can be either a scalar, vector, or more generally an object in a normed space Θ. Suppose there is a family of functions $\psi_\theta \colon \mathcal{Y} \to \mathbb{R}$ indexed by θ with the property

$$\mathrm{E}\big\{\psi_{\theta_0}(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\big\} = 0, \tag{5.1}$$

where θ_0 denotes the true parameter. The function ψ_θ is called *identifying function* and (5.1) the (population version of an) estimating equation. The name estimating equation is motivated by the fact that an estimator for θ_0 can be constructed from a sample version of (5.1).

We shall see that the conditional expectation in (5.1) can be replaced by an unconditional one of the form

$$\mathrm{E}\big\{\psi_{\theta_0}(\boldsymbol{Y}) w_{\boldsymbol{x}}(\boldsymbol{Y})\big\} = 0, \tag{5.2}$$

where $w_{\boldsymbol{x}}$ is a weight function that accounts for the conditioning on $\boldsymbol{X} = \boldsymbol{x}$. This unconditional representation is explained by Sklar's theorem for densities (see Section 2.1). It states that the joint density $f_{\boldsymbol{Z}}$ for any absolutely continuous random vector $\boldsymbol{Z} \in \mathbb{R}^d$ can be represented as

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = c_Z\{F_{Z_1}(z_1), \dots, F_{Z_d}(z_d)\} \prod_{j=1}^{d} f_{Z_j}(z_j), \tag{5.3}$$

where $c_{\boldsymbol{Z}}$ is called *copula density*. More precisely, $c_{\boldsymbol{Z}}$ is the density of the vector of probability integral transforms $\big(F_{Z_1}(Z_1), \dots, F_{Z_d}(Z_d)\big)$, where all components are standard uniform variables. In particular, $c_Z \equiv 1$ for univariate Z.

Using (5.3), we can write the conditional density $f_{\boldsymbol{Y}|\boldsymbol{X}}$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$ as

$$\frac{f_{\boldsymbol{Y},\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{c_{\boldsymbol{Y},\boldsymbol{X}}\{F_{Y_1}(y_1), \dots, F_{Y_q}(y_q), F_{X_1}(x_1), \dots, F_{X_p}(x_p)\}}{c_{\boldsymbol{X}}\{F_{X_1}(x_1), \dots, F_{X_p}(x_p)\}} \left\{\prod_{j=1}^{q} f_{Y_j}(y_j)\right\}.$$

This yields,

$$\begin{aligned} \mathrm{E}\{\psi_{\theta_0}(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\} &= \int_{\mathcal{Y}} \psi_{\theta_0}(\boldsymbol{y}) f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y} \mid \boldsymbol{x}) d\boldsymbol{y} \\ &= \frac{\int_{\mathcal{Y}} \psi_{\theta_0}(\boldsymbol{y}) c_{\boldsymbol{Y},\boldsymbol{X}}\{F_{Y_1}(y_1), \dots, F_{X_p}(x_p)\}\{\prod_{j=1}^{q} f_{Y_j}(y_j)\} d\boldsymbol{y}}{c_{\boldsymbol{X}}\{F_{X_1}(x_1), \dots, F_{X_p}(x_p)\}}. \end{aligned}$$

Since the the denominator does not depend on θ_0, it has no effect when solving for a zero of $\mathrm{E}\{\psi_{\theta_0}(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\}$ (which is our ultimate goal). For the numerator, we apply (5.3) to $f_{\boldsymbol{Y}}$, which yields

$$\begin{aligned} &\int_{\mathcal{Y}} \psi_{\theta_0}(\boldsymbol{y}) c_{\boldsymbol{Y},\boldsymbol{X}}\{F_{Y_1}(y_1), \dots, F_{X_p}(x_p)\}\{\prod_{j=1}^{q} f_{Y_j}(y_j)\} d\boldsymbol{y} \\ &= \int_{\mathcal{Y}} \psi_{\theta_0}(\boldsymbol{y}) \frac{c_{\boldsymbol{Y},\boldsymbol{X}}\{F_{Y_1}(y_1), \dots, F_{X_p}(x_p)\}}{c_{\boldsymbol{Y}}\{F_{Y_1}(y_1), \dots, F_{Y_q}(y_q)\}} f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y} \\ &= \mathrm{E}\{\psi_{\theta_0}(\boldsymbol{Y}) w_{\boldsymbol{x}}(\boldsymbol{Y})\}, \end{aligned}$$

and, hence, (5.2) holds with

$$w_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{c_{\boldsymbol{Y},\boldsymbol{X}}\{F_{Y_1}(y_1), \dots, F_{Y_q}(y_q), F_{X_1}(x_1), \dots, F_{X_p}(x_p)\}}{c_{\boldsymbol{Y}}\{F_{Y_1}(y_1), \dots, F_{Y_q}(y_q)\}}. \tag{5.4}$$

Remark 5.1. *Since $c_{\boldsymbol{Y}}$ is a copula density, it has uniform marginal densities. In particular, if the response is univariate ($q = 1$), it holds $c_Y \equiv 1$ and*

$$w_{\boldsymbol{x}}(y) = c_{Y,\boldsymbol{X}}\{F_Y(y), F_{X_1}(x_1), \dots, F_{X_p}(x_p)\}.$$

5.2.2 Estimators for copula-based estimating equations

Suppose we observe an *iid* sequence of random vectors $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$. We shall use a sample version of the unconditional estimating equation (5.2) to construct copula-based estimators for the parameter θ_0. To do this, all unknown quantities in (5.2) are replaced by estimates.

The class of estimators we consider arises from two approximations. The sample average $n^{-1}\sum_{i=1}^n \psi_{\theta_0}(\boldsymbol{Y}_i)w_{\boldsymbol{x}}(\boldsymbol{Y}_i)$ provides a natural estimate of the expectation in (5.2). Let further $\widehat{w}_{\boldsymbol{x}}(y) = \widehat{w}_{\boldsymbol{x}}(y; \boldsymbol{Y}_1, \boldsymbol{X}_1, \dots, \boldsymbol{Y}_n, \boldsymbol{X}_n)$ be an estimator of $w_{\boldsymbol{x}}$ (examples of such estimators will be discussed in Section 5.3.3). Then we define an estimator $\widehat{\theta} = \widehat{\theta}(\boldsymbol{Y}_1, \boldsymbol{X}_1, \dots, \boldsymbol{Y}_n, \boldsymbol{X}_n)$ of θ_0 as the solution to

$$\frac{1}{n}\sum_{i=1}^n \psi_{\widehat{\theta}}(\boldsymbol{Y}_i)\widehat{w}_{\boldsymbol{x}}(\boldsymbol{Y}_i) = 0. \tag{5.5}$$

Since $w_{\boldsymbol{x}}$ is a ratio of copula densities, the estimator $\widehat{\theta}$ is primarily driven by estimators for the copula density. This allows us to harness the rich toolbox of existing copula models and associated estimating techniques. In particular, we can model $c_{\boldsymbol{Y},\boldsymbol{X}}$ by a vine copula and apply the nonparametric estimators investigated in previous chapters. The proposed framework is much more general, however, and also includes other models and estimators.

5.2.3 Examples

The estimator solving (5.5) is quite versatile. With different choices of the identifying function ψ_{θ_0}, we can solve for a variety of conditional functionals. In the following, we introduce a few popular examples that will be discussed further in later sections.

Example 5.1 (Mean regression). *The classical example with a univariate response is the conditional mean $\theta_0 = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ which is identified by $\psi_\theta(y) = y - \theta$. The estimating equation (5.5) then has the explicit solution*

$$\widehat{\theta} = \frac{\sum_{i=1}^n Y_i\, \widehat{w}_{\boldsymbol{x}}(Y_i)}{\sum_{i=1}^n \widehat{w}_{\boldsymbol{x}}(Y_i)},$$

which is quite similar to the Nadaraya-Watson estimator for the conditional mean. A difference is that the weights $\widehat{w}_{\boldsymbol{x}}$ are also functions of the response Y.

Example 5.2 (Quantile regression). *The conditional quantile $\theta_{0,t} = F_{Y|\boldsymbol{X}}^{-1}(t \mid \boldsymbol{X} = \boldsymbol{x})$ for $t \in T = (0,1)$ can be found by minimizing the conditional expectation of the "check-function" $\rho_{\theta_t}(y) = (y - \theta_t)\{t - \mathbb{1}(y < \theta_t)\}$ or, equivalently, solving (5.2) for $\theta_{0,t}$ with*

$$\psi_{\theta_t,t}(y) = t\mathbb{1}(y \geq \theta_t) - (1-t)\mathbb{1}(y < \theta_t). \tag{5.6}$$

In this context, $\psi_{\theta,t} \in \Psi$, where $\Psi = \{\psi_{\theta,t}, t \in T\}$ is a class of identifying functions indexed by the quantile level t and induces a process of solutions $\{\theta_{0,t} \in T\}$. The corresponding sample version (5.5) can only be solved numerically, although there are efficient algorithms to compute "weighted quantiles".

Example 5.3 (Expectile regression). *Expectiles generalize the mean of a distribution in a way that is similar to how quantiles generalize the median (see, e.g., Newey and Powell, 1987, for details). The identifying function is*

$$\psi_{\theta,t}(y) = t(y-\theta)\mathbb{1}(y \geq \theta) - (1-t)(\theta - y)\mathbb{1}(y < \theta),$$

for $t \in T = (0,1)$, and mean regression is recovered by $t = 1/2$.

Example 5.4 (Exponential family regression). *Suppose $f_{Y|X=x}$ is a one parameter exponential family with canonical parameter θ, i.e.,*

$$f(y;\theta) = h(y)\exp\left\{a(y)\theta - b(\theta)\right\},$$

where h, a, and b are known functions. Then the parameter can be identified via

$$\psi_\theta(y) = \big\{a(y) - b'(\theta)\big\}.$$

Example 5.5 (Instrumental variable regression). *Assume that the goal is to characterize the relationship between a response Y_1 and a treatment Y_2 in the sense that*

$$Y_1 = f(Y_2) + Z,$$

where f is unknown and Z is an error term with zero mean.

When the treatment is endogenous (i.e., not independent of Z) identifying its effect further requires an instrument Y_3 that is exogenous (i.e., independent of Z). More specifically, let $\boldsymbol{Y} = (Y_1, Y_2, Y_3) \in \mathcal{Y} \subseteq \mathbb{R}^3$ and $Z \in \mathcal{Z} \subseteq \mathbb{R}$ be random variables where $\mathrm{E}\big(Z\big) = 0$, and Z is independent of Y_3 but not necessarily of Y_2. In this setting, we can identify f through $\mathrm{E}\big(Y_1|Y_3 = y_3\big) = E\big\{f(Y_2)|Y_3 = y_3\big\}$, which is an ill-posed inverse problem.

This issue can be resolved by assuming that f lives in a compact space (Newey and Powell, 2003), e.g., the space of functions with bounded L_2-norm. With $f(y) \approx \boldsymbol{\theta}^\top b(y)$ for $y \in \mathbb{R}$ an approximation with K coefficients $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ and basis functions $b(y) = (b_1(y), \cdots, b_K(y))$, compactness is imposed by assuming bounded basis coefficients. The estimating equation is then

$$\psi_{\boldsymbol{\theta}}(\boldsymbol{y}) = b(y_3)\big\{y_1 - \boldsymbol{\theta}^\top b(y_2)\big\} \in \mathbb{R}^K.$$

5.3 Asymptotic theory

We now consider the asymptotic properties of $\widehat{\theta}_t$. We use a general framework to encompass a wide range of identifying functions and both parametric and nonparametric estimators of $w_{\boldsymbol{x}}$. In Section 5.3.1, we state and discuss the main results. In Section 5.3.2, we detail our assumptions and point to alternative conditions that are sometimes easier to verify. In Section 5.3.3, we use specific examples to illustrate how the assumptions can be verified in practice. All proofs are relegated to Section 5.6.

5.3.1 Main results

In what follows, we ignore measurability issues, because the functions that we consider are generally well behaved. Recall the definitions of stochastic processes and weak convergence given in Section 2.4.

Consider a collection of identifying functions $\Psi = \{\psi_{\theta,t} \colon \theta \in \Theta, t \in T\}$. Then the estimator $\widehat{\theta}$ and true parameter θ_0 are processes $\{\widehat{\theta}_t \colon t \in T\}$ and $\{\theta_{0,t} \colon t \in T\}$. For example, if $\theta_{0,t}$ is the conditional t-quantile (as in Example 5.2), then $\widehat{\theta}$ and θ_0 are processes indexed by the quantile level $t \in T = (0,1)$. In the case of mean regression, T is a singleton and $\widehat{\theta}$ is an ordinary random variable.

Recall that $\widehat{\theta}_t = \widehat{\theta}_t(\boldsymbol{x})$ and $\theta_{0,t} = \theta_{0,t}(\boldsymbol{x})$ are also functions of the value conditioned upon through $\boldsymbol{X} = \boldsymbol{x}$. This is reflected in their definitions (5.5) and (5.2) through $\widehat{w}_{\boldsymbol{x}}$ and $w_{\boldsymbol{x}}$. However, $\widehat{\theta}_t$ as a process in $\boldsymbol{x}$ cannot have a tight limit in many cases of practical interest. Hence, we assume $\boldsymbol{x}$ fixed for the remainder of this section and simply write $\widehat{\theta}_t$ and $\theta_{0,t}$ instead of $\widehat{\theta}_t(\boldsymbol{x})$ and $\theta_{0,t}(\boldsymbol{x})$, respectively.

In what follows, convergence is always understood as $n \to \infty$. Our first result shows that $\widehat{\theta}_t$ is consistent uniformly in the indexing set T.

Theorem 5.1 (Consistency)**.** *Under (A1)–(A3),* $\sup_{t\in T}\bigl|\widehat{\theta}_t - \theta_{0,t}\bigr| \to_P 0$.

The assumptions will be discussed in detail in Section 5.3.2. One assumption worth mentioning already is that the estimator for the copula density c needs to be consistent. In particular, any parametric model needs to be correctly specified.

While stronger than point-wise consistency, Theorem 5.1 is insufficient for statistical inference. For instance, to test hypotheses and construct confidence bands, an asymptotic distribution is essential. To obtain a weak convergence result, we specify the estimator $\widehat{w}_{\boldsymbol{x}}$ further. In what follows, we assume that $\widehat{w}_{\boldsymbol{x}}$ is asymptotically linear in the sense that

$$\sup_{\boldsymbol{y}\in\mathcal{Y}}\left|\widehat{w}_{\boldsymbol{x}}(\boldsymbol{y}) - \frac{1}{n}\sum_{i=1}^{n} a_{n,\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{Y}_i, \boldsymbol{X}_i)\right| = o_P\bigl(n^{-1/2} r_n\bigr),$$

where $a_{n,\boldsymbol{x}} \colon \mathbb{R}^{2q+p} \to \mathbb{R}$ is a sequence of continuous functions and $r_n^{-1} = O(1)$. This assumption is satisfied by many popular estimators of $w_{\boldsymbol{x}}$, and we refer to Section 5.3.3 for examples. The rate r_n is introduced to encompass both parametric and nonparametric estimators of $w_{\boldsymbol{x}}$ within the same formalism. While r_n is a diverging sequence for nonparametric estimators, $r_n = 1$ gives the standard $\sqrt{n}$ rate for parametric ones.

Theorem 5.2 (Weak convergence). *Under (A1-A8),*

$$r_n^{-1}\sqrt{n}(\widehat{\theta}_t - \theta_{0,t} - \beta_n) \rightsquigarrow -V_{\theta_{0,t},t,w_x}^{-1}\mathbb{G}_t \quad in\ \ell^\infty(T),$$

where $V_{\theta_{0,t},t,w_x}$ *is the Hadamard derivative of* $\mathrm{E}\{\psi_{\theta_{0,t},t}(\boldsymbol{Y})w_{\boldsymbol{x}}(\boldsymbol{Y})\}$ *with respect to* $\theta_{0,t}$, $\beta_n = r_n V_{\theta_{0,t},t,w_x}^{-1}\mathrm{E}\{\omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y},\boldsymbol{X})\}$ *with*

$$\omega_{n,t,\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{x}') = r_n^{-1}\mathrm{E}\{\psi_{\theta_{0,t},t}(\boldsymbol{Y})a_{n,\boldsymbol{x}}(\boldsymbol{Y},\boldsymbol{y},\boldsymbol{x}')\},$$

for all $(\boldsymbol{y},\boldsymbol{x}') \in \mathcal{Y}\times\mathcal{X}$, *and* $\mathbb{G}$ *is a tight, zero-mean Gaussian element with covariance* $\lim_{n\to\infty}\mathrm{cov}\{K_{n,t_1},K_{n,t_2}\}$, *where*

$$K_{n,t} = r_n^{-1}\psi_{\theta_{0,t},t}(\boldsymbol{Y})w_{\boldsymbol{x}}(\boldsymbol{Y}) + \omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y},\boldsymbol{X}).$$

Recall that $\widehat{\theta}_t$ is defined as a plug-in type estimator, obtained by substituting $\widehat{w}_{\boldsymbol{x}}$ for $w_{\boldsymbol{x}}$ to solve the empirical estimating equation. As such, the asymptotic distribution of $\widehat{\theta}_t$ combines the effects of two approximations:

1. replacing $w_{\boldsymbol{x}}$ with $\widehat{w}_{\boldsymbol{x}}$,

2. replacing the true estimating equation by its empirical counterpart (5.5).

Since the approximation in step 2 is unbiased, the bias β_n is driven by the bias of $\widehat{w}_{\boldsymbol{x}}$, but "averaged out" by taking the expectation with respect to $\boldsymbol{Y}$. Furthermore, the sequence converging to the asymptotic variance can be decomposed into $\mathrm{var}\{K_{n,t}\} = A_{n,t}^1 + A_{n,t}^2 + A_{n,t}^3$, where

- $A_{n,t}^1 = r_n^{-2}\mathrm{Var}\{\psi_{\theta_{0,t},t}(\boldsymbol{Y})w_{\boldsymbol{x}}(\boldsymbol{Y})\}$ isolates the contribution of step 2,
- $A_{n,t}^2 = \mathrm{Var}\{\omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y},\boldsymbol{X})\}$ is driven by the variance of $\widehat{w}_{\boldsymbol{x}}$ in step 1, but with a similar "averaging out" as for β_n,
- $A_{n,t}^3 = 2r_n^{-1}E\{\psi_{\theta_{0,t},t}(\boldsymbol{Y})w_{\boldsymbol{x}}(\boldsymbol{Y})\omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y},\boldsymbol{X})\}$ reflects the dependence between the two steps, which is induced by the fact that the same data is used in both.

In the case of nonparametric estimators for $w_{\boldsymbol{x}}$, we have $r_n \to \infty$ and $A_{n,t}^2$ dominates asymptotically.

Theorem 5.2 allows us to compute the exact limiting distribution for specific choices of identifying functions and estimator $\widehat{w}_{\boldsymbol{x}}$. But the limiting distribution also depends on θ_0 and $w_{\boldsymbol{x}}$. In practice, we can only approximate this distribution by substituting $\widehat{\theta}$ and $\widehat{w}_{\boldsymbol{x}}$ for the true values. Such computations are rather involved, especially for complex estimators $\widehat{w}_{\boldsymbol{x}}$. And even for simple ones, they may not suffice: when T is not finite, critical values for confidence bands cannot in general be determined from the limiting distribution.

To remedy this issue, a popular alternative is the *bootstrap* method. The idea is to define a new estimator $\widetilde{\theta}_t$, based on a randomly reweighted version of the data. The distribution of $\sqrt{n}(\widehat{\theta}_t - \theta_{0,t})$ is then approximated by the distribution of $\sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t)$. Although the bootstrap was originally introduced as a resampling technique (Efron, 1979), we use a slightly different formulation that simplifies our asymptotic analysis. Let $\xi_i, i = 1, \dots, n$, be an *iid* sequence of positive random variables independent of the data and satisfying $\mathrm{E}(\xi_1) = \mu \in (0, \infty)$, $\mathrm{var}(\xi_1) = \sigma^2 \in (0, \infty)$, and $\mathrm{E}(|\xi_1|^{2+\epsilon}) < \infty$ for some $\epsilon > 0$. For example, choosing ξ_i to have standard exponential distribution results in the *Bayesian bootstrap* (Rubin, 1981).

Define the bootstrap estimator $\{\tilde{\theta}_t, t \in T\}$ as solving

$$\frac{1}{n}\sum_{i=1}^{n}(\xi_i/\bar{\xi})\psi_{\widetilde{\theta}_t,t}(\mathbf{Y}_i)\widetilde{w}_{\boldsymbol{x}}(\mathbf{Y}_i) = 0,$$

where $\bar{\xi} = n^{-1}\sum_{i=1}^{n}\xi_i$ and

$$\widetilde{w}_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(\xi_i/\bar{\xi})a_{n,\boldsymbol{x}}(\boldsymbol{y}, \mathbf{Y}_i, \boldsymbol{X}_i).$$

For an arbitrary Z_n, let $\widetilde{Z}_n$ be its bootstrapped version with random weights ξ_i, $i = 1, \dots, n$, and Z some tight process. We write $\widetilde{Z}_n \rightsquigarrow_{\xi,P} Z$ if

$$\sup_{h \in BL_1}\left|\mathrm{E}_\xi\{h(\widetilde{Z}_n)\} - \mathrm{E}\{h(Z)\}\right| \to_P 0,$$

where BL_1 is the space of functions with Lipschitz norm bounded by 1 and E_ξ denotes the expectation with respect to the ξ_i's only. Hence, the first expectation above is still a random variable since $\widetilde{Z}_n$ still depends on the observations. This explains why the convergence above is in probability, as opposed to the deterministic convergence in usual weak convergence (Definition 2.3). The convergence $\widetilde{Z}_n \rightsquigarrow_{\xi,P} Z$ can be understood as convergence of the law of $\widetilde{Z}_n$ conditional on the observations (see Section 2.2.3 of Kosorok, 2007).

We can now state the next theorem, which implies that the conditional law of a properly scaled version of $r_n^{-1}\sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t)$ given the observed data converges to the law of $\mathbb{G}_t$.

Theorem 5.3 (Validity of the bootstrap)**.** *Under assumptions (A1)–(A9),*

$$(\mu/\sigma)r_n^{-1}\sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t) \rightsquigarrow_{\xi,P} -V_{\theta_{0,t},t,w_x}^{-1}\mathbb{G}_t \quad \text{in } \ell^\infty(T),$$

where $\mathbb{G}_t$ is as in Theorem 5.2.

Remark 5.2. *The boostrap of Efron (1979) is different in that resampling from the data amounts to choosing $(\xi_1, \dots, \xi_n) \sim \mathrm{Multinomial}(n, 1/n, \dots, 1/n)$, which violates the independence requirement for the bootstrap weights. While this issue can be resolved by a Poissionization argument (see, van der Vaart and Wellner, 1996, Section 3.6), it requires substantial additional effort.*

5.3.2 Assumptions

For an arbitrary class of functions $\mathcal{G}$, a function G is called *envelope* of $\mathcal{G}$ if $G \geq \sup_{g \in G} |g|$ pointwise. To further simplify notation here and in the proofs, we define $g_{\theta_t, t, w} = \psi_{\theta_t, t} w$ so that

$$\mathrm{E}\{g_{\theta_t, t, w_{\boldsymbol{x}}}(\boldsymbol{Y})\} = \mathrm{E}\big\{\psi_{\theta_t, t}(\boldsymbol{Y}) | \boldsymbol{X} = \boldsymbol{x}\big\}.$$

Assumption 5.1. *Suppose that $\boldsymbol{x} \in \mathcal{X}$ is fixed.*

(A1) The estimator $\widehat{w}_{\boldsymbol{x}}$ satisfies $|\widehat{w}_{\boldsymbol{x}}(\boldsymbol{y}) - w_{\boldsymbol{x}}(\boldsymbol{y})| \to_P 0$ for all $\boldsymbol{y} \in \mathcal{Y}$.

(A2) There is a class of weight functions $\mathcal{W}$ and $\delta > 0$ such that

(i) $\mathcal{G}_\delta = \{g_{\theta_t, t, w} \colon |\theta_t - \theta_{0,t}| < \delta, t \in T, w \in \mathcal{W}\}$ is P-Donsker,

(ii) $\mathrm{P}(\widehat{w}_{\boldsymbol{x}} \in \mathcal{W}) \to 1$.

(A3) there exists $\theta_0 \in \Theta$ such that for all $\epsilon > 0$, $\inf_{\|\theta - \theta_0\| > \epsilon} |E\{g_{\theta, t, w_{\boldsymbol{x}}}(\boldsymbol{Y})\}| > 0$ and $E\{g_{\theta_0, t, w_{\boldsymbol{x}}}(\boldsymbol{Y})\} = 0$.

(A4) The map $\theta \mapsto \psi_{\theta, t}$ is continuous in $L_2(P)$ at $\theta_{0,t}$.

(A5) The map $\theta \mapsto \mathrm{E}\{g_{\theta, t, w_{\boldsymbol{x}}}(\boldsymbol{Y})\}$ is Hadamard differentiable in a neighborhood of $\theta_{0,t}$ and the derivatives $V_{\theta_{0,t}, t, w_{\boldsymbol{x}}}$ are invertible.

(A6) for $\Psi_{\theta_{0,t}} = \{\psi_{\theta_{0,t}, t} \colon t \in T\}$, it holds

$$\int_0^{\delta_n} \sqrt{\log N\big\{\epsilon/2, \Psi_{\theta_{0,t}}, L_1(P)\big\}} d\epsilon \to 0, \qquad \text{for every } \delta_n \downarrow 0.$$

(A7) There is a sequence of continuous functions $a_{n,\boldsymbol{x}} \colon \mathbb{R}^{2q+p} \to \mathbb{R}$ and $r_n^{-1} = O(1)$ such that

$$\sup_y \left| \widehat{w}_{\boldsymbol{x}}(\boldsymbol{y}) - \frac{1}{n} \sum_{i=1}^n a_{n,\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{Y}_i, \boldsymbol{X}_i) \right| = o_P\big(n^{-1/2} r_n\big), \quad \sup_{\boldsymbol{s}} |a_{n,\boldsymbol{x}}(\boldsymbol{s}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{x}})| < \infty.$$

(A8) With $\omega_{n,t,\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{x}) = r_n^{-1} \mathrm{E}\{\psi_{\theta_{0,t}, t}(\boldsymbol{Y}) a_{n,\boldsymbol{x}}(\boldsymbol{Y}, \boldsymbol{y}, \boldsymbol{x})\}$, the sequence of function classes $\mathcal{O}_n = \{\omega_{n,t,\boldsymbol{x}} \colon t \in T\}$ admits envelopes $\Omega_n \colon \mathbb{R}^{2q+p} \to \mathbb{R}$ satisfying for every $\epsilon > 0$

$$\mathrm{E}\{\Omega_n^2(\boldsymbol{Y}, \boldsymbol{X})\} = O(1), \qquad \mathrm{E}\{\Omega_n^2(\boldsymbol{Y}, \boldsymbol{X}) \mathbb{1}(\Omega_n > \epsilon\sqrt{n})\} \to 0.$$

(A9) The bootstrap estimator $\widetilde{w}_{\boldsymbol{x}}$ satisfies $|\widetilde{w}_{\boldsymbol{x}}(\boldsymbol{y}) - w_{\boldsymbol{x}}(\boldsymbol{y})| \to_P 0$ and $\mathrm{P}(\widetilde{w}_{\boldsymbol{x}} \in \mathcal{W}) \to 1$ with $\mathcal{W}$ is as in (A2).

First, note that (A1) and (A2ii) together imply that $w_{\boldsymbol{x}} \in \mathcal{W}$. Then for any θ_t, $t \in T$, such that $|\theta_t - \theta_{0,t}| < \delta$, we have $g_{\theta_t, t, w_{\boldsymbol{x}}} \in \mathcal{G}_\delta$. (A2i) is crucial to prove weak-convergence of $\widehat{\theta}_t$ and, along with (A2ii), arguably the hardest to verify. There is a trade off between them: the larger the class $\mathcal{W}$, the easier it is to show (A2ii), but the harder it is to show (A2i) and vice versa. Also, since $w_{\boldsymbol{x}} \in \mathcal{W}$, the class $\mathcal{W}$ must be large enough to include the true weight function.

Assumption (A3) implies that the true estimating equation uniquely identifies the parameter of interest. Assumption (A9) is the bootstrap equivalent of (A1) and (A2ii) and could be replaced by a condition on $a_{n,\boldsymbol{x}}$ that is similar to (A8). Assumptions (A3) and (A9) are easily verified for most cases of practical interest (such as the ones in the Section 5.2.3), and will not be discussed further.

In Section 5.3.2 and Section 5.3.2, we give alternative conditions to (A2) that are less general but easier to verify. Assumptions (A4), (A5) and (A6) are easily verifiable for many identifying functions of practical interest (see Section 5.3.3 for quantiles, expectiles and exponential families). Assumptions (A1), (A7) and (A8) allow to use the same framework to obtain general results for a wide class of estimators. Section 5.3.3 discusses both parametric and nonparametric examples.

Assumption (A2i)

Recalling the definition of $\mathcal{G}_\delta$ and defining $\Psi_\delta = \cup_{\|\theta-\theta_{0,t}\|<\delta}\{\psi_{\theta,t} \colon t \in T\}$, we get that $\mathcal{G}_\delta = \Psi_\delta \cdot \mathcal{W}$. Consider the map $\phi(u,v) = uv$ for each $(u,v) \in \mathbb{R}^2$. (A2i) amounts to showing that $\phi \circ (\Psi_\delta, \mathcal{W})$ is P-Donsker. Define $\bar{\psi}$ and $\bar{w}$ as any envelope function of Ψ_δ and $\mathcal{W}$, respectively. Then for every $\psi_1, \psi_2 \in \Psi_\delta$ and $w_1, w_2 \in \mathcal{W}$,

$$\begin{aligned}&|\psi_1(\boldsymbol{y})w_1(\boldsymbol{y}) - \psi_2(\boldsymbol{y})w_2(\boldsymbol{y})|^2 \\ &\leq 2\big\{\bar{\psi}^2(\boldsymbol{y})|w_1(\boldsymbol{y}) - w_2(\boldsymbol{y})|^2 + \bar{w}^2(\boldsymbol{y})|\psi_1(\boldsymbol{y}) - \psi_2(\boldsymbol{y})|^2\big\}.\end{aligned}$$

We abbreviate $\|P\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |\mathrm{E}\{g(\boldsymbol{Y})\}|$ for some class $\mathcal{G}$ of measurables functions from $\mathcal{Y}$ to $\mathbb{R}$. With the inequality from the last display, Corollarly 2.10.13 of van der Vaart and Wellner (1996) leads to the following refinement.

Lemma 5.1. *If there are envelope functions $\bar{\psi}$ and $\bar{w}$ such that*

(B1) $\bar{\psi}\mathcal{W}$ and $\bar{w}\Psi_\delta$ are P-Donsker,

(B2) $\|P\|_{\bar{\psi}\mathcal{W}} < \infty$ and $\|P\|_{\bar{w}\Psi_\delta} < \infty$,

(B3) $E\{g(\boldsymbol{Y})\}^2 < \infty$ for at least one $g \in \mathcal{G}_\delta$,

then (A2i) holds.

We now give a set of sufficient conditions for (B1)–(B3) to hold and discuss how they can be verified. Denote by $C_M^\beta(\mathcal{Y})$ the class of functions whose partial derivatives up to order $\lfloor\beta\rfloor$ (the greatest integer smaller than β) are uniformly bounded by M and whose highest partial derivatives are Lipschitz of order $\beta - \lfloor\beta\rfloor$.

Lemma 5.2. *Let $\mathcal{Y} \subseteq \mathbb{R}^q$. If*

(C1) $\mathcal{W} = C_M^\beta(\mathcal{Y})$ *with* $\beta > q/2$,

(C2) Ψ_δ *is P-Donsker and* $\|P\|_{\Psi_\delta} < \infty$,

(C3) there exists a partition $\mathcal{Y} = \cup_{j=1}^\infty \mathcal{Y}_j$ *such that* $\bar{\psi}|_{\mathcal{Y}_j} \in C_{M_j}^\beta(\mathcal{Y}_j)$ *and* $\sum_{j=1}^\infty M_j P(\mathcal{Y}_j)^{1/2} < \infty$,

then (B1) holds.

(C1) might seem restrictive since usual copula densities are unbounded in the corners of the unit hypercube, thus implying $c_{\boldsymbol{YX}}(F_{\boldsymbol{Y}}(\boldsymbol{y}), F_{\boldsymbol{X}}(\boldsymbol{x})) \notin C_M^\beta(\mathcal{Y} \times \mathcal{X})$. However, the assumption only requires $w_{\boldsymbol{x}} \in C_M^\beta(\mathcal{Y})$, namely boundedness of the copula density for fixed $\boldsymbol{x}$. (C2) is easily verified for quantiles, expectiles, and exponential families; see Section 5.3.3.

(C3) is used to prove that $\bar{\psi}\mathcal{W}$ is P-Donsker. If $\mathcal{Y}$ is a bounded, convex subset of $\mathbb{R}^q$ with nonempty interior and $\bar{\psi} \in C_M^\beta(\mathcal{Y})$ for some $M < \infty$, then $\mathcal{W}$ has a finite bracketing integral (as in Theorem 2.2) provided $\beta > q/2$, by van der Vaart and Wellner (1996), Theorem 2.7.1. As such, it is P-Donsker and so is $\bar{\psi}\mathcal{W}$.

Furthermore, if $\mathcal{Y} = \mathbb{R}$ and $M_j = M$ for all j, then $\sum_{j=1}^\infty M_j P(\mathcal{Y}_j)^{1/2} < \infty$ can be replaced by a tail condition like $E(|Y|^{2+\epsilon}) < \infty$ for some $\epsilon > 0$ by van der Vaart and Wellner (1996), Corollary 2.7.4. If M_j is not constant, but satisfies $M_j \sim \sup_{y \in \mathcal{Y}_j} |y|$, then a sufficient condition for $\sum_{j=1}^\infty M_j P(\mathcal{Y}_j)^{1/2} < \infty$ is that $E(|Y|^{4+\epsilon}) < \infty$.

Assumption (A2ii)

(A2ii) requires that the estimator $\widehat{w}_{\boldsymbol{x}}$ lies in a P-Donsker class with probability going to one. In the setting of Lemma 5.2, a sufficient condition is that $\widehat{w}_{\boldsymbol{x}}$'s first $\lfloor \beta \rfloor + 1$ derivatives are uniformly consistent for the respective derivatives of $w_{\boldsymbol{x}}$.

Lemma 5.3. *If (C1), (C2), (C3), and*

(C4) $\sup_{\boldsymbol{y}} |\partial^k(\widehat{w}_{\boldsymbol{x}} - w_{\boldsymbol{x}})(\boldsymbol{y})| = o_P(1)$ *for all* $|k| = 0, \ldots, \lfloor \beta \rfloor + 1$,

with $k = (k_1, \cdots, k_q)$, $\partial^k = \partial_1^{k_1} \cdots \partial_q^{k_q}$, *and* $|k| = \sum_1^q k_j$, *then (A2ii) holds.*

An estimator $\widehat{w}_{\boldsymbol{x}}$ will typically have the form

$$\widehat{w}_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{\widehat{c}_{\boldsymbol{Y},\boldsymbol{X}}\{\widehat{F}_{Y_1}(y_1), \ldots, \widehat{F}_{Y_q}(y_q), \widehat{F}_{X_1}(x_1), \ldots, \widehat{F}_{X_p}(x_p)\}}{\widehat{c}_{\boldsymbol{Y}}\{\widehat{F}_{Y_1}(y_1), \ldots, \widehat{F}_{Y_q}(y_q)\}}.$$

where $\widehat{F}_{Y_j}, \widehat{F}_{X_k}$ are estimators of the marginal distributions and $\widehat{c}_{\boldsymbol{Y},\boldsymbol{X}}$, $\widehat{c}_{\boldsymbol{Y}}$ are estimators of the copula density. Then (C4) is satisfied if $\widehat{c}_{\boldsymbol{Y},\boldsymbol{X}}$, $\widehat{c}_{\boldsymbol{Y}}$, and $\widehat{F}_{Y_j}$

(as functions of $\boldsymbol{y}$) have uniformly converging derivatives up to order $\lfloor\beta\rfloor + 1$. This is true for sufficiently smooth parametric models and most classical nonparametric techniques like Bernstein polynomial and kernel estimators (see e.g., Gawronski, 1985, Silverman, 1978). Note that (C4) technically excludes the empirical distribution function as an estimator for the marginal distributions, because it is not differentiable. This is unlikely to be an issue in practice, however, since the empirical distribution is at least first-order equivalent to a smoothed version for which (C4) does hold.

5.3.3 Examples

Identifying functions

Assumption (A4) is obvious in all following examples and not discussed further.

Example 5.2 (Quantile regression, cont.). *For (A5), we see that the map*

$$\theta \mapsto \mathrm{E}\{g_{\theta,t,w_{\boldsymbol{x}}}(Y)\} = t - F_{Y\mid\boldsymbol{X}}(\theta \mid \boldsymbol{x}),$$

is differentiable at θ with invertible derivative $V_{\theta_t,t,w_{\boldsymbol{x}}} = -f_{Y\mid\boldsymbol{X}}(\theta_t \mid \boldsymbol{x})$. The class Ψ_δ has a smooth envelope $\bar{\psi}_{\theta,t}(y) = 1$ and contains only monotone functions from $\mathbb{R}$ to $[-1,1]$. Hence, $\log N_{[]}\{\epsilon, \Psi_\delta, L_r(P)\} = O(\epsilon^{-1})$ for all $r \geq 1$ by Theorem 2.7.5 in van der Vaart and Wellner (1996). In particular, Ψ_δ is P-Donsker (C2), and (A6) follows from the inequality $N(\epsilon/2, \mathcal{G}, \|\cdot\|) \leq N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|)$ for any class $\mathcal{G}$ and norm $\|\cdot\|$. (C3) is trivially satisfied since $\bar{\psi} \equiv 1$; (C1), (B1), and (B3) become mild conditions on $\mathcal{W}$ and F_Y.

Example 5.3 (Expectile regression, cont.). *We have the map (Zwingmann and Holzmann, 2017)*

$$\theta \to \mathrm{E}\{g_{\theta,t,w_{\boldsymbol{x}}}(Y)\} = t\int_\theta^\infty \{1 - F_{Y\mid X}(y \mid \boldsymbol{x})\}dy - (1-t)\int_{-\infty}^\theta F_{Y\mid X}(y \mid \boldsymbol{x})dy,$$

and its invertible derivative is $V_{\theta,t,w_{\boldsymbol{x}}} = (2t-1)F_{Y\mid X}(\theta \mid \boldsymbol{x}) - t$, which implies (A5). For the special case of mean regression ($t = 1/2$), this yields $V_{\theta,1/2,w_{\boldsymbol{x}}} = -1/2$. If t takes values in a compact subset of $(0,1)$, Zwingmann and Holzmann (2017) show that Ψ_δ is P-Donsker (C2) by deriving bounds on its bracketing numbers. One can similarly verify that the covering integral condition (A6) is satisfied. It further holds that

$$\begin{aligned}
|\psi_{\theta_t,t}(y)| &= \bigl|t(y-\theta_t)\mathbb{1}(y \geq \theta_t) - (1-t)(\theta_t - y)\mathbb{1}(y < \theta_t)\bigr| \\
&\leq |t|\,|y-\theta_t|\mathbb{1}(y \geq \theta_t) + |1-t|\,|\theta_t - y|\mathbb{1}(y < \theta_t) \\
&\leq |y-\theta_t| + |\theta_t - y| \\
&\leq 2(|\theta_t| + |y|),
\end{aligned}$$

for all $t \in T$ and $y \in \mathbb{R}$. Using the fact that $|y| \leq 1 + y^2$ for all $y \in \mathbb{R}$. we obtain *a possible smooth envelope for* Ψ_δ *as*

$$\bar{\psi}(y) = \sup_{t \in T} \sup_{|\theta_t - \theta_{0,t}| < \delta} 2\big(|\theta_t| + 1 + y^2\big).$$

With this, (C3), (C1), (B1), and (B3) again become mild conditions on $\mathcal{W}$ *and* F_Y*.*

Example 5.4 (Exponential family regression, cont.)**.** *We have*

$$\theta \to \mathrm{E}\{g_{\theta, w_{\boldsymbol{x}}}(Y)\} = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) - b'(\theta)\mathrm{E}\{w_{\boldsymbol{x}}(Y)\},$$

and its derivative is $V_{\theta,w_{\boldsymbol{x}}} = -b''(\theta)\mathrm{E}\{w_{\boldsymbol{x}}(Y)\}$*, which implies (A5) provided that* $b''(\theta)$ *exists in a neighborhood of* $\theta_{0,t}$ *and* $b''(\theta_0) \neq 0$*. Since there is no indexing parameter* t*, conditions (A2i) and (A6) only relate to the function* b' *and the true weight* $w_{\boldsymbol{x}}$ *which are usually well behaved.*

Example 5.5 (Instrumental variable regression, cont.)**.** *We have*

$$\theta \to \mathrm{E}\{g_{\boldsymbol{\theta}, w_x}(\boldsymbol{Y})\} = \mathrm{E}\big[b(Y_3)\big\{Y_1 - \boldsymbol{\theta}^\top b(Y_2)\big\} \mid \boldsymbol{X} = \boldsymbol{x}\big],$$

and its derivative matrix is $V_{\theta,w_{\boldsymbol{x}}} = -E\big\{b(Y_3)b(Y_2)^\top \mid \boldsymbol{X} = \boldsymbol{x}\big\}$*, which implies (A5) for reasonable choices of basis functions. Similarly, (A2i) and (A6) only relate to the basis, which can be chosen arbitrarily. For instance,* $b(y) = (1, y)^\top$ *gives a linear treatment effect, and*

$$V_{\theta,w_{\boldsymbol{x}}} = \begin{pmatrix} 1 & \mathrm{E}\big(Y_2 \mid X = x\big) \\ \mathrm{E}\big(Y_3 \mid X = x\big) & \mathrm{E}\big(Y_2 Y_3 \mid X = x\big) \end{pmatrix}.$$

Estimators for the weight function

In this section, we discuss parametric and nonparametric estimators of $w_{\boldsymbol{x}}$. For the sake of simplicity, we focus on $\mathcal{Y} \subseteq \mathbb{R}$, while stressing that similar arguments follow for multivariate responses.

Example 5.6 (Fully parametric estimator)**.** *Let* $\boldsymbol{\eta}_Y \in H_Y \subseteq \mathbb{R}^{d_Y}$*,* $\boldsymbol{\eta}_{\boldsymbol{X}} \in H_X \subseteq \mathbb{R}^{d_X}$ *and* $\boldsymbol{\eta}_C \in H_C \subseteq \mathbb{R}^{d_C}$ *be parameters, indexing a family of marginal and copula densities. We further write* $\boldsymbol{\eta} = (\boldsymbol{\eta}_Y, \boldsymbol{\eta}_X, \boldsymbol{\eta}_C)$ *and* $\widehat{\boldsymbol{\eta}}$ *for the maximum-likelihood estimator of the true* $\boldsymbol{\eta}_0$*, i.e.,*

$$\widehat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \sum_{i=1}^n \nabla_{\boldsymbol{\eta}} \log f_{Y\boldsymbol{X}}(Y_i, \boldsymbol{X}_i; \boldsymbol{\eta}).$$

We denote $w_{\boldsymbol{x}}(y) = w_{\boldsymbol{x}}(y; \boldsymbol{\eta}_0)$ *and* $\widehat{w}_{\boldsymbol{x}}(y) = w_{\boldsymbol{x}}(y; \widehat{\boldsymbol{\eta}})$ *the true and estimated weight functions. Assumption (A1) is satisfied except for pathological cases. Provided that* $w_{\boldsymbol{x}}(\cdot; \boldsymbol{\eta})$ *is sufficiently smooth in* $\boldsymbol{\eta}$*, (C4) is also an immediate consequence of the consistency of* $\widehat{\boldsymbol{\eta}}$*. Furthermore, (A7) holds with* $r_n = 1$*.*

Recall that the maximum likelihood estimator satisfies

$$\begin{aligned}&\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\\&=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\big\{-\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}\log f_{Y\boldsymbol{X}}(Y,\boldsymbol{X};\boldsymbol{\eta}_0)\big\}^{-1}\nabla_{\boldsymbol{\eta}}\log f_{Y\boldsymbol{X}}(Y_i,\boldsymbol{X}_i;\boldsymbol{\eta}_0)+o_p(n^{-1/2}).\end{aligned}$$

Using a Taylor expansion for $\widehat{w}_{\boldsymbol{x}}(y)=w_{\boldsymbol{x}}(y;\widehat{\boldsymbol{\eta}})$ around $\boldsymbol{\eta}_0$, we obtain

$$\begin{aligned}&w_{\boldsymbol{x}}(y;\widehat{\boldsymbol{\eta}})-w_{\boldsymbol{x}}(y;\boldsymbol{\eta}_0)\\&=\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\eta}}w_{\boldsymbol{x}}(y)\mathrm{E}\big\{-\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}\log f_{Y\boldsymbol{X}}(Y,\boldsymbol{X};\boldsymbol{\eta}_0)\big\}^{-1}\nabla_{\boldsymbol{\eta}}\log f_{Y\boldsymbol{X}}(Y_i,\boldsymbol{X}_i;\boldsymbol{\eta}_0)\\&\quad+o_p(n^{-1/2}),\end{aligned}$$

uniformly in y and, hence,

$$\begin{aligned}&a_{n,\boldsymbol{x}}(y,\bar{y},\bar{\boldsymbol{x}})\\&=w_{\boldsymbol{x}}(y)+\nabla_{\boldsymbol{\eta}}w_{\boldsymbol{x}}(y)\mathrm{E}\big\{-\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}\log f_{Y\boldsymbol{X}}(Y,\boldsymbol{X};\boldsymbol{\eta}_0)\big\}^{-1}\nabla_{\boldsymbol{\eta}}\log f_{Y\boldsymbol{X}}(\bar{y},\bar{\boldsymbol{x}};\boldsymbol{\eta}_0).\end{aligned}$$

Then condition (A8) is satisfied under a simple regularity condition like

$$\mathrm{E}\bigg\{\int\bar{\psi}(y)a_{n,\boldsymbol{x}}(y,Y,\boldsymbol{X})f_Y(y)dy\bigg\}^2<\infty.$$

Example 5.7 (Simple kernel estimator)**.** *For any univariate margin Z, the classical kernel estimator of the cumulative distribution function is*

$$\widehat{F}_Z(y)=\frac{1}{nb_n}\sum_{i=1}^{n}\int_{-\infty}^{y}K\bigg(\frac{s-Z_i}{b_n}\bigg)ds,$$

where K is a Lipschitz continuous probability density function and $b_n>0$ a bandwidth parameter. And using a multivariate analog for the density, we obtain the estimator

$$\widehat{w}_{\boldsymbol{x}}(y)=\frac{1}{nh_n^{1+p}}\sum_{i=1}^{n}K\bigg(\frac{\widehat{F}_Y(y)-\widehat{F}_Y(Y_i)}{h_n}\bigg)\prod_{k=1}^{p}K\bigg(\frac{\widehat{F}_{X_k}(y_k)-\widehat{F}_{X_k}(X_{i,k})}{h_n}\bigg).,$$

where $h_n>0$ is the bandwidth for $\widehat{w}_{\boldsymbol{x}}$. When b_n vanishes at an appropriate rate, the marginal estimators can be shown to be uniformly $\sqrt{n}$-consistent and, thus, asymptotically negligible for estimating $w_{\boldsymbol{x}}$ (Silverman, 1978) . Hence, we get (A7) with

$$a_{n,\boldsymbol{x}}(y,\bar{y},\bar{\boldsymbol{x}})=\frac{1}{h_n^{1+p}}K\bigg(\frac{F_Y(y)-F_Y(\bar{y})}{h_n}\bigg)\prod_{k=1}^{p}K\bigg(\frac{F_{X_k}(x_k)-F_{X_k}(\bar{x}_k)}{h_n}\bigg)$$

for any $r_n^{-1} = o(1)$. Assumption (A1) follows from the consistency of the classical kernel density estimator. If $h_n \to 0$ and $nh_n^{p+2(\lfloor\beta\rfloor+1)}/\ln n \to \infty$, (C4) can be verified (e.g., Hansen, 2008), but only for suprema over compact interior subsets of $\mathcal{Y}$. The reason is that the simple form of this estimator induces boundary bias for $F_Y(y)$ close to 0 or 1. Nevertheless, it is possible to show that derivatives of $\widehat{w}_{\boldsymbol{x}}$ converge to smooth functions provided $w_{\boldsymbol{x}}$ satisfies standard regularity conditions, i.e., (A2ii) holds when $\mathcal{W}$ is a smoothness class. Assumption (A8) can be verified for $r_n = h_n^{p/2}$ with tedious calculations using standard arguments for kernel smoothers. Taking $h_n \sim n^{-1/(4+p)}$, the estimator $\widehat{\theta}_t$ converges at rate $O_P\{n^{-2/(4+p)}\}$ which is optimal for nonparametric regression problems with p covariates (Stone, 1980).

More sophisticated versions of $\widehat{w}_{\boldsymbol{x}}$ include semiparametric estimators (Genest et al., 1995) and nonparametric techniques that specialize on copula densities (for a review, see, Nagler et al., 2017). They can be treated in a similar manner, but require more effort.

5.4 Simulations

In this section, we illustrate the versatility of our approach and compare them to state-of-the-art methods in various contexts. Based on the theory developed in Section 5.3, we build one parametric and two nonparametric estimators of $\widehat{w}_{\boldsymbol{x}}$. The first *EE Gaussian copula* (EE for estimating equation) is fully parametric and constructed by fitting fitting Gaussian marginal distributions and a Gaussian copula using the maximum-likelihood estimator at each step. The other two are fully nonparametric and use kernel estimators for the marginal distributions with the direct plug-in methodology to select the bandwidth (Sheather and Jones, 1991), but they differ in the copula estimator. The method *EE kernel density*, uses a multivariate Gaussian kernel estimator (see Example 5.6) for the copula, with a bandwidth equal to $n^{-2/(p+4)}\Sigma$, where n is the sample size and Σ the covariance matrix of the data on the copula scale (i.e., transformed through the margins). The method *EE kernel vine*, uses a nonparametric vine estimator for the copula (method `tll2` in Nagler et al., 2017).

5.4.1 Linear Gaussian model and mean regression

In this section, we build intuition about the inner workings of our method by using the simplest example: a linear Gaussian model, that is $Y = \boldsymbol{\beta}\boldsymbol{X} + Z$, with $\boldsymbol{X} \in \mathbb{R}^p$ a vector of *iid* $N(0,1)$ random variables, $\boldsymbol{\beta} = (1,\dots,1)$, and $Z \sim N(0,1)$ independent of X. If the goal is to estimate $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^p$, the main advantage of this example is that the copula is Gaussian. As such, the first estimator described above is correctly specified, and can be compared to the ordinary least-square estimator (OLS).

In Figure 5.1, we show the results of a simulation study using various estimators, sample sizes, and covariates dimension. All results are based on 250 Monte-Carlo

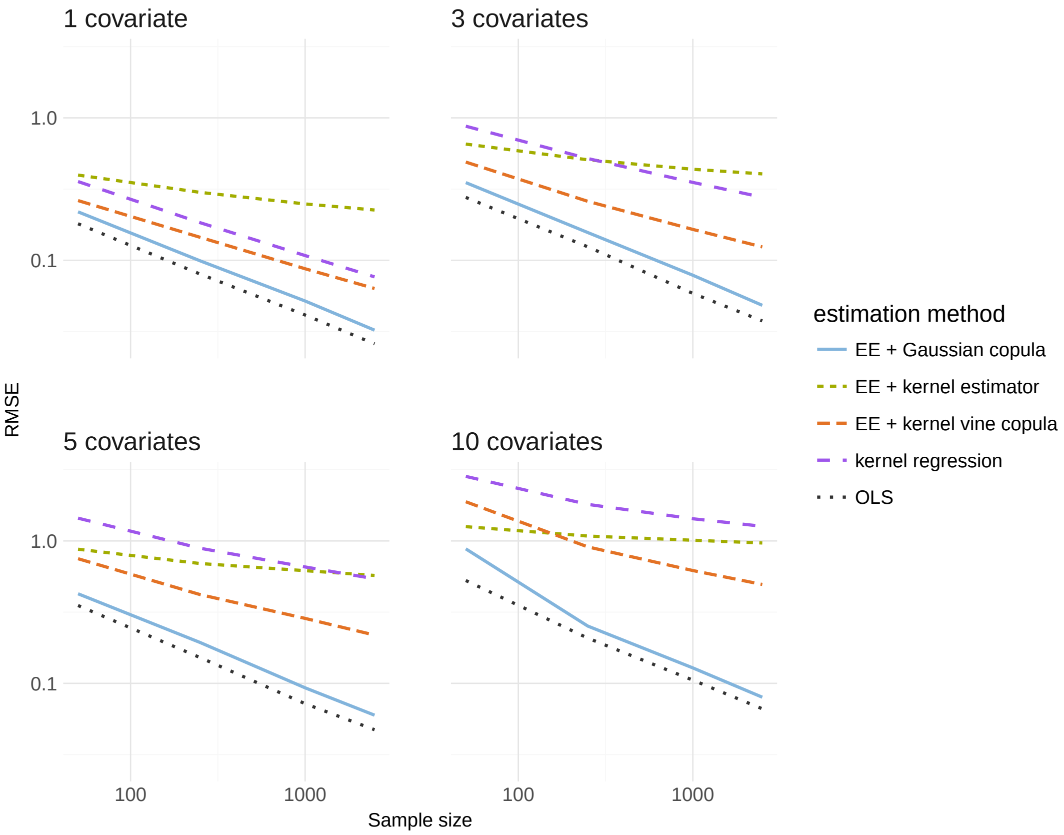


Figure 5.1: Linear Gaussian model and mean regression: root mean squared error for increasing sample sizes (based on 250 replications).

replications. The competitor methods are the OLS estimator and a d-variate kernel regression estimator using least-squares cross validation for the bandwidth (Li and Racine, 2004).

There are two main observations resulting from this simulation study. First, the $\sqrt{n}$ rate in the root-mean-square error (RMSE) clearly appears for both parametric methods. As expected however, the OLS is more efficient, with the relative efficiency of the copula-based estimator being between 60% and 70%. Second, the curse of dimensionality affects the convergence rates of the *EE kernel density* and *kernel regression* estimators. For the kernel estimator based on vine copulas however, the convergence rate of a nonparametric regression with a single covariate is retained. This is no surprise, since it is based on a structural assumption that lifts the curse of dimensionality (Nagler and Czado, 2016).

5.4.2 Quantile regression

In this section, we illustrate how our method performs in the context of quantile regression. We assume that $Y = f(\boldsymbol{X}, Z)$, with $\boldsymbol{X} \in \mathbb{R}^p$ a vector of *iid* $U(-1, 1)$ random variables, $Z \sim N(0, 1)$ independent of $\boldsymbol{X}$, and that the goal is to estimate

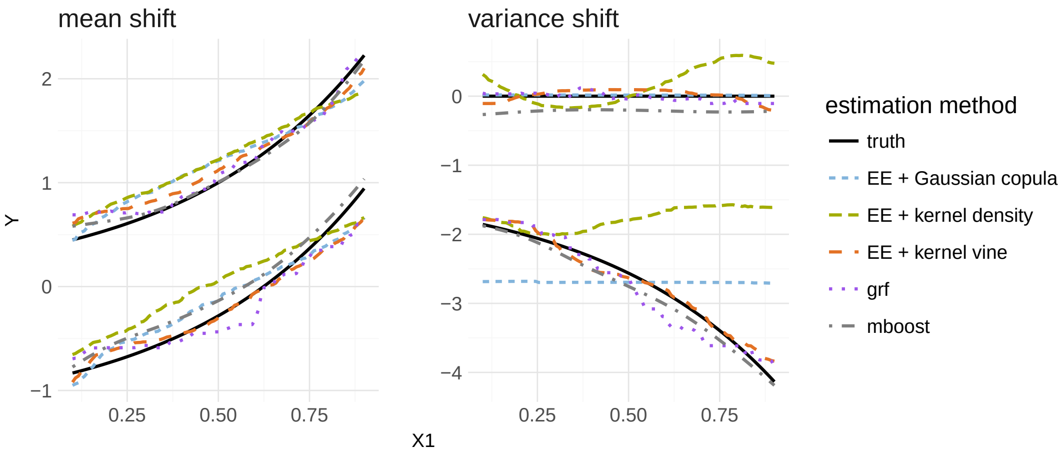


Figure 5.2: Quantile regression with 1'000 observations and 10 covariates for the quantile levels $\alpha = 0.1, 0.5$. The true conditional quantile functions and estimates from various methods.

$F^{-1}_{Y|\boldsymbol{X}}(\alpha \mid \boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^p$ and $\alpha \in (0, 1)$.

In Figure 5.2, we show the results of a small simulation experiment with $n=$ 1'000 observations and $p = 10$ covariates. The true conditional α-quantile ($\alpha = 0.1, 0.5$) is shown as a solid line together with estimates derived by various methods. In the left and right panels, we use $Y = \exp(X_1) + Z$ and $Y = \{1 + \exp(X_1)\}Z$, corresponding respectively to mean and variance shifts. In both cases, Y depends on $\boldsymbol{X}$ only through its first component, and the other covariates are irrelevant. In the legend, *mboost* corresponds to an estimator based on gradient boosting (Hothorn et al., 2010), implemented in the R package Hothorn et al. (2016). Furthermore, *grf* corresponds to an estimator based on generalized random forests (Athey et al., 2017), implemented in the R package Tibshirani et al. (2018).

For the mean shift, it seems that all methods perform appropriately: the estimated quantile curves have the right magnitude and are increasing in X_1. Even predictions from *EE Gaussian copula*, which is misspecified, are reasonably close to the truth. For the variance shift, the picture is different. While it is clear that *EE Gaussian copula*, by assumption, is not suited to capture variance shifts, it appears that *EE kernel density* also struggles, most likely due to both the curse of dimensionality and the choice of bandwidth. However, *EE kernel vine* is on par with state-of-the-art methods. Note that, for *mboost*, the step size (i.e., the ν argument of `boost_control`) was fine-tuned to match the true quantiles, as the variance shift is not captured using the default setup. It is also noteworthy that neither *EE kernel vine* nor *grf* needed parameter tuning to capture the shift in variance, and that predictions from *grf* are considerably more wiggly using the default setup.

5.4.3 Instrumental variable regression

In this section, we illustrate how our method performs in the context of instrumental variable regression. We assume that

$$\begin{aligned} Y_1 &= f(Y_2, \boldsymbol{X}) + Z, \\ Y_2 &= Y_3 + V \\ Z &= -V + W \end{aligned}$$

with $\boldsymbol{X} \in \mathbb{R}^p$ a vector of *iid* $N(0,1)$ random variables, and Y_3, V and W are three other standard Gaussian random variables independent of $\boldsymbol{X}$ and each other, and that the goal is to estimate $f(y, \boldsymbol{x})$ for $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$. In other words, Y_1 is the response, Y_2 is the endogenous treatment, and Y_3 is an exogenous instrument. For instance, when $f(Y_2, \boldsymbol{X}) = g(\boldsymbol{X})Y_2$, the effect of Y_2 on Y_1 is

$$g(\boldsymbol{x}) = \frac{\operatorname{Cov}\left(Y_1, Y_3 \mid \boldsymbol{X} = \boldsymbol{x}\right)}{\operatorname{Cov}\left(Y_2, Y_3 \mid \boldsymbol{X} = \boldsymbol{x}\right)}. \tag{5.7}$$

In Figure 5.3, we show the results resulting from an experiment with n =10’000 observations and $p = 10$ covariates. The panels shows the true regression function f as a function of Y_2 and estimates based on the various methods. The upper panels corresponds to the model $f(Y_2, \boldsymbol{X}) = (1 + X_1)Y_2$ (linear effect) and the lower to $f(Y_2, \boldsymbol{X}) = (1 + X_1)Y_2^2$ (quadratic effect). In both cases, Y_1 depends on $\boldsymbol{X}$ only through its first component, and the other covariates are irrelevant. Columns correspond to two different values of the conditioning variable X_1. In the legend, *grf* corresponds to an estimator based on generalized random forests (Athey et al., 2017), implemented in the R package Tibshirani et al. (2018). Furthermore, *EE* estimators are built by feeding the estimated weights, along with Y_1, Y_2 and Y_3, to the `crsiv` function from the R package Racine and Nie (2018). This function then solves for the basis coefficients (see Example 5.5) using the method of Darolles et al. (2011). Note that `crsiv` could be fed the covariates vector $\boldsymbol{X}$ directly to solve this problem, instead of using weights from the approach of this paper. However, such an approach is computationally burdensome and actually infeasible in practice except for very low dimensional $\boldsymbol{X}$.

For the linear effect, it seems that all methods provide reasonable estimates. Even predictions from *EE Gaussian*, which is misspecified, are close to the truth. For the quadratic effect, the picture is different. It is clear that *grf*, which aims at estimating (5.7), cannot capture such a functional relationship. Furthermore, while *EE Gaussian* and *EE kernel vine* perform appropriately, *EE kernel density* struggles, most likely due to the reasons already described in Section 5.4.2.

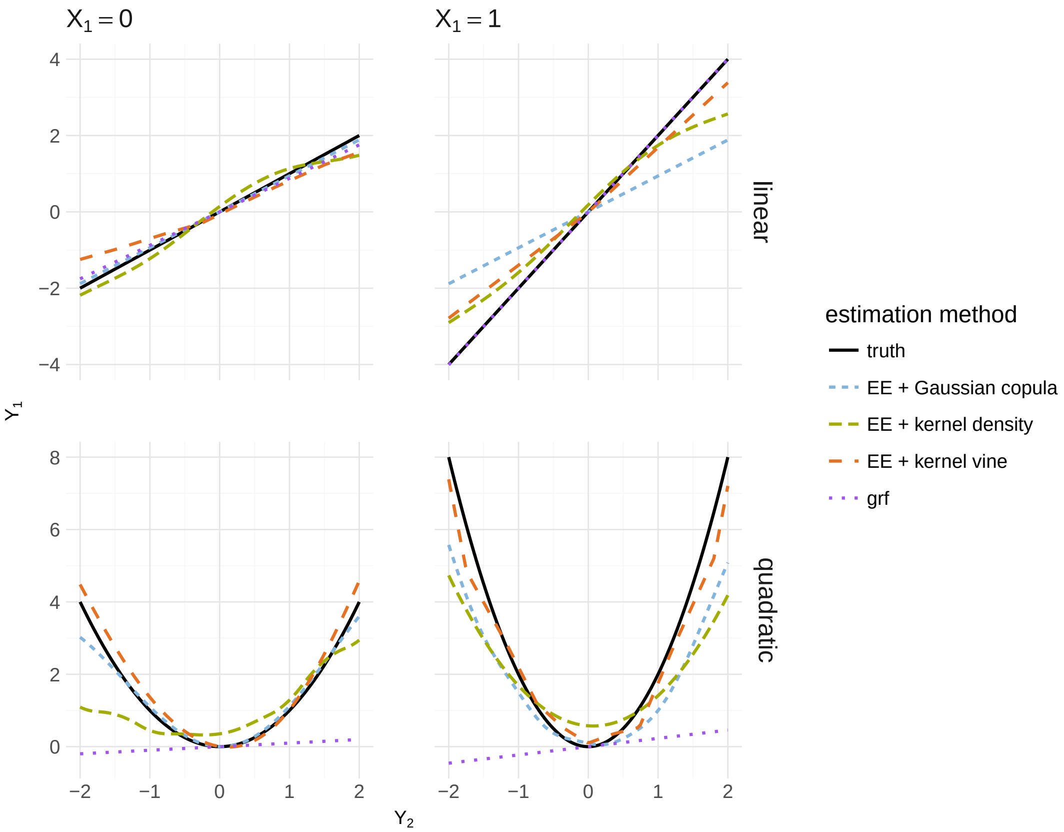


Figure 5.3: Instrumental variable regression with 10'000 observations and 10 covariates: true regression function (solid line) and estimates from various methods. Columns correspond to different values of the covariate X_1 and rows to different simulation models.

5.5 Discussion

5.5.1 Connection to previous results

The results in Section 5.3 provide an umbrella theory for solutions to copula-based estimating equations. They have close connections to several recent results. Noh et al. (2013) shows consistency and asymptotic normality for mean regression. Noh et al. (2015) and De Backer et al. (2017) provide similar results for quantile regression, the latter using a semiparametric method for estimating the copula density. Rémillard et al. (2017) establish weak convergence of the conditional quantile estimate as a process of the quantile level and prove validity of a parametric bootstrap procedure.

Our results generalize and extend the above in several ways:

- While the focus of previous research was on mean and quantile regression with univariate response, we allow for large classes of (potentially multivariate) identifying functions $\psi_{\theta,t}$. This opens new possibilities for copula-based solutions to other regression problems (see Section 5.2.3).

- Most previous results study only parametric or semiparametric estimators of $c_{\boldsymbol{Y},\boldsymbol{X}}$ (and thereby $w_{\boldsymbol{x}}$). Our conditions allow for parametric, semiparametric, and fully nonparametric methods to be dealt with in the same framework.

- We establish weak convergence of $\widehat{\theta}_t$ as a process of the parameter $t \in T$ indexing the identifying functions. This can be used to derive the asymptotic distribution of continuous functionals of $\theta_{0,t}$ or a vector of solutions that correspond to different identifying functions.

- We suggest and validate a multiplier bootstrap scheme for cases where exact calculation of the asymptotic distribution is inconvenient or insufficient.

Being more general, our results are also less explicit. However, it is usually straightforward to bring them in a more explicit form given a specific choice for the class of identifying functions Ψ and the estimator $\widehat{w}_{\boldsymbol{x}}$.

The above references also contain results that go beyond the context of Section 5.3. Noh et al. (2015) relax the *iid*-assumption and De Backer et al. (2017) allow for potentially right-censored Y. We discuss these and other possible extensions in the following.

5.5.2 Extensions

Stationary data

Our results in Section 5.3 rely heavily on empirical process theory for *iid* data. But the literature on empirical processes under serial dependence is much less developed. Nonetheless, a few classical results (e.g. Andrews, 1991) and more recent developments (Dehling et al., 2002) give hope that our results can be extended to stationary sequences under more stringent conditions.

Censored and missing response

Our approach can be extended to allow for censoring or missingness in the response. Such effects can be accounted for by adding a second weight function in the estimating equation. For example, suppose we only observe a right-censored version $Y^c = \min(Y, Z) \in \mathbb{R}$ of Y along with a censoring indicator $\Delta = \mathbb{1}(Y \leq Z)$.

Denoting $\bar{F}_{Z|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \Pr(Z \geq y \mid \boldsymbol{X} = \boldsymbol{x})$, we obtain

$$\begin{aligned}
&\mathrm{E}\bigl\{\psi_\theta(Y)\Delta/\bar{F}_{Z|\boldsymbol{X}}(Y \mid \boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}\bigr\} \\
&= \mathrm{E}\Bigl[\mathrm{E}\bigl\{\psi_\theta(Y)\Delta/\bar{F}_{Z|\boldsymbol{X}}(Y \mid \boldsymbol{x}) \mid Y, \boldsymbol{X} = \boldsymbol{x}\bigr\}\Bigr] \\
&= \mathrm{E}\Bigl[\mathrm{E}\{\Delta \mid Y, \boldsymbol{X} = \boldsymbol{x}\}\psi_\theta(Y)/\bar{F}_{Z|\boldsymbol{X}}(Y \mid \boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}\Bigr] \\
&= \mathrm{E}\bigl\{\bar{F}_{Z|\boldsymbol{X}}(Y \mid \boldsymbol{x})\psi_\theta(Y)/\bar{F}_{Z|\boldsymbol{X}}(Y \mid \boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}\bigr\} \\
&= \mathrm{E}\bigl\{\psi_\theta(Y) \mid \boldsymbol{X} = \boldsymbol{x}\bigr\}.
\end{aligned}$$

From there we can follow the steps in Section 5.2.1, to show that

$$\mathrm{E}\bigl\{\psi_\theta(Y) \mid \boldsymbol{X} = \boldsymbol{x}\bigr\} = 0 \quad \Leftrightarrow \quad E\bigl\{\psi_\theta(Y^c)w^c_{\boldsymbol{x}}(Y^c)\,\zeta_{\boldsymbol{x}}\bigr\} = 0,$$

where $\zeta_{\boldsymbol{x}} = \Delta/\bar{F}_{Z|\boldsymbol{X}}(Y \mid \boldsymbol{x})\}$ and $w^c_{\boldsymbol{x}}$ is defined similar to $w_{\boldsymbol{x}}$ in (5.4), but with Y replaced by Y^c.

This technique was used in De Backer et al. (2017) to allow for right censoring in copula-based quantile regression. It is only one instance of a larger class of methods based on *inverse probability weighting.* Other weight functionals ζ_x can be used similarly to account for other forms of censoring and missingness (see, e.g. Robins et al., 1994, Wooldridge, 2007, Han et al., 2016).

Discrete and mixed data

In applications, one often encounters discrete variables. Classical examples are classification problems (where Y is a class indicator) or count data. However, the ideas in this article are developed under the assumption that $(\boldsymbol{Y}, \boldsymbol{X})$ is absolutely continuous. Several subtle issues arise when discrete data are modeled with copulas (see, Genest and Neslehova, 2007), but our assumption is mainly for convenience.

Indeed, the ideas from Section 5.2 can be extended to discrete and mixed data. To give a simple example, suppose that Y is a Bernoulli variable and X is continuous and univariate. Sklar's theorem guarantees that there is a function $C_{Y,X}$ satisfying $F_{Y,X}(y, x) = C_{Y,X}\{F_Y(y), F_X(x)\}$. Denoting $C^{(2)}_{Y,X} = \partial C_{Y,X}(u, v)/\partial v$, one can verify that (5.2) is satisfied with

$$w_{\boldsymbol{x}}(y) = C^{(2)}_{Y,X}\bigl\{F_Y(y), F_X(x)\bigr\} - C^{(2)}_{Y,X}\bigl\{F_Y(y-1), F_X(x)\bigr\}.$$

Similar expressions exist in more than two dimensions with the simple difference above generalizing to a volume in a multi-dimensional space (see, Panagiotelis et al., 2012).

A simpler but slightly awkward solution for nonparametric estimation methods is *jittering.* By adding noise to discrete variables, one can transform the general regression problem into a purely continuous one. Nagler (2018b) shows that, under a suitable choice of noise density, the two problems become equivalent:

asymptotic properties carry over from the purely continuous to the mixed data setting. In particular, consistent estimators for the latter are automatically consistent for the former.

5.6 Proofs

Most arguments below are framed in the language of empirical processes, see Section 2.4 for an introduction.

5.6.1 Proof of Theorem 5.1

Denote $\Theta_\delta = \{\theta \colon \sup_{t\in T} |\theta_t - \theta_{0,t}| < \delta\}$ for some $0 < \delta < \infty$. For any $(\theta, \boldsymbol{y}) \in \Theta_\delta \times \mathcal{Y}$, let

$$\widehat{f}_\theta(\boldsymbol{y}) = \sup_{t\in T} |g_{\theta,t,\widehat{w}_{\boldsymbol{x}}}(\boldsymbol{y}) - g_{\theta,t,w_{\boldsymbol{x}}}(\boldsymbol{y})| \le \bar{\psi}_\delta(\boldsymbol{y})|\widehat{w}_{\boldsymbol{x}}(\boldsymbol{y}) - w_{\boldsymbol{x}}(\boldsymbol{y})|,$$

where $\bar{\psi}_\delta(\boldsymbol{y})$ is an envelope of the set $\{\psi_{\theta,t} \colon \theta \in \Theta_\delta, t \in T\}$. By (A1), we have that for every $\boldsymbol{y} \in \mathcal{Y}$, $\sup_{\theta\in\Theta_\delta} \widehat{f}_\theta(\boldsymbol{y}) \to_P 0$, which implies $\widehat{f}_{\theta_n}(\boldsymbol{y}) \to_P 0$, for any random sequence $\theta_n \in \Theta_\delta$. Furthermore, if G is an integrable envelope for $\mathcal{G}_\delta$, it holds $\widehat{f}_\theta(\boldsymbol{y}) \le 2G(\boldsymbol{y}) < \infty$. As $G(\boldsymbol{y})$ does not depend on θ, $\widehat{f}_{\theta_n}(\boldsymbol{y}) \to_P 0$ and $\widehat{f}_{\theta_n}(\boldsymbol{y}) \le 2G(\boldsymbol{y})$ together imply that $\widehat{f}_{\theta_n}(\boldsymbol{y})$ is uniformly integrable. Therefore, it converges to zero in mean, namely $\big(P_{\{\mathbf{Y}_i,\boldsymbol{X}_i\}_{i=1}^n} \widehat{f}_{\theta_n}\big)(\boldsymbol{y}) \to 0$, where $P_{\{\mathbf{Y}_i,\boldsymbol{X}_i\}_{i=1}^n}$ denote the expectation with respect to $\{\mathbf{Y}_i, \boldsymbol{X}_i\}_{i=1}^n$. By Fubini's theorem and with P the expectation with respect to $\boldsymbol{y}$, we have that $P_{\{\mathbf{Y}_i,\boldsymbol{X}_i\}_{i=1}^n}(P\widehat{f}_{\theta_n}) = P(P_{\{\mathbf{Y}_i,\boldsymbol{X}_i\}_{i=1}^n}\widehat{f}_{\theta_n})$, which implies that $P_{\{\mathbf{Y}_i,\boldsymbol{X}_i\}_{i=1}^n}(P\widehat{f}_{\theta_n}) \to 0$ by the dominated convergence theorem. And since $P\widehat{f}_{\theta_n}$ is non-negative by definition, this also implies $P\widehat{f}_{\theta_n} \to_P 0$. Using $\widehat{\theta}$ instead of the arbitrary θ_n, we have $P\widehat{f}_{\widehat{\theta}} \to_P 0$ and hence

$$\sup_{t\in T} |\mathbb{P}_n\, g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}} - Pg_{\widehat{\theta}_t,t,w_{\boldsymbol{x}}}| \le \sup_{t\in T} |(\mathbb{P}_n - P)g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}}| + P\widehat{f}_{\widehat{\theta}} \to_P 0, \tag{5.8}$$

where the first term goes to zero in probability since, with probability going to one, $g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}}$ is contained in a Glivenko-Cantelli class by (A2). Since $\mathbb{P}_n\, g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}} = 0$ by definition of $\widehat{\theta}_t$, (5.8) implies $\sup_{t\in T} |Pg_{\widehat{\theta}_t,t,w_{\boldsymbol{x}}}| \to_P 0$, which, along with (A3), finally yields $\sup_{t\in T} |\widehat{\theta}_t - \theta_{0,t}| \to_P 0$. □

5.6.2 Proof of Theorem 5.2

To improve readability, we make use of the following two lemmas. Their proofs are deferred to later sections.

Lemma 5.4. *Under (A1)–(A5), it holds uniformly in $t \in T$,*

$$\sqrt{n}(\widehat{\theta}_t - \theta_{0,t}) = -V_{\theta_{0,t},t,w_x}^{-1}\big(\mathbb{G}_n\, g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \sqrt{n}Pg_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}}\big) + o_P\big(1 + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\big).$$

Lemma 5.5. *Suppose that (A2) and (A6)–(A8) hold. Define* $h_n = r_n^{-1} g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \omega_{n,t,\boldsymbol{x}}$*,* $\mathcal{G}_n = r_n^{-1}\mathcal{G}_\delta$*,* $\mathcal{O}_n = \{\omega_{n,t,\boldsymbol{x}} \colon t \in T\}$*,* $\mathcal{H}_n = \mathcal{G}_n + \mathcal{O}_n$*, and the semi-metric* $\rho(h_1, h_2) = \|h_1 - h_2\|_{P,2} = \{P(h_1 - Ph_1 - h_2 + Ph_2)^2\}^{1/2}$*. Then* $\mathcal{H}_n$ *is totally bounded by* ρ *and for all* $\epsilon > 0$*,*

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \mathrm{P}\left\{ \sup_{h_1,h_2 \in \mathcal{H}_n, \rho(h_1,h_2)<\delta} \left|\mathbb{G}_n\left(h_1 - h_2\right)\right| > \epsilon \right\} = 0.$$

First note that

$$\begin{aligned} Pg_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}} &= \mathrm{E}_{\boldsymbol{Y}}\{\psi_{\theta_{0,t},t}(\boldsymbol{Y})\widehat{w}_{\boldsymbol{x}}(\boldsymbol{Y})\} \\ &= \frac{1}{n}\mathrm{E}_{\boldsymbol{Y}}\left\{\psi_{\theta_{0,t},t}(\boldsymbol{Y})\sum_{i=1}^n a_{n,\boldsymbol{x}}(\boldsymbol{Y}, \boldsymbol{Y}_i, \boldsymbol{X}_i)\right\} + o_P(n^{-1/2}r_n) \\ &= \frac{r_n}{n}\sum_{i=1}^n \omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y}_i, \boldsymbol{X}_i) + o_P(n^{-1/2}r_n) \\ &= r_n\mathbb{P}_n\,\omega_{n,t,\boldsymbol{x}} + o_P(n^{-1/2}r_n). \end{aligned} \tag{5.9}$$

Using this together with Lemma 5.4 yields

$$\begin{aligned} &\sqrt{n}(\widehat{\theta}_t - \theta_{0,t} - \beta_n) \\ &= \sqrt{n}(\widehat{\theta}_t - \theta_{0,t} - r_n V^{-1}_{\theta_{0,t},t,w_x} P\omega_{n,t,\boldsymbol{x}}) \\ &= -V^{-1}_{\theta_{0,t},t,w_x}\left(\mathbb{G}_n\, g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \sqrt{n}Pg_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}} - r_n\sqrt{n}P\omega_{n,t,\boldsymbol{x}}\right) + o_P\left(1 + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\right) \\ &= -V^{-1}_{\theta_{0,t},t,w_x}\left(\mathbb{G}_n\, g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + r_n\sqrt{n}\mathbb{P}_n\,\omega_{n,t,\boldsymbol{x}} - r_n\sqrt{n}P\omega_{n,t,\boldsymbol{x}}\right) + o_P\left(r_n + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\right) \\ &= -V^{-1}_{\theta_{0,t},t,w_x}\mathbb{G}_n\left(g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + r_n\omega_{n,t,\boldsymbol{x}}\right) + o_P\left(r_n + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\right). \end{aligned}$$

We now need to show that $\mathbb{G}_n\left(r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \omega_{n,t,\boldsymbol{x}}\right) \rightsquigarrow \mathbb{G}_t$. If this is the case, then $\sqrt{n}(\widehat{\theta}_t - \theta_{0,t}) = o_P(r_n)$ and the o_P term for the remainder becomes asymptotically negligible. Convergence of finite-dimensional distributions is an immediate consequence of conditions (A2i) and (A8) and the Lindeberg-Feller central limit theorem (van der Vaart, 1998, Theorem 2.27). Asymptotic equicontinuity is established in Lemma 5.5, which concludes the proof. □

5.6.3 Proof of Theorem 5.3

We start with the following lemma, whose proof can be found at the end of this section.

Lemma 5.6. *Suppose (A1)–(A5), (A7) and (A9) hold and define* $\tilde{\mathbb{P}}_n = n^{-1}\sum_{i=1}^n(\xi_i/\bar{\xi})\delta_{\boldsymbol{Y}_i,\boldsymbol{X}_i}$ *and* $\tilde{\mathbb{G}}_n = \sqrt{n}(\tilde{\mathbb{P}}_n - \mathbb{P}_n)$. *Then,*

$$\begin{aligned}&\sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t)\\ &= -V^{-1}_{\theta_{0,t},t,w_{\boldsymbol{x}}}\tilde{\mathbb{G}}_n\left(g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + r_n\omega_{n,t,\boldsymbol{x}}\right) + o_P\{1 + \sqrt{n}|\widetilde{\theta}_t - \widehat{\theta}_t| + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\}.\end{aligned}$$

Now we need to show that $(\mu/\sigma)\tilde{\mathbb{G}}_n\left(r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \omega_{n,t,\boldsymbol{x}}\right) \rightsquigarrow_{\xi,P} \mathbb{G}_t$. We shall first show that it converges weakly. Using $\mu/\bar{\xi} \to_P 1$ and the notation

$$\widetilde{h}_n(\xi, y, x) = (\xi - \mu)\{r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \omega_{n,t,\boldsymbol{x}}\}/\sigma$$

we obtain

$$\begin{aligned}&(\mu/\sigma)\tilde{\mathbb{G}}_n\left(r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \omega_{n,t,\boldsymbol{x}}\right)\\ &= \frac{\mu}{\sigma}\frac{1}{\sqrt{n}}\sum_{i=1}^n(\xi_i/\bar{\xi})\{r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}}(\boldsymbol{Y}_i) + \omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y}_i)\}\\ &\quad - \frac{\mu}{\sigma}\frac{1}{\sqrt{n}}\sum_{i=1}^n\{r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}}(\boldsymbol{Y}_i) + \omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y}_i)\}\\ &= \frac{1}{\sigma}\frac{1}{\sqrt{n}}\sum_{i=1}^n(\xi_i - \mu)\{r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}}(\boldsymbol{Y}_i) + \omega_{n,t,\boldsymbol{x}}(\boldsymbol{Y}_i)\}\{1 + o_P(1)\}\\ &= \sqrt{n}\mathbb{P}_n\widetilde{h}_n\{1 + o_P(1)\}\\ &= \mathbb{G}_n\widetilde{h}_n\{1 + o_P(1)\},\end{aligned}$$

where the last equality holds because ξ_i has mean μ and is independent of $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ (i.e., $P\widetilde{h}_n = 0$). Weak convergence of $\mathbb{G}_n\widetilde{h}_n$ can be established similarly to the convergence of $\mathbb{G}_n\left(r_n^{-1}g_{\theta_{0,t},t,w_{\boldsymbol{x}}} + \omega_{n,t,\boldsymbol{x}}\right)$ in the proof of Theorem 5.2, provided that additionally (A2), (A6) and (A8) hold with $g_{\theta_{0,t},t,w_{\boldsymbol{x}}}$ and $\omega_{n,t,\boldsymbol{x}}$ replaced by $\widetilde{g}_{\theta_{0,t},t,w_{\boldsymbol{x}}}$ and $\widetilde{\omega}_{n,t,\boldsymbol{x}}$. These conditions follow immediately from the fact that the sequence ξ_i is independent from the data and has finite variance. In particular, $\mathbb{G}_n\widetilde{h}_n$ converges weakly to $\mathbb{G}_t$ and, thus, is asymptotically tight. The remaining steps to show $\rightsquigarrow_{\xi,P}$-convergence are identical to the last paragraph of the proof of Theorem 2 in Kosorok (2003). □

5.6.4 Proof of Lemma 5.2

First, $\bar{w}$ being bounded along with (C2) imply that $\bar{w}\Psi_\delta$ is P-Donsker by van der Vaart and Wellner (1996), Example 2.10.10. Second, (C3) implies that $\bar{\psi}\mathcal{W}|_{\mathcal{Y}_j} \in C^\beta_{A\,M\,M_j}$ for some constant $A = A(\beta)$ and $\sum_{j=1}^\infty A\,MM_jP(\mathcal{Y}_j)^{1/2} < \infty$. Thus, $\bar{\psi}\mathcal{W}$ is P-Donsker by van der Vaart and Wellner (1996), Example 2.10.25. □

5.6.5 Proof of Lemma 5.3

This follows from the triangular inequality (first inequality), the mean value theorem (second inequality), and C4 (third inequality):

$$\begin{aligned}
&\sup_{\boldsymbol{y},\boldsymbol{y}'} \frac{|\partial_{\lfloor\beta\rfloor}\widehat{w}_{\boldsymbol{x}}(\boldsymbol{y}) - \partial_{\lfloor\beta\rfloor}\widehat{w}_{\boldsymbol{x}}(\boldsymbol{y}')|}{\|\boldsymbol{y}-\boldsymbol{y}'\|^{\beta-\lfloor\beta\rfloor}}\\
&\le \sup_{\boldsymbol{y},\boldsymbol{y}'}\left\{\frac{|\partial_{\lfloor\beta\rfloor}w_{\boldsymbol{x}}(\boldsymbol{y}) - \partial_{\lfloor\beta\rfloor}w_{\boldsymbol{x}}(\boldsymbol{y}')|}{\|\boldsymbol{y}-\boldsymbol{y}'\|^{\beta-\lfloor\beta\rfloor}} + \frac{|\partial_{\lfloor\beta\rfloor}(\widehat{w}_{\boldsymbol{x}}-w_{\boldsymbol{x}})(\boldsymbol{y}) - \partial_{\lfloor\beta\rfloor}(\widehat{w}_{\boldsymbol{x}}-w_{\boldsymbol{x}})(\boldsymbol{y}')|}{\|\boldsymbol{y}-\boldsymbol{y}'\|^{\beta-\lfloor\beta\rfloor}}\right\}\\
&\le M + \sup_{\boldsymbol{y}} |\partial_{\lfloor\beta\rfloor+1}(\widehat{w}_{\boldsymbol{x}}-w_{\boldsymbol{x}})(\boldsymbol{y})|\\
&\le M + o_P(1),
\end{aligned}$$

and similarly for lower order derivatives. □

5.6.6 Proof of Lemma 5.4

Because $(\widehat{\theta}_t, \widehat{w}_{\boldsymbol{x}})$ and $(\theta, w_{\boldsymbol{x}})$ are such that $\mathbb{P}_n\, g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}} = 0$, it holds

$$\begin{aligned}
-\mathbb{G}_n\, g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}} &= \sqrt{n}(P g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}} - \mathbb{P}_n\, g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}})\\
&= \sqrt{n} P g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}}\\
&= \sqrt{n} P (g_{\widehat{\theta}_t,t,\widehat{w}_{\boldsymbol{x}}} - g_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}}) + \sqrt{n} P g_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}}.
\end{aligned}$$

Linearizing the first term on the right-hand side yields

$$\sqrt{n}(\widehat{\theta}_t - \theta_{0,t}) = -V^{-1}_{\theta_{0,t},t,w_x}\big(\mathbb{G}_n\, g_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}} + \sqrt{n} P g_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}}\big) + o_P\big(\sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\big).$$

Then the consistency of $\widehat{w}_{\boldsymbol{x}}$, (A2), and (A4) imply

$$\mathbb{G}_n\,(g_{\theta_{0,t},t,\widehat{w}_{\boldsymbol{x}}} - g_{\theta_{0,t},t,w_{\boldsymbol{x}}}) \to_P 0,$$

and we obtain the desired result. □

5.6.7 Proof of Lemma 5.5

First note that $\mathcal{G}_n$, $\mathcal{O}_n$, and $\mathcal{H}_n$ are all totally bounded by ρ. By Lemma 7.20 in Kosorok (2007), it is enough to show that for every $\epsilon, \eta > 0$, there is a finite partition $\cup_{j=1}^m \mathcal{H}_{n,j}$ such that

$$\limsup_{n\to\infty} \mathrm{P}\left\{\sup_{1\le j\le m}\ \sup_{h_1,h_2\in\mathcal{H}_{n,j}} \big|\mathbb{G}_n\,(h_1-h_2)\big| > \epsilon\right\} < \eta. \tag{5.10}$$

For any fixed $\delta_1, \delta_2 > 0$, we can construct finite partitions $\mathcal{G}_n = \cup_{j=1}^m \mathcal{G}_{n,j}$ and $\mathcal{O}_n = \cup_{k=1}^m \mathcal{O}_{n,k}$ such that $\rho(g_1, g_2) < \delta_1$ and $\rho(\omega_1, \omega_2) < \delta_2$ for any two $g_1, g_2 \in \mathcal{G}_{n,j}$, $1 \le j \le m$, and similarly for $\mathcal{O}_n$. Then we define $\cup_{j,k=1}^m \mathcal{H}_{n,j,k} = \cup_{j,k=1}^m (\mathcal{G}_{n,j} + \mathcal{O}_{n,k})$, which is a finite partition of $\mathcal{H}_n$ such that

$$\begin{aligned}
&\mathrm{P}\bigg\{\sup_{1\le j,k\le m}\ \sup_{h_1,h_2\in\mathcal{H}_{n,j,k}}\big|\mathbb{G}_n\left(h_1-h_2\right)\big|>\epsilon\bigg\}\\
\le{}&\mathrm{P}\bigg\{\sup_{1\le j\le m}\ \sup_{g_1,g_2\in\mathcal{G}_{n,j}}\big|\mathbb{G}_n\left(g_1-g_2\right)\big|>\epsilon/2\bigg\}\\
&+\mathrm{P}\bigg\{\sup_{1\le k\le m}\ \sup_{\omega_1,\omega_2\in\mathcal{O}_{n,k}}\big|\mathbb{G}_n\left(\omega_1-\omega_2\right)\big|>\epsilon/2\bigg\}.
\end{aligned}$$

Since $\mathcal{G}_\delta$ is P-Donsker, (5.10) holds for $\mathcal{H}_{n,j}$ replaced with $\mathcal{G}_{n,j}$. Therefore, we can assume that δ_1 was such that for n sufficiently large,

$$\mathrm{P}\bigg\{\sup_{1\le j\le m}\ \sup_{g_1,g_2\in\mathcal{G}_{n,j}}\big|\mathbb{G}_n\left(g_1-g_2\right)\big|>\epsilon/2\bigg\}<\eta/2.$$

Further, Markov's inequality yields

$$\begin{aligned}
&\mathrm{P}\bigg\{\sup_{1\le k\le m}\ \sup_{\omega_1,\omega_2\in\mathcal{O}_{n,k}}\big|\mathbb{G}_n\left(\omega_1-\omega_2\right)\big|>\epsilon/2\bigg\}\\
\le{}&\frac{2}{\epsilon}\mathrm{E}\bigg\{\sup_{1\le k\le m}\ \sup_{\omega_1,\omega_2\in\mathcal{O}_{n,k}}\big|\mathbb{G}_n\left(\omega_1-\omega_2\right)\big|\bigg\},
\end{aligned}\tag{5.11}$$

and using Lemma 19.34 of van der Vaart (1998), there is $0<a(\delta_2)<\infty$ such that (5.11) is bounded from above by a universal constant times

$$\int_0^{\delta_2}\sqrt{\log N_{[\,]}\big\{\epsilon\|\Omega_n\|_{P,2},\mathcal{O}_n,L_2(P)\big\}}d\epsilon+\sqrt{n}\mathrm{E}\big[\Omega_n\mathbb{1}\{\Omega_n>a(\delta_2)\sqrt{n}\}\big].\tag{5.12}$$

For the second term in (5.12), we obtain the bound

$$\sqrt{n}\mathrm{E}\big[\Omega_n\mathbb{1}\{\Omega_n>a(\delta_2)\sqrt{n}\}\big]\le\mathrm{E}\big[\Omega_n^2\mathbb{1}\{\Omega_n>a(\delta_2)\sqrt{n}\}\big]/a(\delta_2),$$

which, by (A8), becomes arbitrarily small for any δ_2 and sufficiently large n. For the first term in (5.12), observe that

$$\begin{aligned}
&\big|\omega_{n,t_1,\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{x})-\omega_{n,t_2,\boldsymbol{x}}(y,\boldsymbol{x})\big|\\
\le{}&r_n^{-1}\int|\psi_{\theta_{0,t},t_1}(\boldsymbol{s})-\psi_{\theta_{0,t},t_2}(\boldsymbol{s})|\,|a_{n,\boldsymbol{x}}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{x})|f_{\boldsymbol{Y}}(\boldsymbol{s})d\boldsymbol{s}\\
\le{}&r_n^{-1}\sup_{\boldsymbol{s}}|a_{n,\boldsymbol{x}}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{x})|\int|\psi_{\theta_{0,t},t_1}(\boldsymbol{s})-\psi_{\theta_{0,t},t_2}(\boldsymbol{s})|f_{\boldsymbol{Y}}(\boldsymbol{s})d\boldsymbol{s}\\
={}&r_n^{-1}\sup_{\boldsymbol{s}}|a_{n,\boldsymbol{x}}(\boldsymbol{s},\boldsymbol{y},\boldsymbol{x})|\times\|\psi_{\theta_{0,t},t_1}-\psi_{\theta_{0,t},t_2}\|_{P,1}.
\end{aligned}$$

Hence, $\omega_{n,t,\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{x})$ is Lipschitz in $\psi_{\theta_{0,t},t}$ with respect to the $L_1(P)$ metric. Then

Theorem 2.7.11 of van der Vaart and Wellner (1996) yields

$$\int_0^{\delta_2} \sqrt{\log N_{[\,]}\{\epsilon \|\Omega_n\|_{P,2}, \mathcal{O}_n, L_2(P)\}} d\epsilon \leq \int_0^{\delta_2} \sqrt{\log N\{\epsilon/2, \Psi_{\theta_{0,t}}, L_1(P)\}} d\epsilon.$$

Now (A6) implies that we can choose δ_2 such that the left hand side is asymptotically less than $\eta/2$ and, thus, (5.10) holds. □

5.6.8 Proof of Lemma 5.6

First note that assumptions (A2)–(A5) continue to hold when $\widehat{w}_{\boldsymbol{x}}$ and $\psi_{\theta,t}(\boldsymbol{y})$ are replaced with $\widetilde{w}_{\boldsymbol{x}}$ and $\widetilde{\psi}_{\theta,t}(\xi, \boldsymbol{y}) = \xi\psi_{\theta,t}(\boldsymbol{y})$, respectively. Then $\sup_{t \in T} |\widetilde{\theta}_t - \theta_{0,t}| \to_P$ can be established using (A9) and the same arguments as in the proof of Theorem 5.1. Using a similar expansion as in the proof of Theorem 5.2, we get

$$\begin{aligned} &- \tilde{\mathbb{G}}_n\, g_{\widetilde{\theta}_t, \widetilde{w}_{\boldsymbol{x}}} \\ &= \mathbb{G}_n\, g_{\widetilde{\theta}_t, \widetilde{w}_{\boldsymbol{x}}} + \sqrt{n}(Pg_{\widetilde{\theta}_t, \widetilde{w}_{\boldsymbol{x}}} - Pg_{\widehat{\theta}_t, \widetilde{w}_{\boldsymbol{x}}}) + \sqrt{n}(Pg_{\widehat{\theta}_t, \widetilde{w}_{\boldsymbol{x}}} - Pg_{\theta_{0,t}, \widetilde{w}_{\boldsymbol{x}}}) + \sqrt{n}Pg_{\theta_{0,t}, \widetilde{w}_{\boldsymbol{x}}}. \end{aligned}$$

After linearizing the second and third term, we get

$$\begin{aligned} &- \tilde{\mathbb{G}}_n\, g_{\widetilde{\theta}_t, \widetilde{w}_{\boldsymbol{x}}} \\ &= \mathbb{G}_n\, g_{\widetilde{\theta}_t, t, \widetilde{w}_{\boldsymbol{x}}} + V_{\widehat{\theta}_t, t, \widetilde{w}_{\boldsymbol{x}}} \sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t) + V_{\theta_{0,t}, t, \widetilde{w}_{\boldsymbol{x}}} \sqrt{n}(\widehat{\theta}_t - \theta_{0,t}) + \sqrt{n}Pg_{\theta_{0,t}, t, \widetilde{w}_{\boldsymbol{x}}} \\ &\quad + o_P\big(1 + \sqrt{n}|\widetilde{\theta}_t - \widehat{\theta}_t| + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\big). \end{aligned}$$

Since by (A5) $V_{\theta,t,w}$ is continuous in θ and w, the consistency of $\widehat{\theta}_t$ and $\widetilde{w}_{\boldsymbol{x}}$ (Theorem 5.1 and A9) imply $V_{\widehat{\theta}_t, t, \widetilde{w}_{\boldsymbol{x}}} = V_{\theta_0, t, w_{\boldsymbol{x}}} + o_p(1)$. Then using Lemma 5.4 for $\sqrt{n}(\widehat{\theta}_t - \theta_{0,t})$ yields

$$\begin{aligned} &- \tilde{\mathbb{G}}_n\, g_{\widetilde{\theta}_t, \widetilde{w}_{\boldsymbol{x}}} \\ &= \mathbb{G}_n\, g_{\widetilde{\theta}_t, t, \widetilde{w}_{\boldsymbol{x}}} + V_{\theta_{0,t}, t, w_x} \sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t) - \mathbb{G}_n\, g_{\theta_{0,t}, t, w_{\boldsymbol{x}}} - \sqrt{n}Pg_{\theta_{0,t}, t, \widehat{w}_{\boldsymbol{x}}} + \sqrt{n}Pg_{\theta_{0,t}, t, \widetilde{w}_{\boldsymbol{x}}} \\ &\quad + o_P\big(1 + \sqrt{n}|\widetilde{\theta}_t - \widehat{\theta}_t| + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\big). \end{aligned}$$

Consistency of $(\widetilde{\theta}_t, \widetilde{w}_{\boldsymbol{x}})$, (A2), and (A4) imply $\tilde{\mathbb{G}}_n \left(g_{\widetilde{\theta}_t, t, \widetilde{w}_{\boldsymbol{x}}} - g_{\theta_{0,t}, t, w_{\boldsymbol{x}}}\right) = o_P(1)$. Using (5.9) in the second equality below yields

$$\begin{aligned} &- \tilde{\mathbb{G}}_n\, g_{\theta_{0,t}, w_{\boldsymbol{x}}} \\ &= V_{\theta_{0,t}, t, w_x} \sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t) + \sqrt{n}(Pg_{\theta_{0,t}, t, \widetilde{w}_{\boldsymbol{x}}} - Pg_{\theta_{0,t}, t, \widehat{w}_{\boldsymbol{x}}}) \\ &\quad + o_P\big(1 + \sqrt{n}|\widetilde{\theta}_t - \widehat{\theta}_t| + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\big) \\ &= V_{\theta_{0,t}, t, w_x} \sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t) + r_n \sqrt{n}(\tilde{\mathbb{P}}_n\, \omega_{n,t,\boldsymbol{x}} - \mathbb{P}_n\, \omega_{n,t,\boldsymbol{x}}) \\ &\quad + o_P\big(1 + \sqrt{n}|\widetilde{\theta}_t - \widehat{\theta}_t| + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\big) \\ &= V_{\theta_{0,t}, t, w_x} \sqrt{n}(\widetilde{\theta}_t - \widehat{\theta}_t) + r_n \tilde{\mathbb{G}}_n\, \omega_{n,t,\boldsymbol{x}} + o_P\big(1 + \sqrt{n}|\widetilde{\theta}_t - \widehat{\theta}_t| + \sqrt{n}|\widehat{\theta}_t - \theta_{0,t}|\big). \end{aligned}$$

Rearranging terms proves our claim. □

6 The jittering technique for nonparametric function estimation with mixed data

6.1 Introduction

In applications of statistics, data containing discrete variables are omnipresent. An online retailer records information on how many purchases a customer made in the past. Social scientists typically use discrete scales on which study participants rate their satisfaction, attitude, or feelings. Another common example is where data describe unordered categories, like gender or business sectors.

Suppose that $(\boldsymbol{Z}, \boldsymbol{X})$ is a random vector with discrete component $\boldsymbol{Z} \in \mathbb{Z}^p$ and continuous component $\boldsymbol{X} \in \mathbb{R}^q$. This includes the cases $p \geq 1$, $q = 0$ (all variables are discrete) and $p = 0$, $q \geq 1$ (all variables are continuous). We consider problems where one aims at estimating a functional T of the density/probability mass function $f_{\boldsymbol{Z},\boldsymbol{X}}$ based on observations $(\boldsymbol{Z}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. This formulation is general enough to include many common problems in nonparametric function estimation, in particular: density estimation, regression, and classification.

Some nonparametric estimation techniques have been specifically designed to allow for mixed continuous and discrete data (Ahmad and Cerrito, 1994, Li and Racine, 2003, Hall et al., 1983, Efromovich, 2011), but the number is small and the more sophisticated methods are often developed in a purely continuous framework. Examples are local polynomial methods (Fan and Gijbels, 1996, Loader, 1999) or copula-based estimators (e.g., Otneim and Tjøstheim, 2017, Nagler and Czado, 2016, Kauermann and Schellhase, 2014). These methods are no longer consistent when applied to mixed data types.

There is a popular trick among practitioners to get an approximate answer nevertheless: just make the data continuous by adding noise to each discrete variable. This trick is sometimes called *jittering* or *adding jitter*. Examples where it has been successfully applied are: avoiding overplotting in data visualization (Few, 2008), adding intentional bias to complex machine learning models (Zur et al., 2004), deriving theoretical properties of concordance measures (Denuit and Lambert, 2005), or nonparametric copula estimation for mixed data (Genest et al., 2017). An example of its misuse was pointed out by Nikoloulopoulos (2013) in the context of parametric copula models. Generally, the trick lacks theoretical justification because it can introduce bias. But we shall see that this issue is resolved under a suitable choice of noise distribution.

This chapter aims to formalize this trick and to provide a starting point for a more nuanced investigation of its properties. Some open questions and partial answers will be given at the end.

6.2 Jittering mixed data

6.2.1 Preliminaries and notation

We assume throughout that all random variables live in a space with a natural concept of ordering. Unordered categorical variables can always be coded into a set of binary dummy variables (for which $0 < 1$ gives a natural ordering). We further assume throughout that any discrete random variable, say Z, is supported on a set $\Omega_Z \subseteq \mathbb{Z}$. This is without loss of generality: if Ω_Z is an arbitrary (ordered) countable set, we can identify its elements with corresponding elements in $\mathbb{Z}$ in a way that the ordering is preserved. For any continuous random vector $\boldsymbol{X}$, we write $f_{\boldsymbol{X}}$ for its joint density. In case $\boldsymbol{Z}$ is a discrete random vector, $f_{\boldsymbol{Z}}$ denotes its density with respect to the counting measure, i.e., $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \Pr(\boldsymbol{Z} = \boldsymbol{z})$. A random vector with mixed types will be partitioned into $(\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$. Then $f_{\boldsymbol{Z},\boldsymbol{X}}$ is the density with respect to the product of the counting and Lebesgue measures,

$$f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) = \frac{\partial^q}{\partial x_1 \cdots \partial x_q} \Pr\big(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} \leq \boldsymbol{x}\big).$$

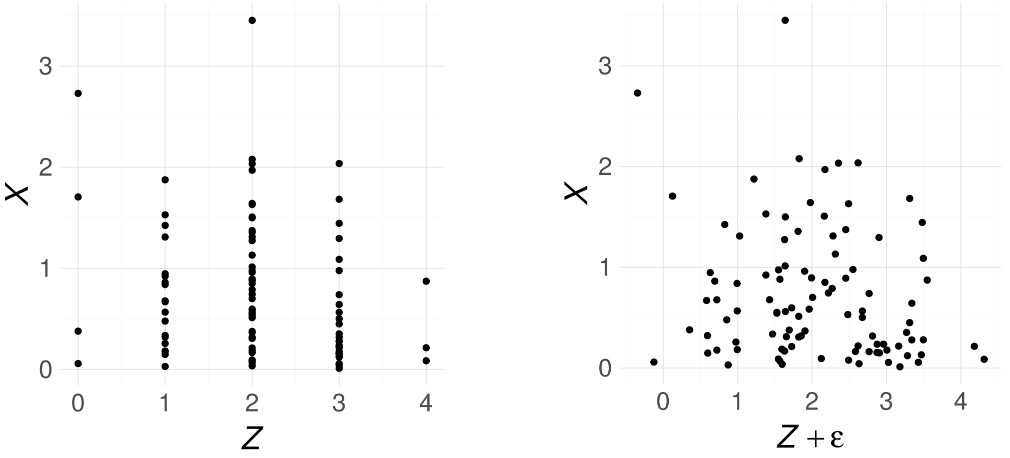
6.2.2 Jittering random vectors

The jittered version of a random vector is defined by adding noise to all discrete variables.

Definition 6.1. *Let $f_{\boldsymbol{\epsilon}}$ be a bounded density function that is continuous on $\mathbb{Z}^p$ and almost everywhere on $\mathbb{R}^p$. The jittered version of the random vector $(\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$ is defined as $(\boldsymbol{Z} + \boldsymbol{\epsilon}, \boldsymbol{X})$, where $\boldsymbol{\epsilon} \in \mathbb{R}^p$ has density $f_{\boldsymbol{\epsilon}}$ and is independent of $(\boldsymbol{Z}, \boldsymbol{X})$.*

The concept is illustrated in Figure 6.1 for a bivariate vector (Z, X) with mixed types. Figure 6.1a shows random samples from this vector. We see that X is continuously distributed on $\mathbb{R}_+$, and Z is discretely distributed on the set $\{0, 1, 2, 3, 4\}$. Figure 6.1b shows the jittered version of the same observations using Uniform$(-0.5, 0.5)$ noise. While all X values remain unchanged, the Z values are randomly shifted to left and right.

Remark 6.1. *Jittering is only meaningful when $p \geq 1$, i.e., at least one of the variables is discrete. The jittered version of a continuous vector $\boldsymbol{X}$ is the vector $\boldsymbol{X}$ itself.*



(a) original observations (b) jittered observations

Figure 6.1: Jittering observations with mixed types. The discrete component Z is jittered with Uniform$(-0.5, 0.5)$ noise.

Remark 6.2. *The noise density $f_{\boldsymbol{\epsilon}}$ may exhibit jumps at a countable number of points. An example of such a density is the uniform density on* $(-0.5, 0.5)$*, which jumps at* -0.5 *and* 0.5*.*

6.2.3 The density of a jittered random vector

Provided that $f_{\boldsymbol{Z},\boldsymbol{X}}$ exists, the density of the jittered vector $(\boldsymbol{Z} + \boldsymbol{\epsilon}, \boldsymbol{X})$ is simply the discrete-continuous convolution of $f_{\boldsymbol{Z},\boldsymbol{X}}$ and the noise density $f_{\boldsymbol{\epsilon}}$:

$$f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) = \sum_{\boldsymbol{z}' \in \mathbb{Z}^p} f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}', \boldsymbol{x}) f_{\boldsymbol{\epsilon}}(\boldsymbol{z} - \boldsymbol{z}'), \tag{6.1}$$

for almost all $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{R}^{p+q}$. To see this, write

$$\begin{aligned}\Pr(\boldsymbol{Z} + \boldsymbol{\epsilon} \leq \boldsymbol{z}, \boldsymbol{X} \leq \boldsymbol{x}) &= \sum_{\boldsymbol{z}' \in \mathbb{Z}^p} \Pr(\boldsymbol{Z} = \boldsymbol{z}', \boldsymbol{X} \leq \boldsymbol{x}, \boldsymbol{\epsilon} \leq \boldsymbol{z} - \boldsymbol{z}') \\ &= \sum_{\boldsymbol{z}' \in \mathbb{Z}^p} \Pr(\boldsymbol{Z} = \boldsymbol{z}', \boldsymbol{X} \leq \boldsymbol{x}) \Pr(\boldsymbol{\epsilon} \leq \boldsymbol{z} - \boldsymbol{z}'),\end{aligned}$$

and take the derivative with respect to $(\boldsymbol{z}, \boldsymbol{x})$.

We observe a close relationship between the densities $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ and $f_{\boldsymbol{Z},\boldsymbol{X}}$. If we know $f_{\boldsymbol{Z},\boldsymbol{X}}$ at all values $(\boldsymbol{z}', \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$, we can immediately compute $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ at all values $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{R}^{p \times q}$. The other direction is more interesting for our purposes: can we recover $f_{\boldsymbol{Z},\boldsymbol{X}}$ from known values of $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$? In general, this poses a rather challenging deconvolution problem. But we can make things easier by a suitable choice of noise density η. In fact, there is a large class of noise densities densities for which no deconvolution is necessary and $f_{\boldsymbol{Z},\boldsymbol{X}}$ and $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ coincide on $\mathbb{Z}^p \times \mathbb{R}^q$.

Proposition 6.1. *It holds*

$$f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x}) = f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x}) \tag{6.2}$$

for any joint density $f_{\boldsymbol{Z},\boldsymbol{X}}$ and all $(\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$, if and only if the following two conditions are satisfied:

(i) $f_{\boldsymbol{\epsilon}}(\boldsymbol{0}) = 1$,

(ii) $f_{\boldsymbol{\epsilon}}(\boldsymbol{z}) = 0$ *for all* $\boldsymbol{z} \in (\mathbb{Z} \setminus \{0\})^q$.

Proof. It is obvious that conditions (i) and (ii) imply (6.2). For the reverse implication, fix $\boldsymbol{x} \in \mathbb{R}^q$. Then (6.1) and (6.2) imply that

$$f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x}) = \sum_{\boldsymbol{z}' \in \mathbb{Z}^p} f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}',\boldsymbol{x}) f_{\boldsymbol{\epsilon}}(\boldsymbol{z}-\boldsymbol{z}'), \quad \text{for all } \boldsymbol{z} \in \mathbb{Z}^p. \tag{6.3}$$

Assuming that $\boldsymbol{Z}$ is almost surely equal to a constant $\boldsymbol{z} \in \mathbb{Z}^p$, all summands in (6.3) except the one with $\boldsymbol{z}' = \boldsymbol{z}$ are zero. Then (6.3) yields

$$f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x}) = f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x}) f_{\boldsymbol{\epsilon}}(\boldsymbol{0}),$$

and, hence, condition (i) must hold. Under this condition, (6.2) becomes

$$0 = \sum_{\boldsymbol{z}' \in \mathbb{Z}^p \setminus \{\boldsymbol{z}\}} f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}',\boldsymbol{x}) f_{\boldsymbol{\epsilon}}(\boldsymbol{z}-\boldsymbol{z}'), \quad \text{for all } \boldsymbol{z} \in \mathbb{Z}^p. \tag{6.4}$$

Now suppose that $f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x})$ is a probability density that is strictly positive for all $\boldsymbol{z} \in \mathbb{Z}$. Since $f_{\boldsymbol{\epsilon}}$ is also a probability density, (6.4) can only hold if condition (ii) is satisfied. □

A simple, but powerful implication is: under the conditions of Proposition 6.1, we can estimate the discrete-continuous density $f_{\boldsymbol{Z},\boldsymbol{X}}$ by estimating the purely continuous density $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$. This finding will be generalized further in Section 6.3.

6.2.4 A convenient class of noise distributions

In the following we give a particularly convenient class of noise densities.

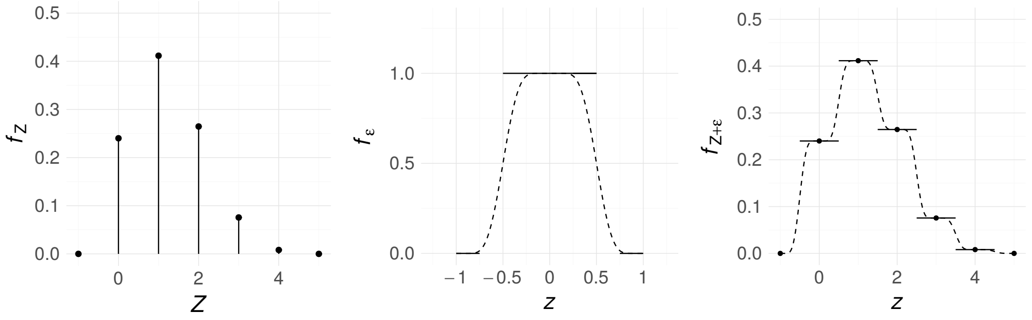
Definition 6.2. *We say that $f_{\boldsymbol{\epsilon}} \in \mathcal{E}_{\gamma_1,\gamma_2}$ for some $0 < \gamma_1 \leq 0.5 \leq \gamma_2 < 1$, if*

(i) $f_{\boldsymbol{\epsilon}}(\boldsymbol{x}) = \prod_{j=1}^p \eta(x_p)$ *for all* $\boldsymbol{x} \in \mathbb{R}^p$,

(ii) η *is a continuous probability density function,*

(iii) $\eta(x) = 1$ *for all* $x \in [-\gamma_1, \gamma_1]$,

(iv) $\eta(x) = 0$ *for all* $x \in \mathbb{R} \setminus (-\gamma_2, \gamma_2)$.



(a) original density (b) noise density (c) convoluted density

Figure 6.2: Jittering of a density function: (a) Binomial$(4, 0.3)$ density; (b) noise densities f_U (solid) and $f_{U_{\theta,\nu}}$ noise density (dashed), see Examples 6.1 and 6.2; (c) the convolution of the densities in (a) and (b).

Two elements of this class are illustrated in Figure 6.2b.

The class $\mathcal{E}_{\gamma_1,\gamma_2}$ satisfies (6.2), but adds two notable restrictions to the conditions given in Proposition 6.1: the random noise is componentwise independent, and it is constant in a neighborhood of zero. The first restriction is made purely for convenience and will be discussed further in Section 6.5.2. The second ensures that the derivatives of $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})$ with respect to $\boldsymbol{z}$ vanish for all $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$. This property is particularly useful in nonparametric density estimation, since an estimators' bias is usually proportional to derivatives of the target density.

Proposition 6.2. *If $f_{\boldsymbol{\epsilon}} \in \mathcal{E}_{\gamma_1,\gamma_2}$, $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$, and $\boldsymbol{m} \in \mathbb{N}^p$ such that $\sum_{k=1}^p m_k = m_+$, then*

$$\frac{\partial^{m_+} f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})}{\partial z_1^{m_1} \cdots \partial z_p^{m_p}} = 0.$$

Let us consider two examples for distribution classes contained in $\mathcal{E}_{\gamma_1,\gamma_2}$.

Example 6.1. *Let $f_U(x) = \mathbb{1}(|x| < 0.5)$ denote the uniform density on the set $(-0.5, 0.5)$. Then $f_U \in \mathcal{E}_{0.5,0.5}$. Furthermore, f_U is a piecewise constant function that jumps at $x = -0.5$ and $x = 0.5$. Figure 6.2 (solid lines) illustrates its use for jittering a discrete random variable $Z \sim \mathrm{Binomial}(4, 0.3)$. The density $f_{Z+\epsilon}$ of the jittered random variable $Z + \epsilon$ is shown in Figure 6.2c. It is a piecewise constant function and (because $f_U(0) = 1$) coincides with f_Z for all $z \in \mathbb{Z}$.* □

Example 6.2. *Let $\nu \in \mathbb{N}$ and $0 \leq \theta < 1$. Set $U_{\theta,\nu} = U + \theta(B_\nu - 0.5)$ where $U \sim$ Uniform$(-0.5, 0.5)$ and $B_\nu \sim \mathrm{Beta}(\nu, \nu)$ with corresponding distribution function F_{B_ν}. The density of $U_{\theta,\nu}$ can be calculated as*

$$f_{U_{\theta,\nu}}(x) = \begin{cases} \mathbb{1}(|x| < 0.5), & \theta = 0, \\ F_{B_\nu}\{(x + 0.5)/\theta + 0.5\} - F_{B_\nu}\{(x - 0.5)/\theta + 0.5\}, & \theta > 0. \end{cases}$$

The case where $\theta = 0$ is trivial. For $\theta > 0$, we first derive the density of the random variable $\theta(B_\nu - 0.5)$ as $f_{\theta(B_\nu - 0.5)}(x) = f_{B_\nu}(x/\theta + 0.5)/\theta$. The density of $U_{\theta,\nu}$ is then the convolution of the densities f_U and $f_{\theta(B_\nu-0.5)}$, i.e.,

$$f_{U_{\theta,\nu}}(x) = \int \mathbb{1}(-0.5 < s < 0.5) f_{B_\nu}((x-s)/\theta + 0.5)/\theta ds.$$

The change of variables $t = (x-s)/\theta + 0.5$ then yields

$$\begin{aligned} f_{U_{\theta,\nu}}(x) &= \int \mathbb{1}(-0.5 < x - \theta(t - 0.5) < 0.5) f_{B_\nu}(t)dt \\ &= \int_{(x-0.5)/\theta+0.5}^{(x+0.5)/\theta+0.5} f_{B_\nu}(t)dt \\ &= F_{B_\nu}\big\{(x+0.5)/\theta + 0.5\big\} - F_{B_\nu}\big\{(x-0.5)/\theta+0.5\big\}. \end{aligned}$$

Observe that if $|x| \leq (1-\theta)/2$, we get

$$\frac{x+0.5}{\theta} + 0.5 \geq \frac{-(1-\theta)/2 + 0.5}{\theta} + 0.5 = -\frac{1}{2\theta} + 0.5 + \frac{1}{2\theta} + 0.5 = 1,$$

and, hence, $F_{B_\nu}\big\{(x+0.5)/\theta + 0.5\big\} = 1$. Similarly,

$$\frac{x-0.5}{\theta} + 0.5 \leq \frac{(1-\theta)/2 - 0.5}{\theta} + 0.5 = \frac{1}{2\theta} - 0.5 - \frac{1}{2\theta} + 0.5 = 0,$$

and, hence, $F_{B_\nu}\big\{(x-0.5)/\theta + 0.5\big\} = 0$. Altogether this yields $f_{U_{\theta,\nu}}(x) = 1$ for $|x| \leq (1-\theta)/2$. Similar calculations show that $f_{U_{\theta,\nu}}(x) = 0$ if $|x| \geq (1+\theta)/2$ and, thus, $f_{U_{\theta,\nu}} \in \mathcal{E}_{(1-\theta)/2,(1+\theta)/2}$. Furthermore, $f_{U_{\theta,\nu}}$ is $\nu - 1$ times continuously differentiable everywhere on $\mathbb{R}$. Hence, if $f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x})$ is m times continuously differentiable in $\boldsymbol{x}$ for all $(\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$, $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ is $\min\{\nu-1, m\}$ times continuously differentiable everywhere on $\mathbb{R}^{p+q}$. Also, $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ coincides with $f_{\boldsymbol{Z},\boldsymbol{X}}$ everywhere on $\mathbb{Z}^p \times \mathbb{R}^q$. This is illustrated with dashed lines in Figure 6.2 for $Z \sim \mathrm{Binomial}(4, 0.3)$, $\nu = 5$, and $\theta = 0.3$. □

6.3 Nonparametric function estimation via jittering

6.3.1 Jittering estimators

Suppose we want to estimate a functional T of $f_{\boldsymbol{Z},\boldsymbol{X}}$, where $(\boldsymbol{Z},\boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$. Let $(\boldsymbol{Z}_1,\boldsymbol{X}_1)$, ..., $(\boldsymbol{Z}_n,\boldsymbol{X}_n)$ be a sequence of random vectors having the same distribution as $(\boldsymbol{Z},\boldsymbol{X})$. This sequence represents the observations that are used to estimate $T(f_{\boldsymbol{Z},\boldsymbol{X}})$ and is often assumed to be *iid*. However, independence is not required in what follows; we only assume that the sequence has a stationary distribution. Let further $\boldsymbol{\epsilon}_i$, $i = 1, \dots, n$, be independent and identically distributed vectors that have the same distribution as $\boldsymbol{\epsilon}$ (as in Definition 6.1) and are independent of $(\boldsymbol{Z}_1,\boldsymbol{X}_1), \dots, (\boldsymbol{Z}_n,\boldsymbol{X}_n)$.

Definition 6.3. *An estimator $\widetilde{\tau}_n$ of $T(f_{\boldsymbol{Z},\boldsymbol{X}})$ is called jittering estimator if it is a measurable function of the jittered data, i.e., $\widetilde{\tau}_n = \widetilde{\tau}_n(\boldsymbol{Z}_1 + \boldsymbol{\epsilon}_1, \boldsymbol{X}_1, \dots, \boldsymbol{Z}_n + \boldsymbol{\epsilon}_n, \boldsymbol{X}_n)$.*

Under the conditions of see Proposition 6.1, we have $f_{\boldsymbol{Z},\boldsymbol{X}} \equiv f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ on $\mathbb{Z}^p \times \mathbb{R}^q$ and, thus, $T(f_{\boldsymbol{Z},\boldsymbol{X}}) = T(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$. Now if $\widetilde{\tau}$ is an estimator of $T(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$, then it is also an estimator of $T(f_{\boldsymbol{Z},\boldsymbol{X}})$. This means that we can use any estimator that works in a purely continuous setting to estimate the target functional $T(f_{\boldsymbol{Z},\boldsymbol{X}})$, even though $f_{\boldsymbol{Z},\boldsymbol{X}}$ is the density of a mixed data model.

More generally, suppose that there is another functional T^* such that $T(f_{\boldsymbol{Z},\boldsymbol{X}}) = T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$. We shall call T^* a *jittering equivalent* of T. T is always a jittering equivalent of itself, but sometimes other functionals are more convenient to work with (as we will see shortly in examples). Again, if $\widetilde{\tau}_n$ is an estimator of $T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$, then it is also an estimator of $T(f_{\boldsymbol{Z},\boldsymbol{X}})$.

We already discussed the example of density estimation, where $T(f_{\boldsymbol{Z},\boldsymbol{X}}) = f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})$ for some $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$. But the setup is much more general and also covers regression problems, as we shall see in the following.

6.3.2 Examples: Estimating regression functions via jittering

Generally speaking, regression models describe a conditional relationship between a *response* variable and a vector of *covariates* (also called *predictors*). Let the response $Y \in \Omega_Y$ be either discrete or continuous and denote the vector of predictors as $(\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$. Since the joint density $f_{Y,\boldsymbol{Z},\boldsymbol{X}}$ fully characterizes the conditional relationship of Y and $(\boldsymbol{Z}, \boldsymbol{X})$, regression functions can be equivalently stated as a functional T of $f_{Y,\boldsymbol{Z},\boldsymbol{X}}$. Hence, regression functions involving discrete variables are amenable to estimation using the jittering technique.

The following examples illustrate the concept of jittering in the most common types of regression problems. In all examples, η is the noise term used to jitter Y in case Y is discrete and defined as in Definition 6.1. We assume that both f_η and $f_{\boldsymbol{\epsilon}}$ satisfy the conditions of Proposition 6.1.

Example 6.3 (Mean regression)**.** *Suppose we want to estimate the conditional mean* $\mathrm{E}(Y \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x})$*. We get:*

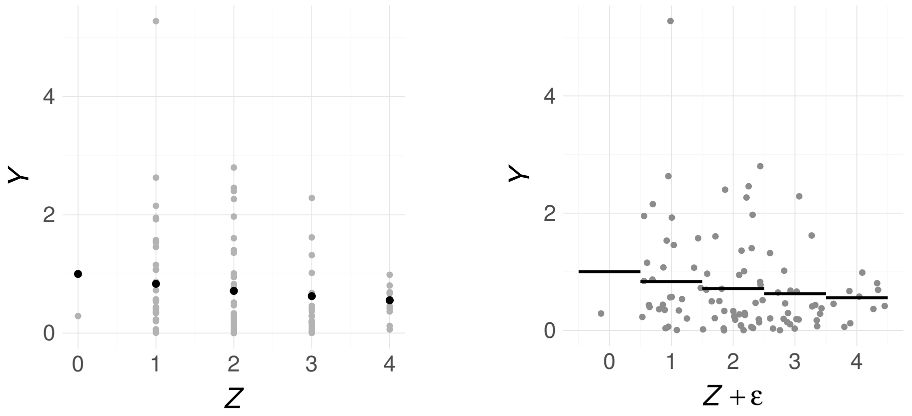
(i) continuous response:

$$T_{m,c}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \frac{\int_{\mathbb{R}} s f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x}) ds}{\int_{\mathbb{R}} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x}) ds}.$$

A jittering equivalent is $T^*_{m,c}(f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = T_{m,c}(f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$.

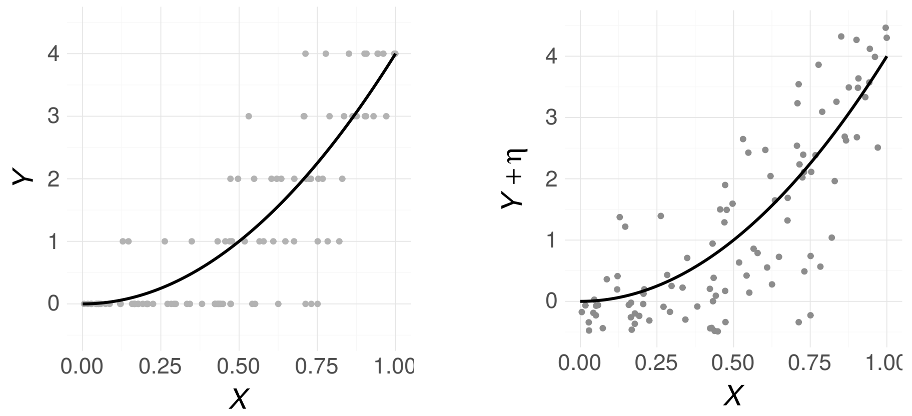
(ii) discrete response:

$$T_{m,d}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \frac{\sum_{s\in\mathbb{Z}} s f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x})}{\sum_{s\in\mathbb{Z}} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x})}.$$



(a) original (b) jittered

Figure 6.3: Mean regression: continuous response, discrete covariate. Black points and solid lines indicate the true regression function.



(a) original (b) jittered

Figure 6.4: Mean regression: discrete response, continuous covariate. Solid lines indicate the true regression function.

An example for a jittering equivalent is

$$T^*_{m,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = T_{m,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}),$$

which is motivated by the equality

$$f_{Y,\boldsymbol{Z},\boldsymbol{X}}(y,\boldsymbol{z},\boldsymbol{x}) = f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(y,\boldsymbol{z},\boldsymbol{x}), \quad \textit{for all } (y,\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^{p+1} \times \mathbb{R}^q.$$

Hence, this jittering equivalent simply acts as if $f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ were a discrete-continuous density. Sometimes this is computationally or conceptually inconvenient. In such cases we can use

$$T^*_{m,d}(f_{Y+\eta,\boldsymbol{Z}+\epsilon,\boldsymbol{X}}) = \frac{\int_{\mathbb{R}} s f_{Y+\eta,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}{\int_{\mathbb{R}} f_{Y+\eta,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds} = T_{m,c}(f_{Y+\eta,\boldsymbol{Z},\boldsymbol{X}}),$$

which is another jittering equivalent that treats $f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ as a continuous density.

The use of jittering in mean regression is illustrated in Figure 6.3 for the case of a continuous response Y and a discrete covariate X. Figure 6.3a shows simulated data from such a regression model in gray along with the true regression function in black. Note that the regression function is integer-valued because the covariate is. Figure 6.3b shows the jittered version (using Uniform$(-0.5, 0.5)$ noise) of the simulated data and the regression function in the jittered model as solid line. The regression function in the jittered model is a step function whose values equal to the original regression function for all $z \in \mathbb{Z}$. The situation is similar for the case of a discrete response and continuous covariate (Figure 6.4) with the difference that the true regression curve is smooth in both the original and jittered domain.

Example 6.4 (Distribution regression)**.** *Suppose we want to estimate the conditional distribution function* $\Pr(Y \le y \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x})$.

(i) continuous response:

$$T_{p,c}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \frac{\int_{-\infty}^{y} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}{\int_{\mathbb{R}} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}.$$

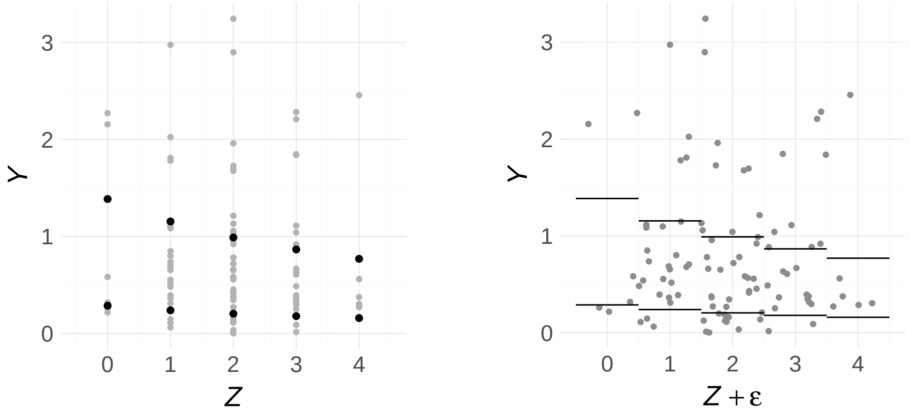
A jittering equivalent is $T^*_{p,c}(f_{Y,Z+\boldsymbol{\epsilon},\boldsymbol{X}}) = T_{p,c}(f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$.

(ii) discrete response:

$$T_{p,d}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \frac{\sum_{s=-\infty}^{y} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})}{\sum_{s\in\mathbb{Z}} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})}.$$

Two examples for jittering equivalents are

$$T^*_{p,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = T_{p,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$$



(a) original (b) jittered

Figure 6.5: Quantile regression ($\alpha = 0.25, 0.75$): continuous response, discrete covariate. Black points and solid lines indicate the true quartile functions.

and

$$\begin{aligned} &T^*_{p,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) \\ &= \frac{\int_{-\infty}^{y} f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds + f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(y,\boldsymbol{z},\boldsymbol{x}) \int_0^1 f_\eta(s)ds}{\int_{\mathbb{R}} f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds} \\ &= T_{p,c}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) + \frac{f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(y,\boldsymbol{z},\boldsymbol{x}) \int_0^1 f_\eta(s)ds}{\int_{\mathbb{R}} f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}. \end{aligned}$$

The second example again treats $f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ as a continuous density. Recall that jittering spreads the probability mass originally located on y onto the interval $(y-1, y+1)$. The mass in $(y-1, y]$ is collected by integrating $f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ up to the value y (first term above). The correction term

$$f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(y,\boldsymbol{z},\boldsymbol{x}) \int_0^1 f_\eta(s)ds,$$

then collects the remaining mass in the interval $(y, y+1)$.

Example 6.5 (Quantile regression)**.** *For $\alpha \in [0,1]$, the conditional quantile function corresponding to $\Pr(Y \leq y \mid \boldsymbol{Z} = z, \boldsymbol{X} = \boldsymbol{x})$ is defined as*

$$Q(\alpha \mid \boldsymbol{z}, \boldsymbol{x}) = \inf\bigl\{y \in \mathbb{R} \colon \Pr(Y \leq y \mid \boldsymbol{Z} = z, \boldsymbol{X} = \boldsymbol{x}) \geq \alpha\bigr\}.$$

Recall the definitions of $T_{p,c}, T_{p,d}, T^*_{p,c}$, and $T^*_{p,d}$ from Example 6.4.

(i) continuous response: We are interested in the functional

$$T_{q,c}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \inf\bigl\{y \in \mathbb{R} \colon T_{p,c}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) \geq \alpha\bigr\}.$$

An example for the jittering equivalent is

$$T^*_{q,c}(f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = \inf\big\{y \in \mathbb{R} \colon T^*_{p,c}(f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) \geq \alpha\big\}.$$

(ii) discrete response: We are interested in the functional

$$T_{q,d}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \inf\big\{y \in \mathbb{R} \colon T_{p,d}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) \geq \alpha\big\}.$$

An example for the jittering equivalent is

$$T^*_{q,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = \inf\big\{y \in \mathbb{R} \colon T^*_{p,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) \geq \alpha\big\}.$$

The use of jittering in quantile regression is illustrated in Figure 6.5 for the case of a continuous response Y and a discrete covariate X. Figure 6.5a shows simulated data from such a regression model in gray along with the true quartile functions (i.e., $\alpha = 0.25, 0.75$) in black. Figure 6.3b shows the jittered version (using Uniform$(-0.5, 0.5)$ noise) of the simulated data and the quartile functions in the jittered model as solid lines. We again observe that the conditional quartiles in the original and jittered models coincide for all $z \in \mathbb{Z}$.

Example 6.6 (Estimating equations)**.** *Suppose we are in the setting of Chapter 5 and want to solve*

$$\mathrm{E}\big\{\psi_\theta(Y) \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x})\big\} = 0,$$

where ψ_θ is an identifying function for the parameter of interest θ. Recall that this is equivalent to solving

$$\mathrm{E}\big\{\psi_\theta(Y) \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x})\big\} = 0,$$

(i) continuous response: The target functional is

$$T_{ee,c}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \frac{\int_{\mathbb{R}} \psi_\theta(s) f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x}) ds}{\int_{\mathbb{R}} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x}) ds}.$$

From Example 6.3, we know that

$$\mathrm{E}\{\psi_\theta(Y) \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}) = \mathrm{E}\{\psi_\theta(Y) \mid \boldsymbol{Z} + \boldsymbol{\epsilon} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}).$$

Hence, a jittering equivalent is given by

$$T^*_{ee,c}(f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = \frac{\int_{\mathbb{R}} \psi_\theta(s) f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x}) ds}{\int_{\mathbb{R}} f_{Y,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s, \boldsymbol{z}, \boldsymbol{x}) ds}.$$

(ii) discrete response: The target functional is

$$T_{ee,d}(f_{Y,\boldsymbol{Z},\boldsymbol{X}}) = \frac{\sum_{s\in\mathbb{Z}} \psi_\theta(s) f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})}{\sum_{s\in\mathbb{Z}} f_{Y,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})}.$$

The natural jittering equivalent is

$$T^*_{ee,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = \frac{\sum_{s\in\mathbb{Z}} \psi_\theta(s) f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})}{\sum_{s\in\mathbb{Z}} f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})},$$

but treats $f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ as a discrete density. This can be inconvenient, especially when taking a copula approach to solve the estimating equation (Chapter 5). As in the case of distribution regression (Example 6.4), there are small complications. One issue is that ψ_θ is only defined on $\mathbb{Z}$. This can be addressed by defining $\psi^\circ_\theta(y) = \psi_\theta([y])$ for all $y \in \mathbb{R}$, where $[s]$ denotes rounding to the closest integer. But even then, it is not generally true that

$$\mathrm{E}\{\psi_\theta(Y) \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}) = \mathrm{E}\{\psi^\circ_\theta(Y+\eta) \mid \boldsymbol{Z}+\boldsymbol{\epsilon} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}).$$

One solution is to set

$$\psi^*_\theta(y) = \frac{\psi_\theta([y])\mathbb{1}\left(\big|y-[y]\big| < 1-\gamma_2\right)}{\int_{-(1-\gamma_2)}^{1-\gamma_2} f_\eta(s)ds},$$

with γ_2 as in Proposition 6.1. Then, one can show that

$$\mathrm{E}\{\psi_\theta(Y) \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}) = \mathrm{E}\{\psi^*_\theta(Y+\eta) \mid \boldsymbol{Z}+\boldsymbol{\epsilon} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}), \tag{6.5}$$

for all $(\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$ and we obtain the alternative jittering equivalent

$$T^*_{ee,d}(f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}) = \frac{\int_{\mathbb{R}} \psi^*_\theta(s) f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}{\int_{\mathbb{R}} f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}.$$

To see that (6.5) holds, we start with

$$\begin{aligned}
&\mathrm{E}\{\psi^*_\theta(Y+\eta) \mid \boldsymbol{Z}+\boldsymbol{\epsilon} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}) \\
&= \frac{\int_{\mathbb{R}} \psi^*_\theta(s) f_{Y+\eta,\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}{f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x})} \\
&= \frac{\int_{\mathbb{R}} \psi^*_\theta(s) f_{Y+\eta,\boldsymbol{Z},\boldsymbol{X}}(s,\boldsymbol{z},\boldsymbol{x})ds}{f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x})} \\
&= \sum_{y'\in\mathbb{Z}} f_{Y|\boldsymbol{Z},\boldsymbol{X}}(y',\boldsymbol{z},\boldsymbol{x}) \int_{\mathbb{R}} \psi^*_\theta(s) f_\eta(y'-s)ds,
\end{aligned} \tag{6.6}$$

where the last equality is due to (6.1). Recalling that $f_\eta(y) = 0$ for $|y| \geq \gamma_2$, *we get*

$$\int_{\mathbb{R}} \psi_\theta^*(s) f_\eta(y' - s) ds = \int_{y'-\gamma_2}^{y'+\gamma_2} \psi_\theta^*(s) f_\eta(y' - s) ds.$$

Since $\gamma_2 \in [0.5, 1)$*, it holds for all* $s \in (y' - \gamma_2, y' + \gamma_2)$ *that*

$$\big|s - [s]\big| < 1 - \gamma_2 \quad \Leftrightarrow \quad [s] = y' \text{ and } |s - y'| < 1 - \gamma_2,$$

Therefore,

$$\begin{aligned}
\int_{\mathbb{R}} \psi_\theta^*(s) f_\eta(y' - s) ds &= \frac{\int_{y'-\gamma_2}^{y'+\gamma_2} \psi_\theta([s]) \mathbb{1}\big(|s - [s]| < 1 - \gamma_2\big) f_\eta(y' - s) ds}{\int_{-(1-\gamma_2)}^{1-\gamma_2} f_\eta(s) ds} \\
&= \frac{\psi_\theta(y') \int_{y'-(1-\gamma_2)}^{y'+(1-\gamma_2)} f_\eta(y' - s) ds}{\int_{-(1-\gamma_2)}^{1-\gamma_2} f_\eta(s) ds} \\
&= \psi_\theta(y'),
\end{aligned}$$

and, hence,

$$(6.6) = \sum_{y' \in \mathbb{Z}} f_{Y|\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(y', \boldsymbol{z}, \boldsymbol{x}) \psi_\theta(y') = \mathrm{E}\{\psi_\theta(Y) \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x}),$$

as claimed.

6.3.3 A note on asymptotic properties

A convenient fact about jittering estimators is that asymptotic properties for estimating $T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$ directly translate into properties for estimating $T(f_{\boldsymbol{Z},\boldsymbol{X}})$. Recall the definition of a jittering estimator τ (Definition 6.3). The following result is trivial, but important enough to be stated formally.

Proposition 6.3. *Let* T *and* T^* *be two functionals such that* $T(f_{\boldsymbol{Z},\boldsymbol{X}}) = T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$*. If for some sequence* $r_n \to 0$ *and random variable* W*,* $r_n^{-1}\{\widetilde{\tau} - T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})\} \to W$ *almost surely, in probability, or in distribution, then also* $r_n^{-1}\{\widetilde{\tau} - T(f_{\boldsymbol{Z},\boldsymbol{X}})\} \to W$ *almost surely, in probability, or in distribution.*

In particular, any (strongly) consistent estimator of $T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$ is at the same time a (strongly) consistent estimator of $T(f_{\boldsymbol{Z},\boldsymbol{X}})$. Even better: since we can choose the noise distribution η we gain some control over the local behavior of the jittered density $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$. If T^* is sufficiently well-behaved, this allows us to control the local behavior of the estimation target $T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$, too. For example, the form of the regression functionals in Section 6.3.2 and Proposition 6.2 imply that all derivatives of $T^*(f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}})$ with respect to $\boldsymbol{z}$ vanish in a γ_1-neighborhood of $\boldsymbol{z} \in \mathbb{Z}^p$. This allows to estimate regression functionals without bias for the discrete part and, thus, to improve the convergence rates of the estimator $\widetilde{\tau}$; see Chapter 7 for an in-depth analysis of the jittering kernel density estimator.

6.3.4 Examples of jittering estimators

Jittering estimators are extremely easy to implement: all one needs is a way to generate random noise and an estimator that works for continuous data. The following examples introduce jittering analogues of popular estimators that, in their original version, are only applicable to continuous data.

Example 6.7 (Kernel density estimator)**.** *Let K be a symmetric probability density function, and abbreviate the product kernel $K(\boldsymbol{w}) = \prod_{j=1}^{k} K(w_j)$ for any $\boldsymbol{w} \in \mathbb{R}^k$, $k \in \mathbb{N}$. The jittering kernel density estimator of $f_{\boldsymbol{Z},\boldsymbol{X}}$ is*

$$\widetilde{f}(\boldsymbol{z}, \boldsymbol{x}) = \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{z}}{h_n}\right) K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{b_n}\right).$$

where $h_n, b_n > 0$ are the bandwidths. The classical kernel density estimator of Parzen (1962) and Rosenblatt (1956) is recovered when $\boldsymbol{\epsilon}_i = 0$ for all $i = 1, \ldots, n$.

Example 6.8 (Local likelihood density estimator)**.** *Local polynomial likelihood density estimators (Loader, 1996) are a generalization of the kernel density estimator in the following sense. The idea is to locally fit a polynomial, say $g(\cdot; \boldsymbol{a})$, to the log-likelihood, where $\boldsymbol{a}$ is the vector of polynomial coefficients (see Section 3.2.1 for more details in the bivariate case). The local fit is obtained by minimizing the locally weighted likelihood function*

$$\begin{aligned}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{z}}{h_n}\right) K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{b_n}\right) g(\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \boldsymbol{X}_i; \boldsymbol{a}) \\ - \int K\left(\frac{\boldsymbol{s} - \boldsymbol{z}}{h_n}\right) K\left(\frac{\boldsymbol{t} - \boldsymbol{x}}{b_n}\right) \exp\{g(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{a})\} d\boldsymbol{s} d\boldsymbol{t}.\end{aligned}$$

Denoting $\widetilde{\boldsymbol{a}} = (\tilde{a}_1, \tilde{a}_2, \ldots)$ as the minimizer, the density $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})$ is estimated by $\exp(\widetilde{a}_1)$. For a local constant fit, we recover the classical kernel density estimator.

Example 6.9 (Orthogonal series density estimator)**.** *Another popular method for nonparametric density estimation is based on expansions in an orthonormal function basis. See Schwartz (1967), Watson (1969) for some early contributions and Efromovich (2010) for a recent overview. Assume that $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})$ is a square-integrable density with bounded, convex support $\mathcal{S} \subset \mathbb{Z}^p \times \mathbb{R}^q$. Let $(\varphi_j)_{j \in \mathbb{N}}$ be any orthonormal basis system for the space of square-integrable functions on $\mathcal{S}$. Then one can approximate the density $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ with an arbitrary degree of* *accuracy by a partial sum of the form*

$$\sum_{j=1}^{J} \theta_j \varphi_j(\boldsymbol{z}, \boldsymbol{x}), \qquad \textit{where } \theta_j = \int_{\mathcal{S}} \varphi_j(\boldsymbol{z}, \boldsymbol{x}) f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) d\boldsymbol{z} d\boldsymbol{x}.$$

The cutoff *parameter J controls the complexity of the expansions, and thereby the accuracy of the approximation. Possible choices $(\varphi_j)_{j\in\mathbb{N}}$ include spline, polynomial, and wavelet systems (see, e.g.,* Walter and Shen, 2000*, for an overview).*

In order to use such an expansion for density estimation, the coefficients θ_j have to be estimated from data. Noting that $\theta_j = \mathrm{E}\{\varphi_j(\boldsymbol{Z}+\boldsymbol{\epsilon}, \boldsymbol{X})\}$ suggest a straightforward estimator,

$$\widetilde{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \boldsymbol{X}_i).$$

Example 6.10 (Local linear regression)**.** *The jittering local linear regression estimator $\widetilde{m}(\boldsymbol{z}, \boldsymbol{x})$ of $m(\boldsymbol{z}, \boldsymbol{x}) = E(Y \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x})$ is solving*

$$\sum_{i=1}^{n} \{Y_i - \widetilde{m}(\boldsymbol{z}, \boldsymbol{x}) - \boldsymbol{\beta}^\top (\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \boldsymbol{X}_i) + \boldsymbol{\beta}^\top (\boldsymbol{z}, \boldsymbol{x})\} w_i(\boldsymbol{z}, \boldsymbol{x}) = 0,$$

with random weights

$$w_i(\boldsymbol{z}, \boldsymbol{x}) = K\left(\frac{\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{z}}{h_n}\right) K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{b_n}\right),$$

and h_n, b_n and K are as in Example 6.7*. With $\boldsymbol{\epsilon}_i = 0$ for all $i = 1, \ldots, n$, we recover the classical local linear regression estimator (e.g.,* Fan and Gijbels, 1996*). Higher-order local polynomial estimators can be constructed similarly.*

Example 6.11 (Orthogonal series regression)**.** *Similar to* Example 6.9*, let $(\varphi_j)_{j\in\mathbb{N}}$ be a basis of the space of square-integrable functions. If we approximate the regression function $m(\boldsymbol{z}, \boldsymbol{x}) = E(Y \mid \boldsymbol{Z} + \boldsymbol{\epsilon} = \boldsymbol{z}, \boldsymbol{X} = \boldsymbol{x})$ by*

$$m(\boldsymbol{z}, \boldsymbol{x}) = \sum_{j=1}^{J} \theta_j \varphi_j(\boldsymbol{z}, \boldsymbol{x}),$$

we can estimate the coefficients by solving

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{J} \widetilde{\theta}_j \varphi_j(\boldsymbol{Z}_i + \epsilon, \boldsymbol{X}_i) \right\} = 0.$$

This formulation can be easily extended to include penalties on the coefficient vector $\widetilde{\boldsymbol{\theta}}$.

6.4 Application: diagnosis of retinopathy

6.4.1 The problem

We consider a classification problem from the medical sciences. The goal is to diagnose diabetic retinopathy (a disease resulting from diabetes mellitus) from images of the retina. The retinal images have been preprocessed and a total of 19 features have been extracted. Three features are binary categories, six are integer valued count variables, and the remaining 10 features are continuous measurements. For more information about the pre-processing and features, we refer to Antal and Hajdu (2014). The data set contains diagnosis and characteristics of 1151 patients and can be downloaded from the UC Irvine Machine Learning Repository (Lichman, 2013).

6.4.2 A classification model based on the joint density

We will derive a prediction for the diagnosis from the joint density of the features and the diagnosis. Define $Y = \mathbb{1}(\text{patient suffers from retinopathy})$, and $(\boldsymbol{Z}, \boldsymbol{X})$ as the discrete-continuous vector of predictors. The conditional probability of Y can be expressed in terms of the joint density of $(Y, \boldsymbol{Z}, \boldsymbol{X})$:

$$\Pr(Y_i = 1 | \boldsymbol{Z}_i, \boldsymbol{X}_i) = \frac{f_{Y,\boldsymbol{Z},\boldsymbol{X}}(1, \boldsymbol{Z}_i, \boldsymbol{X}_i)}{f_{Y,\boldsymbol{Z},\boldsymbol{X}}(0, \boldsymbol{Z}_i, \boldsymbol{X}_i) + f_{Y,\boldsymbol{Z},\boldsymbol{X}}(1, \boldsymbol{Z}_i, \boldsymbol{X}_i)}.$$

A sensible rule is to diagnose retinopathy when $\Pr(Y = 1 | \boldsymbol{Z}_i, \boldsymbol{X}_i) > \alpha$, where $\alpha \in [0, 1]$ is a tuning parameter: increasing α reduces both the *false positive rate* and *true positive rate*.

The joint density $f_{\boldsymbol{Z},\boldsymbol{X}}$ is unknown and needs to be estimated. We will compare two different methods: the mixed data kernel estimator of Li and Racine (2003) and a version of the vine copula based density estimator from Chapter 4. Implementations of both methods are publicly available as R packages, see Hayfield and Racine (2008) and Nagler (2017).

The theory justifying the vine copula based estimator requires that all variables are continuous. To make it applicable to mixed data, we consider a jittering version of this estimator. That also allows us to bypass several complications arising in copula models for discrete data, cf., Genest and Neslehova (2007) and Panagiotelis et al. (2012). Similar to Section 4.5, we estimate the marginal densities by the classical kernel density estimator and the pair-copulas by a transformation kernel approach (see, Geenens et al., 2017). To allow for consistent estimation of the vine copula density, the marginal kernel estimator must be uniformly consistent (see Assumption 4.1). We facilitate this by choosing $\eta = f_{U_{0.1,5}}$ which ensures that the marginal densities are sufficiently smooth (see Example 6.2).

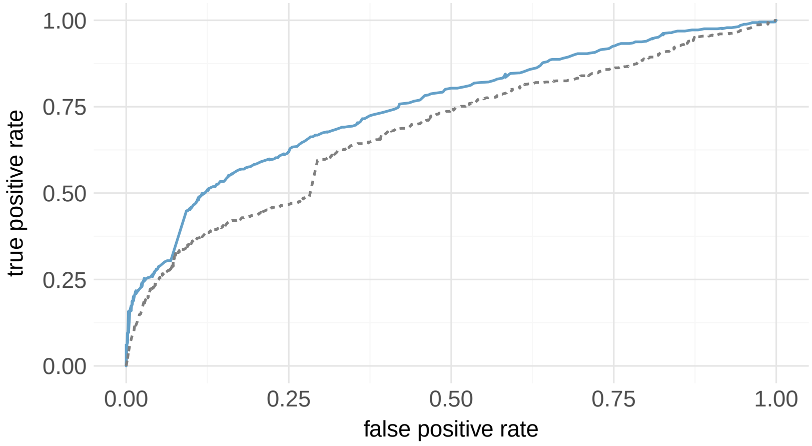


Figure 6.6: The true positive rate as a function of the false positive rate for the joint density classifiers based on Li and Racine (2003, dashed) and vine copulas (solid).

6.4.3 Results

Figure 6.6 shows the true positive rate (proportion of correctly classified retinas with retinopathy) as a function of the false positive rate (proportion of misclassified retinas without retinopathy) achieved by the two methods. All results are based on ten-fold cross-validation. The vine copula based estimator performs uniformly better than the one of Li and Racine (2003), yielding a better true positive rate for each level of the false positive rate. This illustrates that the jittering trick combined with sophisticated estimators can yield much better results than methods specialized to the mixed data setting. The effort to achieve this is minimal: we just add artificial noise to the original data and apply an existing estimator.

6.5 Discussion

6.5.1 Benefits

The most obvious benefit of jittering estimators is convenience. For their implementation, all one needs is an estimator that works in the continuous setting and a way to simulate random noise. This is easily achieved in modern statistical software. At second glance, the method opens many possibilities to extend existing estimators to the mixed data setting. This is increasingly useful with increasing complexity of the estimators. In many cases, there is otherwise no straightforward way to adapt an estimator to mixed data.

A less obvious benefit arises for studying general properties of a nonparametric function estimation problem. In the continuous setting, asymptotic arguments are often easier and well-established. For example, jittering arguments make it straightforward to derive minimax-optimal rates of convergence in nonparametric mixed data models; see Section 7.5.

6.5.2 Issues and open questions

Curse of dimensionality

A key issue for nonparametric estimators is the *curse of dimensionality*. In a continuous setting, the speed of convergence decreases exponentially in the dimension. For example, the classical convergence rate for estimating a d-dimensional continuous density is $n^{-2/(4+d)}$. A discrete density on the other hand can always be estimated with $n^{-1/2}$ rate. It is not obvious, which regime jittering estimators fall into, since a discrete density is estimated by exchanging it with a continuous surrogate.

Unfortunately, this question has no general answer and depends on the estimators' characteristics. The main criterion is how "local" the estimator operates; or more specifically, if the estimator is only affected by data in a shrinking neighborhood. For example, B-spline methods and kernel estimators with a compact kernel function will usually fall into the discrete regime, whereas Bernstein polynomials and kernel estimators with unbounded kernels fall into the continuous one. But we should stress that such considerations are only asymptotic and the behavior on finite samples will likely fall somewhere in between.

Efficiency

Typically, adding noise brings about some unnecessary variance. The magnitude of this effect depends on the characteristics of the estimator. Generally, this additional variance can be reduced by averaging estimates over multiple independent jitters (cf., Genest et al., 2017). In specific cases, a jittering estimator can be inherently efficient, with no need for averaging (see Section 7.4.1).

Choice of noise distribution

When using the jittering technique, an immediate question is which noise distribution to choose. The necessary conditions given in Proposition 1 are fairly broad and allow for a variety of noise distributions.

A referee asked whether it would be possible to preserve some dependence characteristics of the data. Unfortunately, dependence between discrete variables and its connection to the continuous counterpart is a highly subtle issue. One such subtlety is that there is no density when continuous variables are perfectly dependent, but the probability mass function for perfectly dependent variables exists. Genest and Neslehova (2007) address many other interesting issues. The article also provides some arguments for using independent noise, because it is the only way to preserve the equality between probabilistic and analytical definitions of some margin-free dependence measures like Kendall's τ and Spearman's ρ (their equation 7) or tie-corrected versions (p. 495).

In any case, one should understand jittering as an estimation technique rather than a modeling technique. Interpreting the jittered model independently of the "true" one is unlikely to be beneficial. In this chapter, the only criterion for validity of jittering was consistency of estimators. But we should expect that a data-driven choice of noise distribution would improve estimators' accuracy. A closer examination of the noise distribution's effect will be a promising path for future research.

Restriction to nonparametric techniques

Finally, we should warn that this methodology is only valid for nonparametric estimators. Usually, the shape of functionals of the jittered density can not be captured by parametric models, leading to estimators that are inconsistent.

Supplementary material

- https://github.com/tnagler/cctools: an R package providing tools for jittering.
- https://github.com/tnagler/jdify: an R package providing functionality for joint density classification.
- https://gist.github.com/tnagler/843f5c658e1139ff669d33614cc727e6: R code replicating the results from Section 6.4.

7 Asymptotic analysis of the jittering kernel density estimator

7.1 Introduction

Multivariate density estimation is a central field in nonparametric statistics. Yet many popular methods have a significant drawback in applications: they can only be applied to continuous data. Some estimators have been specifically designed to allow for mixed continuous and discrete data (Ahmad and Cerrito, 1994, Li and Racine, 2003, Hall et al., 1983, Efromovich, 2011), but the number is small compared to the methods available in a purely continuous framework.

A common trick among practitioners is to make the discrete variables continuous by adding a small amount of noise. The noisy data is continuous and the usual nonparametric estimators apply. But the addition of random noise can introduce bias, so this procedure generally lacks justification. In Chapter 6, we showed that adding noise still allows for valid estimates when the noise comes from a certain class of distributions. Then any nonparametric density estimator can be used in the mixed data setting. The resulting estimators are called *jittering estimators*.

Jittering estimators have so far been neglected in academic research, likely due to the widespread concern that jittering causes a loss in efficiency. The main objective of this chapter is to demonstrate that this concern is usually unjustified. We give an in-depth analysis of a simple instance from the class of jittering estimators: the *jittering kernel density estimator*, which is the jittering analog of the classical kernel density estimator (Parzen, 1962, Rosenblatt, 1956, Wand, 1992). We shall show that it maintains all the properties expected from a good nonparametric density estimator:

1. It is asymptotically normal and asymptotically unbiased for discrete variables (Theorem 7.1).
2. It is strongly and uniformly consistent (Theorem 7.2).
3. It can be fully efficient (Section 7.4.1).
4. It converges at minimax-optimal rates for a large class of target densities (Theorem 7.3 and Theorem 7.4). To the best of the author's knowledge, these are the first results on minimax-optimality of nonparametric density estimators for mixed data.

Although focus is on only one instance of the class of jittering estimators, we can expect that others have similar properties.

The remainder of this chapter is organized as follows. Section 7.2 introduces the the jittering estimator and some assumptions. Section 7.3 gives a comprehensive asymptotic analysis which is complemented by a study of the asymptotic efficiency and finite sample bias in the univariate discrete setting (Section 7.4). Section 7.5 establishes with minimax-optimal rates for density estimation in a nonparametric mixed data model. Section 7.6 supports demonstrates that the estimator is also competitive on finite samples; Section 7.7 offers conclusions. Proofs of all theorems are deferred to Section 7.8.

7.2 The estimator

Suppose that $(\boldsymbol{Z}, \boldsymbol{X})$ is a random vector with discrete component $\boldsymbol{Z} \in \mathbb{Z}^p$ and continuous component $\boldsymbol{X} \in \mathbb{R}^q$. We explicitly allow for the cases where $p \geq 1$, $q = 0$ (all variables are discrete) and $p = 0$, $q \geq 1$ (all variables are continuous). Our goal is to estimate the density f of $(\boldsymbol{Z}, \boldsymbol{X})$ based on 'observations' $(\boldsymbol{Z}_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$, which are *iid* random vectors having the same distribution as $(\boldsymbol{Z}, \boldsymbol{X})$. In this context, f is the density with respect to the product of the counting and Lebesgue measures, i.e.,

$$f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) = \frac{\partial^q}{\partial x_1 \cdots \partial x_q} \Pr(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} \leq \boldsymbol{x}).$$

Let K be a real-valued function, called *kernel*, and abbreviate $K(\boldsymbol{w}) = \prod_{j=1}^k K(w_j)$ for any $\boldsymbol{w} \in \mathbb{R}^k$, $k \in \mathbb{N}$. The classical kernel density estimator is defined as

$$\widehat{f}(\boldsymbol{z}, \boldsymbol{x}) = \frac{1}{n h_n^p b_n^q} \sum_{i=1}^n K\left(\frac{\boldsymbol{Z}_i - \boldsymbol{z}}{h_n}\right) K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{b_n}\right), \tag{7.1}$$

where $h_n, b_n > 0$ are called *bandwidth parameters* and control the amount of smoothing. The above definition of the estimator is simplified to ease our exposition: we use only one parameter (h_n) for smoothing all components of $\boldsymbol{Z}$ and one parameter (b_n) for smoothing the components of $\boldsymbol{X}$. In practice, one would use a single parameter for each variable or even a bandwidth matrix (see, e.g., Scott, 2008).

The estimator $\widehat{f}$ only works for continuous random vectors. To make it applicable to mixed data, we make all discrete variables continuous by adding noise. Let $\boldsymbol{\epsilon}_i \in \mathbb{R}^p$, $i = 1, \dots, n$, be *iid* random vectors independent from $(\boldsymbol{Z}_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$. Suppose further that the p components of $\boldsymbol{\epsilon}_i$ are *iid* with density η and denote the joint density of $(\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \boldsymbol{X}_i)$ by f_η. The jittering kernel density estimator is defined as the classical kernel density estimator applied to

$(\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$:

$$\widetilde{f}(\boldsymbol{z}, \boldsymbol{x}) = \frac{1}{n h_n^p b_n^q} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{z}}{h_n}\right) K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{b_n}\right). \tag{7.2}$$

To facilitate our analysis, the following conditions are imposed on the kernel function:

Assumption 7.1.

K1: $K \colon [-1, 1] \to \mathbb{R}_{\geq 0}$ is a continuous function satisfying $\int K(t)dt = 1$.

K2: There is $\ell \in \mathbb{N}$, $\ell \geq 2$, such that for $k = 1, \dots, \ell - 1$,

$$\int_{[0,1]} t^k K(t)dt = 0, \qquad \int_{[0,1]} t^\ell K(t)dt > 0.$$

Remark 7.1. *A kernel function satisfying K2 is called ℓ-th order kernel (see, e.g., Marron, 1994).* □

We further assume that the noise density η belongs to the class $\mathcal{E}_{\gamma_1,\gamma_2}$, as defined in Definition 6.2. The first equality implies that we can equivalently estimate f_η instead of f, which is convenient because f_η is the density of a purely continuous random vector. Additionally, all derivatives with respect to $\boldsymbol{z}$ vanish, which makes estimation even easier (see Proposition 6.2).

Remark 7.2. *The estimator $\widetilde{f}$ is similar to the estimators of Ahmad and Cerrito 1994 and Li and Racine (2003). The difference lies in the kernel function for discrete data. The estimators of Ahmad and Cerrito 1994 and Li and Racine (2003) use a deterministic kernel function which is defined on the integers. In contrast, the jittering kernel density estimator (7.2) uses a random kernel $K\{(\cdot + \boldsymbol{\epsilon}_i)/b_n\}$ defined on a compact subset of $\mathbb{R}^p$, where randomness is induced by $\boldsymbol{\epsilon}_i$.*

7.3 Asymptotic analysis in the general setting

7.3.1 Asymptotic distribution

We first study the asymptotic distribution of the jittering kernel density estimator. To motivate our first theorem, we recall a a classical result from kernel density estimation in the purely continuous setting (e.g., Wand, 1992). If f is the density of a continuous random vector $(\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$, sufficiently smooth, $\ell = 2$ in

Assumption 7.1, and $h_n, b_n \to 0$, $nh_n^p b_n^q \to \infty$, then

$$\begin{aligned}
\mathrm{E}\big\{\widehat{f}(\boldsymbol{z},\boldsymbol{x})\big\} &= f(\boldsymbol{z},\boldsymbol{x}) + \frac{h_n^2\sigma_2}{2}\sum_{k=1}^{p}\frac{\partial^2 f(\boldsymbol{z},\boldsymbol{x})}{\partial z_k^2} + \frac{b_n^2\sigma_2}{2}\sum_{j=1}^{q}\frac{\partial^2 f(\boldsymbol{z},\boldsymbol{x})}{\partial x_j^2} \\
&\quad + o\big(h_n^2 + b_n^2\big), \\
\mathrm{Var}\big\{\widehat{f}(\boldsymbol{z},\boldsymbol{x})\big\} &= \frac{\kappa^{p+q} f(\boldsymbol{z},\boldsymbol{x})}{nh_n^p b_n^q} + o\bigg(\frac{1}{nh_n^p b_n^q}\bigg),
\end{aligned} \tag{7.3}$$

where κ and σ_2 are constants defined in Theorem 7.1 below.

Recall that $\widetilde{f}$ is nothing else than $\widehat{f}$ applied to $(\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \boldsymbol{X}_i)$, $i = 1,\dots,n$. Proposition 6.2 showed that $f_\eta(\boldsymbol{z},\boldsymbol{x})$, the density of $(\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \boldsymbol{X}_i)$, has vanishing derivatives with respect to $\boldsymbol{z}$. We can thus expect the first sum in the bias term in (7.3) to vanish asymptotically. In fact, it becomes exactly zero when $h_n \leq \min\{\gamma_1, 1-\gamma_2\}$. The following result improves upon the properties implied by (7.3) by taking these considerations into account.

Assumption 7.2.

A1: $f(\boldsymbol{z},\boldsymbol{x})$ is $\ell+1$ times continuously differentiable with respect to $\boldsymbol{x}$.

A2: $K1$ and $K2$ hold with $\ell \geq 2$.

A3: $\eta \in \mathcal{E}_{\gamma_1,\gamma_2}$.

A4: $b_n \to 0$ and $nh_n^p b_n^q \to \infty$ as $n \to \infty$.

A5: There exists an $n_0 \in \mathbb{N}$, such that $h_n \leq \min\{\gamma_1, 1-\gamma_2\}$ for all $n \geq n_0$.

Theorem 7.1. *Under assumptions A1-A5, it holds for any $(\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$,*

$$\begin{aligned}
\mathrm{E}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} &= f(\boldsymbol{z},\boldsymbol{x}) + \frac{b_n^\ell \sigma_\ell}{\ell!}\sum_{j=1}^{q}\frac{\partial^\ell f(\boldsymbol{z},\boldsymbol{x})}{\partial x_j^\ell} + o(b_n^\ell), \\
\mathrm{Var}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} &= \frac{f(\boldsymbol{z},\boldsymbol{x})}{nb_n^q}\big\{h_n^{-p}\kappa^{p+q} - b_n^q f(\boldsymbol{z},\boldsymbol{x})\big\} + o\bigg(\frac{1}{nh_n^p b_n^q}\bigg),
\end{aligned}$$

where $\sigma_\ell = \int_{-1}^{1} s^\ell K(s)ds$ and $\kappa = \int_{-1}^{1} K^2(s)ds$. If further $nh_n^p b_n^{q+2\ell} = O(1)$,

$$\frac{\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - \mathrm{E}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\}}{\mathrm{Var}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\}} \xrightarrow{d} \mathcal{N}(0,1).$$

Remark 7.3. *The assumptions in Theorem 7.1 differ from those usually made in the continuous framework. There are no assumptions on the smoothness of $\widehat{f}_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z},\boldsymbol{x})$ with respect to $\boldsymbol{z}$, because its local behavior is controlled by $\eta \in \mathcal{E}_{\gamma_1,\gamma_2}$.*

Further, h_n is not required to vanish asymptotically, but should be less than $\min\{\gamma_1, 1-\gamma_2\}$ *for large n. This is sufficient to ensure that there is no bias with respect to $\boldsymbol{z}$. Further decreasing h_n does not change the bias, but inflates the variance.* □

Remark 7.4. *The asymptotic variance does not involve on η or its class parameters γ_1 and γ_2 (and neither does the asymptotic bias). Intuitively, we would expect an increase in the estimator's variance because we are adding random noise. Apparently this effect is dominated by the sampling variability in the original data and asymptotically negligible. So there should be no benefit from averaging over multiple jitters (at least asymptotically). This is in contrast to empirical processes of jittered data (Genest et al., 2017).*

7.3.2 Asymptotically optimal bandwidths

A standard tool for studying optimal bandwidths is the *asymptotic mean squared error*,

$$\mathrm{AMSE}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} = \big[\mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} - f(\boldsymbol{z},\boldsymbol{x})\big]^2 + \mathrm{Var}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\}.$$

Under the assumptions of Theorem 7.1, we get

$$\mathrm{AMSE}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} \approx \frac{b_n^{2\ell}\sigma_\ell^2}{(\ell!)^2}\left(\sum_{j=1}^q \frac{\partial^\ell f(\boldsymbol{z},\boldsymbol{x})}{\partial x_j^\ell}\right)^2 + \frac{f(\boldsymbol{z},\boldsymbol{x})}{nb_n^q}\big\{h_n^{-p}\kappa^{p+q} - b_n^q f(\boldsymbol{z},\boldsymbol{x})\big\}.$$

For $h_n = O(1)$, it is easy to check that the bandwidth b_n minimizing the AMSE satisfies $b_n \sim n^{-1/(2\ell+q)}$. This is well-known as the optimal rate for the classical kernel density estimator when $p = 0$. The AMSE further suggests that it is optimal to choose h_n as large as possible. The largest h_n allowed by A5 is $h_n = \min\{\gamma_1, 1-\gamma_2\}$. Asymptotically, this is the optimal bandwidth. We shall see shortly that this choice means that we are not smoothing the discrete variables at all. This is not unreasonable: in contrast to the continuous case, smoothing discrete variables is not necessary for consistent nonparametric estimation (for a discussion, see, Simar et al., 2011).

On finite samples $h_n = \min\{\gamma_1, 1-\gamma_2\}$ can be too small. Recall that K is zero unless $|x| < 1$ and suppose that $h_n \leq 1-\gamma_2$. Then $K\{(\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{z})/h_n\} > 0$ implies $|Z_{i,k} + \epsilon_{i,k} - z_k| < 1-\gamma_2$ for all $k = 1,\dots,p$. Since $|\epsilon_{i,k}| < \gamma_2$, that is only possible when $|Z_{i,k} - z_k| < 1-\gamma_2 + |\epsilon_{i,k}| < 1$ for all k, which implies $\boldsymbol{Z}_i = \boldsymbol{z}$. In this case the estimator can be written as

$$\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) = \frac{1}{nh_n^p b_n^q}\sum_{i:\,\boldsymbol{Z}_i=\boldsymbol{z}} K\left(\frac{\boldsymbol{\epsilon}_i}{h_n}\right)K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{b_n}\right).$$

It neglects all observations where $\boldsymbol{Z}_i \neq \boldsymbol{z}$ and, thus, does not smooth with respect to the discrete variables. This also means that $\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) = 0$ if $\boldsymbol{Z}_i \neq \boldsymbol{z}$ for all $i = 1,\dots,n$. Theorem 7.1 implicitly assumes that n is large enough to provide sufficiently many observations with $\boldsymbol{Z}_i = \boldsymbol{z}$. This is guaranteed asymptotically whenever $P(\boldsymbol{Z} = \boldsymbol{z}) > 0$, but often demands sample sizes much larger than what is common.

We conclude that Theorem 7.1 is not useful for bandwidth selection on samples of small or moderate size. Cross-validation techniques are more appropriate tools in the mixed data setting (see, e.g., Aitchison and Aitken, 1976, Racine and Li, 2004, Hall et al., 2004).

7.3.3 Consistency

Theorem 7.1 implies pointwise consistency of the jittering kernel density estimator, but assumption A1 is more strict than necessary. The following result weakens this assumption and additionally establishes strong uniform consistency.

Assumption 7.3.

A1′: The $(\ell-1)$th derivative of $f(\boldsymbol{z},\boldsymbol{x})$ exists and is uniformly Lipschitz on $S \subseteq \mathbb{Z}^p \times \mathbb{R}^q$.

Theorem 7.2. *Suppose that assumptions A1′, A2–A5 hold. Then, for all $(\boldsymbol{z},\boldsymbol{x}) \in S$,*

$$\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x}) = O_p\{b_n^\ell + (nh_n^p b_n^q)^{-1/2}\}, \tag{7.4}$$

$$\sup_S \bigl|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x})\bigr| = O_{a.s.}\left\{b_n^\ell + \left(\frac{\max\{\ln\ln n, \ln h_n^{-1}, \ln b_n^{-1}\}}{nh_n^p b_n^q}\right)^{1/2}\right\}. \tag{7.5}$$

Remark 7.5. *If there are $h_0 > 0, n_0 \in \mathbb{N}$ such that $h_n \in (h_0, \min\{\gamma_1, 1-\gamma_2\}]$ for all $n \geq n_0$, the rates of convergence in Theorem 7.2 become*

$$\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x}) = O_p\{b_n^\ell + (nb_n^q)^{-1/2}\},$$

$$\sup_S \bigl|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x})\bigr| = O_{a.s.}\left\{b_n^\ell + \left(\frac{\max\{\ln\ln n, \ln b_n^{-1}\}}{nb_n^q}\right)^{1/2}\right\},$$

which do not involve p, the dimension of the discrete variables. So adding more discrete variables does not change the convergence rate of the estimator. In particular, there is no cost for recoding unordered categorical variables into several binary variables. □

Remark 7.6. *To find the best possible rates in Theorem 7.2, we minimize the expressions with respect to h_n and b_n under the constraint that $h_n = O(1)$.*

(i) The best rate in (7.4) is $n^{-\ell/(2\ell+q)}$ and achieved when $h_n \sim 1$ and $b_n \sim n^{-1/(2\ell+q)}$.

(ii) *For* $q > 0$*, the best rate in* (7.5) *is* $(n/\ln n)^{-\ell/(2\ell+q)}$ *and achieved when* $h_n \sim 1$*,* $b_n \sim (n/\ln n)^{-1/(2\ell+q)}$*.*

(iii) *For* $q = 0$*, the best rate in* (7.5) *is* $(n/\ln\ln n)^{-1/2}$ *and achieved when* $h_n \sim 1$*.*

7.4 A closer look at the univariate discrete setting

The jittering kernel density estimator $\widetilde{f}$ handles continuous variables just like the classical kernel density estimator. How it smooths discrete variables is less obvious. To gain a better understanding, we study its asymptotic efficiency and finite sample bias when there is only one discrete variable ($p = 1$, $q = 0$).

7.4.1 Asymptotic efficiency

For convenience, set $h_n \equiv \min\{\gamma_1, 1 - \gamma_2\}$. The expectation and variance in Theorem 7.1 become

$$\mathrm{E}\big\{\widetilde{f}(z)\big\} = f(z), \quad \mathrm{Var}\big\{\widetilde{f}(z)\big\} = \frac{f(z)}{n}\big[\min(\gamma_1, 1-\gamma_2)^{-1}\kappa - f(z)\big] + o(n^{-1}),$$

The most efficient point estimator for a discrete probability $f(z) = \Pr(Z = z)$ is the sample frequency $f_n(z) = n^{-1}\sum_{i=1}^n \mathbb{1}(Z_i = z)$. It satisfies

$$\mathrm{E}\big\{f_n(z)\big\} = f(z), \quad \mathrm{Var}\big\{f_n(z)\big\} = \frac{f(z)}{n}\big\{1 - f(z)\big\}.$$

The *asymptotic relative efficiency (ARE)* of $\widetilde{f}$ relative to f_n is defined as

$$\mathrm{ARE}\big\{\widetilde{f}(z) : f_n(z)\big\} = \frac{\mathrm{AVar}\{f_n(z)\}}{\mathrm{AVar}\{\widetilde{f}(z)\}},$$

where AVar denotes the leading term of an asymptotic expansion of the variance. The ARE is interpreted as follows: If the estimator $\widetilde{f}$ is used with n observations, then one needs ARE × n observations to obtain the same accuracy with f_n. If the ARE is less than one, then f_n needs less observations, i.e., f_n is more efficient than $\widetilde{f}$. If the ARE is greater then one, it is the other way around. If it is exactly one, the two estimators are equally efficient.

Straightforward calculations yield

$$\begin{aligned}\mathrm{ARE}\big\{\widetilde{f}(z) : f_n(z)\big\} &= \frac{1 - f(z)}{\min\{\gamma_1, 1-\gamma_2\}^{-1}\kappa - f(z)} \\ &= \left(1 + \frac{\min\{\gamma_1, 1-\gamma_2\}^{-1}\kappa - 1}{1 - f(z)}\right)^{-1} \leq 1.\end{aligned}$$

The relative efficiency depends on three quantities:

- It is increasing in $\min\{\gamma_1, 1-\gamma_2\}$ and the most efficient choice is $\gamma_1 = \gamma_2 = 1/2$, which corresponds to the uniform error density on $(-1/2, 1/2)$. On the other hand, the relative efficiency approaches 0 for $\gamma_1 \to 0$ or $\gamma_2 \to 1$.

- It is decreasing in κ, which is the roughness of the kernel K. The 'least rough' kernel is the is the uniform kernel, i.e., $K(x) = 2^{-1}\mathbb{1}(|x| \leq 1)$, for which $\kappa = 1/2$. But this kernel is rather unpopular in practice. A more widely used kernel is the Epanechnikov kernel, $K(x) = 3/4(1-x^2)\mathbb{1}(|x| \leq 1)$, for which $\kappa = 0.6$.

- It is decreasing in $f(z)$. The worst case is that $f(z) = 1$, for which the ARE is zero unless $\gamma_1 = \gamma_2 = \kappa = 1/2$. For a Bernoulli(1/2) variable, Uniform$(-1/2, 1/2)$ noise, and the Epanechnikov kernel, we get ARE $\approx$ 0.71.

Remark 7.7. *Suppose η is the uniform density on $(-1/2, 1/2)$ (for which $\gamma_1 = \gamma_2 = 1/2$), $h_n = 1/2$, and K is the uniform kernel (for which $\kappa = 1/2$). Then, the two estimators are equally efficient. In fact, since $|Z_i - z| \geq 1$ if $Z_i \neq z$, the estimator $\widetilde{f}$ becomes*

$$\begin{aligned}
\widetilde{f}(z) &= \frac{1}{nh_n}\sum_{i=1}^n 2^{-1}\mathbb{1}(|Z_i + \epsilon_i - z| \leq h_n) \\
&= \frac{2}{n}\sum_{i=1}^n 2^{-1}\mathbb{1}(|Z_i + \epsilon_i - z| \leq 1/2) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z_i = z),
\end{aligned}$$

which is exactly the sample frequency estimator f_n. □

7.4.2 Finite sample bias

Assuming $h_n \leq \min\{\gamma_1, 1-\gamma_2\}$, Theorem 7.1 shows that $\widetilde{f}$ is unbiased in a purely discrete setting. On small samples, it is often necessary to choose a larger bandwidth (see Section 7.3.2). When $h_n > \min\{\gamma_1, 1-\gamma_2\}$, the estimator $\widetilde{f}$ is usually biased.

Lemma 7.1. *Suppose that $\eta \in \mathcal{E}_{\gamma_1,\gamma_2}$ and K satisfies $K1$–$K2$. Then,*

$$\begin{aligned}&\mathrm{E}\{\widetilde{f}(z)\} - f(z)\\ &= \sum_{k=1}^{\lceil h_n - 1/2\rceil} \frac{\rho_k^\eta(h_n) f(z+k) - \{\rho_k^\eta(h_n) + \rho_{-k}^\eta(h_n)\} f(z) + \rho_{-k}^\eta(h_n) f(z-k)}{k^2},\end{aligned}$$

where $\rho_k^\eta(h_n) = k^2 \int_{A_k^\eta(h_n)} K(t)\eta(k - h_n t)dx$ and

$$A_k^\eta(h_n) = \left[(1-\gamma_2-k)h_n^{-1},\, (-1+\gamma_2-k)h_n^{-1}\right] \cap [-1,1].$$

To interpret the bias, it is helpful to focus on a simple case first. For $\eta(x) = \mathbb{1}(|x| \leq 1/2)$ and symmetric K, we have $\rho_{-k}^\eta = \rho_k^\eta$ and therefore the following corollary.

Corollary 7.1. *Suppose that $\eta(x) = \mathbb{1}(|x| \leq 1/2)$ (i.e., $\eta \in \mathcal{E}_{0.5,0.5}$) and K is a symmetric function satisfying $K1$–$K2$. Then for all $z \in \mathbb{Z}$,*

$$\mathrm{E}\{\widetilde{f}(z)\} = f(z) + \sum_{k=1}^{\lceil h_n - 1/2\rceil} \rho_k(h_n)\Delta_k^2 f(z),$$

where

$$\Delta_k^2 f(z) = \frac{f(z+k) - 2f(z) + f(z-k)}{k^2},$$

and $\rho_k(h_n) = k^2 \int_{A_k(h_n)} K(t)dt$ with

$$A_k(h_n) = \left[(1/2-k)h_n^{-1},\, (-1/2-k)h_n^{-1}\right] \cap [-1,1].$$

The operator Δ_k^2 is known as the *second order central difference operator* (e.g., Monahan, 2011). It is commonly used as numerical approximation of second order derivative of real-valued functions, which is

$$\frac{d^2 f(x)}{dx^2} = \lim_{s\to 0} \frac{f(x+s) - 2f(x) + f(x-s)}{s^2}.$$

We can interpret $\Delta_k^2 f$ as a discrete analogue to the second order derivative of a real-valued function. In this aspect, the discrete setting is similar to the continuous one (where the bias of $\widetilde{f}$ is proportional to the second order derivative).

The parameter k is called the *step size* and determines how local the derivative approximation is. The bias of $\widetilde{f}$ is a weighted sum of such 'derivatives' for several values of k. The bandwidth h_n limits the maximal step size and thereby controls the locality of the bias. Although not universally true, smaller values of h_n typically correspond to a smaller bias. A simple counter example is when

$f(z+k) = f(z-k) = f(z)$ for all $k \leq \lceil h_n - 1/2 \rceil$, where the bias is zero for all $h_n' \leq h_n$. There are also situations where decreasing h_n leads to a larger bias. This phenomenon also exists in the continuous setting, but is disguised by asymptotic approximations. When $h_n \leq 1/2$ as in Theorem 7.1, the estimator is unbiased.

The bias in Lemma 7.1 can be interpreted similarly. But Δ_k^2 is replaced by a weighted approximation of the derivative. If η or K are asymmetric, different weights will be assigned to the 'forward derivative' $k^{-1}\{f(z+k) - f(z)\}$ and the 'backward derivative' $k^{-1}\{f(z-k) - f(z)\}$.

7.5 Minimax rate optimality

The *maximum risk* associated with a class of densities $\mathcal{F}$ and a (semi-) distance d is defined as

$$\mathcal{R}_n(\widehat{f}, \mathcal{F}, d) = \sup_{f \in \mathcal{F}} \mathrm{E}_f\big\{d^2(\widehat{f}, f)\big\}, \tag{7.6}$$

We consider two semi-distances that relate to pointwise and uniform consistency of $\widehat{f}$, respectively:

$$\begin{aligned} d_{(\boldsymbol{z},\boldsymbol{x})}(\widehat{f}, f) &= \big|\widehat{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x})\big|, \quad \text{for some } (\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q, \\ d_{\infty,\mathcal{S}}(\widehat{f}, f) &= \sup_{\mathcal{S}} \big|\widehat{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x})\big|, \quad \text{for some } \mathcal{S} \subset \mathbb{Z}^p \times \mathbb{R}^q. \end{aligned}$$

For $\mathcal{F}$, we shall consider all bounded density functions whose continuous part belongs to a Hölder class. For $\boldsymbol{a} \in \mathbb{N}_0^q$, we use the multi-index notations $|\boldsymbol{a}| = \sum_{j=1}^q a_j$, $\boldsymbol{x}^{\boldsymbol{a}} = x_1^{a_1} \cdots x_q^{a_q}$, and denote the partial derivatives of f with respect to $\boldsymbol{x}$ as

$$D_{\boldsymbol{x}}^{\boldsymbol{a}} f(\boldsymbol{z},\boldsymbol{x}) = \frac{\partial^{|\boldsymbol{a}|} f(\boldsymbol{z},\boldsymbol{x})}{\partial^{a_1} x_1 \cdots \partial^{a_q} x_q}. \tag{7.7}$$

Definition 7.1. *For $\lambda < \infty$ and $\beta = r + \alpha$, $r \in \mathbb{N}_0$, $0 < \alpha \leq 1$, the class $\mathcal{H}(\beta, \lambda)$ is defined as all functions $f\colon \mathbb{Z}^p \times \mathbb{R}^q \to \mathbb{R}$ such that for all $\boldsymbol{a} \in \mathbb{N}_0$ with $|\boldsymbol{a}| \leq r$,*

(i) f is a probability density on $\mathbb{Z}^p \times \mathbb{R}^q$,

(ii) $D_{\boldsymbol{x}}^{\boldsymbol{a}} f(\boldsymbol{z},\boldsymbol{x})$ exists for all $(\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$ and

$$\sup_{\boldsymbol{z} \in \mathbb{Z}^p, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^q} \left\{ \frac{\big|D_{\boldsymbol{x}}^{\boldsymbol{a}} f(\boldsymbol{z},\boldsymbol{x}) - D_{\boldsymbol{x}}^{\boldsymbol{a}} f(\boldsymbol{z},\boldsymbol{x}')\big|}{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^{\alpha}} + f(\boldsymbol{z},\boldsymbol{x}) \right\} \leq \lambda$$

Remark 7.8. *If $p \geq 1$ and $q = 0$, $\mathcal{H}(\beta, \lambda)$ contains all densities on $\mathbb{Z}^p$. If $p = 0$ and $q \geq 1$, it is a Hölder class on $\mathbb{R}^q$.* □

The following result establishes convergence rates of the jittering kernel density estimator with respect to the maximum risk.

Theorem 7.3. *Denote $\widetilde{f}$ as the estimator defined in Equation 7.2. Suppose $f \in \mathcal{H}(\beta, \lambda)$ and assumptions A2–A4 of Theorem 7.1 hold with $\ell \geq r + 1$, $\beta = r + \alpha$, $0 < \alpha \leq 1$, $\lambda < \infty$. Assume further that there are $h_0 > 0, n_0 \in \mathbb{N}$ such that $h_n \in (h_0, \min\{\gamma_1, 1 - \gamma_2\}]$ for all $n \geq n_0$. Then there exists $\bar{c} \in (0, \infty)$ such that*

$$\limsup_{n\to\infty} r_n^{-2} \mathcal{R}_n(\widehat{f}, \mathcal{F}, d) \leq \bar{c},$$

in each of the following cases:

(i) $r_n = n^{-\beta/(2\beta+q)}$, $d = d_{(\boldsymbol{z},\boldsymbol{x})}$

(ii) $r_n = (n/\ln n)^{-\beta/(2\beta+q)}$, $d = d_{\infty,\mathcal{S}}$, $q \geq 1$,

(iii) $r_n = n^{-1/2}$, $d = d_{\infty,\mathcal{S}}$, $q = 0$, $|\mathcal{S}| < \infty$,

(iv) $r_n = (n/\ln\ln n)^{-1/2}$, $d = d_{\infty,\mathcal{S}}$, $q = 0$, $|\mathcal{S}| = \infty$,

for arbitrary $(\boldsymbol{x}, \boldsymbol{z}) \in \mathbb{Z}^p \times \mathbb{R}^q$ and $\mathcal{S} \subset \mathbb{Z}^p \times \mathbb{R}^q$.

We shall see that the rates in Theorem 7.3 (i)–(iii) are optimal in a minimax sense. The minimax risk is defined as

$$\mathcal{R}_n^*(\mathcal{F}, d) = \inf_{\widehat{f}} \mathcal{R}_n(\widehat{f}, \mathcal{F}, d) = \inf_{\widehat{f}} \sup_{f\in\mathcal{F}} \mathrm{E}_f\big\{d^2(\widehat{f}, f)\big\},$$

where the infimum is taken over all possible estimators $\widehat{f}$ of f. In our context, an 'estimator' is any measurable function of $(\boldsymbol{Z}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$.

Definition 7.2. *A sequence of positive real numbers r_n is called*

(i) an upper bound on the minimax rate if there is $\bar{c}$ such that

$$\limsup_{n\to\infty} r_n^{-2} \mathcal{R}_n^*(\mathcal{F}, d) \leq \bar{c}.$$

(ii) a lower bound on the minimax rate if there is $\underline{c} > 0$ such that

$$\liminf_{n\to\infty} r_n^{-2} \mathcal{R}_n^*(\mathcal{F}, d) \geq \underline{c},$$

(iii) a minimax-optimal rate of convergence if both (i) and (ii) hold.

In a purely continuous setting, optimal rates have long been established (Stone, 1980, 1983, Ibragimov and Khas' minskii, 1983). To the best of the author's knowledge, there are no results on optimal rates in the mixed data setting.

To show that a rate is minimax-optimal, we have to check that it is both an upper and lower bound on the minimax rate. Theorem 7.3 already gives us an upper bound, since, for any estimator $\widehat{f}$,

$$\mathcal{R}_n^*(\mathcal{F}, d) = \inf_{\widehat{f}} \mathcal{R}_n(\widehat{f}, \mathcal{F}, d) \leq \mathcal{R}_n(\widehat{f}, \mathcal{F}, d).$$

Lower bounds on the minimax rate can be deduced easily by considering subsets of $\mathcal{H}(\beta, \lambda)$ for which lower bounds are known (see Section 7.8.4).

Theorem 7.4. *Let $\mathcal{S} \subset \mathbb{Z}^p \times \mathbb{R}^q$ and $(\boldsymbol{z}, \boldsymbol{x}) \in \mathcal{S}$. The minimax-optimal rate of convergence r_n^* associated with the class $\mathcal{H}(\beta, \lambda)$ and distance d satisfies*

(i) $r_n^* = n^{-\beta/(2\beta+q)}$, *for* $d = d_{(\boldsymbol{z},\boldsymbol{x})}$

(ii) $r_n^* = (n/\ln n)^{-\beta/(2\beta+q)}$, *for* $d = d_{\infty,\mathcal{S}}$, $q \geq 1$,

(iii) $r_n^* = n^{-1/2}$, *for* $d = d_{\infty,\mathcal{S}}$, $q = 0$, $|\mathcal{S}| < \infty$,

(iv) $r_n^* \in [n^{-1/2}, (n/\ln\ln n)^{-1/2}]$, *for* $d = d_{\infty,\mathcal{S}}$, $q = 0$, $|\mathcal{S}| = \infty$,

Remark 7.9. *Theorem 7.3 and Theorem 7.4 imply that the jittering kernel density estimator converges at minimax-optimal rates for cases (i)–(iii).* □

Remark 7.10. *Theorem 7.4 only provides an interval for the optimal rate in case (iv). Minimax analysis for this setting is surprisingly har; see (Han et al., 2015) for minimax rates with respect to the ℓ_1 distance. The interval is quite narrow, differing only by a factor of size* $\ln\ln n$. *The exact rate, however, remains an open problem.* □

7.6 Simulation experiments

The jittering kernel density estimator has appealing asymptotic properties. This may come as a surprise: since we are adding noise to the data, we could expect that the data become less informative and uncertainty increases. We complement our asymptotic arguments with a small numerical experiment that illustrates the small sample performance of the estimator. Because of its wide use and close resemblance to our approach, we will use the estimator of Li and Racine (2003) as a benchmark.

We use the following setup:

- We compare three estimators
  (i) `jkde`: the jittering kernel density estimator with noise density $\eta(x) = \mathbb{1}(|x| < 1/2)$, for which $\gamma_1 = \gamma_2 = 1/2$.
  (ii) `jkde2`: the jittering kernel density estimator with noise density $\eta(x) = f_{U_{1/4,5}}(x)$ (as in Example 6.2), for which $\gamma_1 = 3/8$, $\gamma_2 = 5/8$.

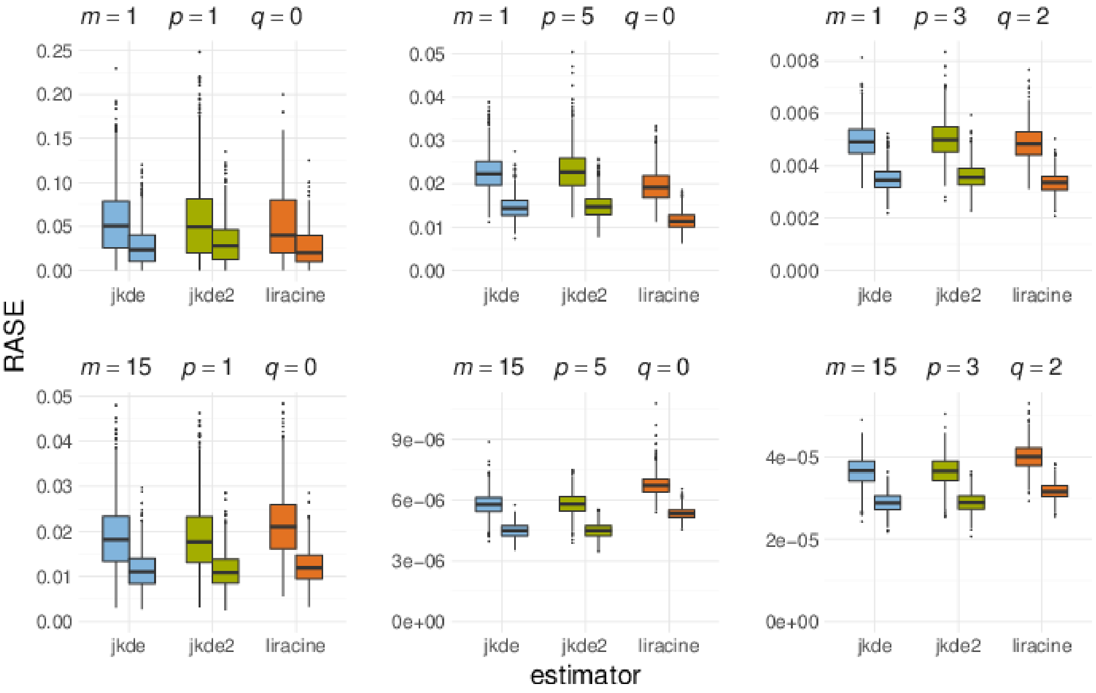


Figure 7.1: RASE achieved by the two estimators for various choices of p, q, and m. Each estimator is represented by two boxes; the left box corresponds to $n = 50$, the right to $n = 200$.

(iii) `liracine`: the estimator of Li and Racine (2003) as implemented in the `np` package (Hayfield and Racine, 2008).

Contrary to (7.2), we use one bandwidth parameter for each variable. All estimators use likelihood cross-validation for bandwidth selection.

- We estimate the density f of a vector $(\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$, where $Z_j \sim \text{Binomial}(m, 0.3)$ for all $j = 1, \dots, p$, $X_j \sim \mathcal{N}(0, 1)$ for all $j = 1, \dots, q$. For sake of simplicity, all variables are simulated independently.
- Results are based on $N_{\text{sim}} = 1000$ simulated data sets with sample sizes $n = 50, 200$.
- As a performance measure we use the *root average square error (RASE)* computed over a grid in $\mathbb{Z}^p \times \mathbb{R}^q$. More specifically, we use $\mathcal{Z} = \{0, \dots, m\}$, $\mathcal{X} = \{-2, -1.6, \dots, 2\}$, and

$$\text{RASE}(\widehat{f}, f) = \sqrt{\sum_{z_1 \in \mathcal{Z}} \cdots \sum_{z_p \in \mathcal{Z}} \sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_q \in \mathcal{X}} \left\{\widehat{f}(\boldsymbol{z}, \boldsymbol{x}) - f(\boldsymbol{z}, \boldsymbol{x})\right\}^2}.$$

Figure 7.1 shows the estimators' performance for various values of p, q and m. Each estimator is represented by two boxes, where the left box corresponds to $n = 50$ and the right box to $n = 200$. The choice of noise density seems to be of minor importance: `jkde` and `jkde2` give almost identical results. Compared to `liracine`, the two estimator show only small differences. The two jittering estimators are slightly more accurate in all scenarios with $m = 15$, and slightly less accurate when $m = 1$. This is related to our observation from Section 7.4.1 that the efficiency is worse when $f(z)$ is large. The relative performance of the three estimators is consistent across the two sample sizes under consideration. Overall, the jittering estimators are competitive with the benchmark estimator `liracine`. We found no evidence that adding artificial noise negatively affects the accuracy of the estimates. This confirms what was suggested by the estimator's asymptotic properties.

7.7 Conclusion

This article gave an in-depth analysis of the behavior of the jittering estimator. It was shown to have appealing large-sample properties and perform well on small samples.

Although our focus was on a particular instance of the class of jittering estimators, we also learned something about the class as a whole. Adding noise to discrete variables does not have a negative impact on estimation accuracy. This is true for both large samples (as confirmed by our asymptotic analysis) and small samples (as illustrated by simulations). More specifically, it allows for estimators that are optimal in terms of convergence rates and efficiency. It is likely that these findings generalize to more sophisticated density estimators or estimators of functionals of the density, such as regression functions.

Supplementary material

- https://github.com/tnagler/cctools: an R package implementing the jittering kernel density estimator and likelihood cross-validation for the bandwidths.

- https://gist.github.com/tnagler/786465cee2c774a844ff1846e7cdacd8: code for the simulation study in Section 7.6.

7.8 Proofs

7.8.1 Proof of Theorem 7.1

We first calculate the bias term. Using a change of variables, we get

$$\begin{aligned}\mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} &= \frac{1}{h_n^p b_n^q}\mathrm{E}\bigg\{K\bigg(\frac{\boldsymbol{Z}+\boldsymbol{\epsilon}-\boldsymbol{z}}{h_n}\bigg)K\bigg(\frac{\boldsymbol{X}-\boldsymbol{x}}{b_n}\bigg)\bigg\}\\ &= \frac{1}{h_n^p b_n^q}\int_{\mathbb{R}^{p+q}} K\bigg(\frac{\boldsymbol{s}-\boldsymbol{z}}{h_n}\bigg)K\bigg(\frac{\boldsymbol{t}-\boldsymbol{x}}{b_n}\bigg)f_\eta(\boldsymbol{s},\boldsymbol{t})d\boldsymbol{s}d\boldsymbol{t}\\ &= \int_{\mathbb{R}^{p+q}} K(\boldsymbol{u})K(\boldsymbol{v})f_\eta(\boldsymbol{z}+h_n\boldsymbol{u},\boldsymbol{x}+b_n\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}\end{aligned}$$

Since $\eta \in \mathcal{E}_{\gamma_1,\gamma_2}$, $f_\eta(\boldsymbol{z},\boldsymbol{x})$ is constant in a small neighborhood of $\boldsymbol{z}$. More specifically, it holds for all $(\boldsymbol{z},\boldsymbol{x}) \in \mathbb{Z}^q \times \mathbb{R}^q$, $0 \le \delta \le \min\{\gamma_1, 1-\gamma_2\}$ and $\boldsymbol{u} \in [-1,1]$ that $f_\eta(\boldsymbol{z}+\delta\boldsymbol{s},\boldsymbol{x}) = f(\boldsymbol{z},\boldsymbol{x})$. Furthermore, K is zero outside of $[-1,1]$. Applying this for $h_n \le \min\{\gamma_1, 1-\gamma_2\}$,

$$\begin{aligned}\mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} &= \int_{[-1,1]^{p+q}} K(\boldsymbol{u})K(\boldsymbol{v})f(\boldsymbol{z}+h_n\boldsymbol{u},\boldsymbol{x}+b_n\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}\\ &= \int_{[-1,1]^q} K(\boldsymbol{v})f(\boldsymbol{z},\boldsymbol{x}+b_n\boldsymbol{v})d\boldsymbol{v}. \end{aligned} \tag{7.8}$$

Recall the derivative notation from (7.7). An ℓ-th order Taylor expansion of f with mean-value remainder yields that

$$\begin{aligned}\mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} - f(\boldsymbol{z},\boldsymbol{x}) &= \sum_{1\le|\boldsymbol{a}|\le\ell}\frac{b_n^{|\boldsymbol{a}|}}{|\boldsymbol{a}|!}\int_{[-1,1]^q} K(\boldsymbol{v})\boldsymbol{v}^{\boldsymbol{a}}D_{\boldsymbol{x}}^{|\boldsymbol{a}|}f(\boldsymbol{z},\boldsymbol{x})d\boldsymbol{v}\\ &\quad+\sum_{|\boldsymbol{a}|=\ell+1}\frac{b_n^{\ell+1}}{(\ell+1)!}\int_{[-1,1]^q} K(\boldsymbol{v})\boldsymbol{v}^{\boldsymbol{a}}D_{\boldsymbol{x}}^{|\boldsymbol{a}|}f(\boldsymbol{z},\boldsymbol{x}+\tau_{\boldsymbol{a}}\boldsymbol{v})d\boldsymbol{v}\\ &= \frac{b_n^\ell}{\ell!}\sum_{j=1}^q\int_{[-1,1]} K(v_j)v_j^\ell\frac{\partial^\ell f(\boldsymbol{z},\boldsymbol{x})}{\partial x_j^\ell}dv_j,\\ &\quad+\sum_{|\boldsymbol{a}|=\ell+1}\frac{b_n^{\ell+1}}{(\ell+1)!}\int_{[-1,1]^q} K(\boldsymbol{v})\boldsymbol{v}^{\boldsymbol{a}}D_{\boldsymbol{x}}^{|\boldsymbol{a}|}f(\boldsymbol{z},\boldsymbol{x}+\tau_{\boldsymbol{a}}\boldsymbol{v})d\boldsymbol{v}\end{aligned}$$

for some $\tau_{\boldsymbol{a}} \in [0,1]$, where the second equality is due to K2. The second sum is $o(b_n^\ell)$ because all terms are bounded by A1 and K1, and $b_n \to 0$ as $n \to \infty$. In summary,

$$\mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} - f(\boldsymbol{z},\boldsymbol{x}) = \frac{b_n^\ell\sigma_\ell}{\ell!}\sum_{j=1}^q\frac{\partial^\ell f(\boldsymbol{z},\boldsymbol{x})}{\partial x_j^\ell} + o(b_n^\ell),$$

as claimed.

For the variance, we get

$$\begin{aligned}\mathrm{Var}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} &= \frac{1}{nh_n^{2p}b_n^{2q}}\mathrm{Var}\bigg\{K\bigg(\frac{\boldsymbol{Z}+\boldsymbol{\epsilon}-\boldsymbol{z}}{h_n}\bigg)K\bigg(\frac{\boldsymbol{X}-\boldsymbol{x}}{b_n}\bigg)\bigg\}\\ &= \frac{1}{n}\bigg[\frac{1}{h_n^{2p}b_n^{2q}}\mathrm{E}\bigg\{K\bigg(\frac{\boldsymbol{Z}+\boldsymbol{\epsilon}-\boldsymbol{z}}{h_n}\bigg)^2K\bigg(\frac{\boldsymbol{X}-\boldsymbol{x}}{b_n}\bigg)^2\bigg\}\\ &\qquad - \frac{1}{h_n^{2p}b_n^{2q}}\mathrm{E}\bigg\{K\bigg(\frac{\boldsymbol{Z}+\boldsymbol{\epsilon}-\boldsymbol{z}}{h_n}\bigg)K\bigg(\frac{\boldsymbol{X}-\boldsymbol{x}}{b_n}\bigg)\bigg\}^2\bigg].\end{aligned}$$

The second term in square brackets has already been calculated for the bias. Using similar arguments, we can show with a first-order tailor expansion

$$\begin{aligned}&\frac{1}{nh_n^pb_n^q}\int_{\mathbb{R}^{p+q}}K^2\bigg(\frac{\boldsymbol{s}-\boldsymbol{z}}{h_n}\bigg)K^2\bigg(\frac{\boldsymbol{t}-\boldsymbol{x}}{b_n}\bigg)f_\eta(\boldsymbol{s},\boldsymbol{t})d\boldsymbol{s}d\boldsymbol{t}\\ &= \kappa^p\int_{[-1,1]^q}K^2(\boldsymbol{v})f(\boldsymbol{z},\boldsymbol{x}+b_n\boldsymbol{v})d\boldsymbol{v}\\ &= \kappa^{p+q}f(\boldsymbol{z},\boldsymbol{x})+o(1),\end{aligned}$$

where $\kappa=\int_{-1}^1K^2(s)ds$. Together,

$$\begin{aligned}\mathrm{Var}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} &= \frac{\kappa^{p+q}}{nh_n^pb_n^q}f(\boldsymbol{z},\boldsymbol{x})+\frac{f^2(\boldsymbol{z},\boldsymbol{x})}{n}+o\bigg(\frac{1}{nh_n^qb_n^q}\bigg)\\ &= \frac{f(\boldsymbol{z},\boldsymbol{x})}{nb_n^q}\big\{h_n^{-p}\kappa^{p+q}-b_n^qf(\boldsymbol{z},\boldsymbol{x})\big\}+o\bigg(\frac{1}{nh_n^pb_n^q}\bigg).\end{aligned}$$

To show that the estimator is asymptotically normal, define

$$Y_{i,n}=\frac{1}{nb_n^q}K\bigg(\frac{\boldsymbol{Z}_i+\boldsymbol{\epsilon}_i-\boldsymbol{z}}{h_n}\bigg)K\bigg(\frac{\boldsymbol{X}_i-\boldsymbol{x}}{b_n}\bigg).$$

Then $\widetilde{f}(\boldsymbol{z},\boldsymbol{x})=\sum_{i=1}^nY_{i,n}$. which is asymptotically normal if the Lyapunov condition,

$$\bigg\{\sum_{i=1}^n\mathrm{E}\big(|Y_{i,n}|^3\big)\bigg\}^{1/3}\bigg\{\sum_{i=1}^n\mathrm{Var}(Y_{i,n})\bigg\}^{-1/2}\to 0,$$

is fulfilled. With arguments similar to the derivation of $\mathrm{Var}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\}$, we get $\mathrm{E}(|Y_{i,n}|^3)=O(n^{-1}h_n^{-2p}b_n^{-2q})$ and $\mathrm{Var}(Y_{i,n})=O(h_n^{-p}b_n^{-q})$. Thus,

$$\bigg\{\sum_{i=1}^n\mathrm{E}\big(|Y_{i,n}|^3\big)\bigg\}^{1/3}\bigg\{\sum_{i=1}^n\mathrm{Var}(Y_{i,n})\bigg\}^{-1/2}=O\big\{(nh_n^pb_n^q)^{-1/6}\big\},$$

which is $o(1)$ due to assumption A4.

7.8.2 Proof of Theorem 7.2

From the triangle inequality, we get the bound

$$\big|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x})\big| \leq \big|\mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} - f(\boldsymbol{z},\boldsymbol{x})\big| + \big|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - \mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\}\big|. \tag{7.9}$$

We start as in the proof of Theorem 7.1, but expand (7.8) as a Taylor polynomial of order $\ell - 2$. We can then show that for some $\tau \in [0,1]$,

$$\begin{aligned}
&\mathrm{E}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} - f(\boldsymbol{z},\boldsymbol{x}) \\
&= \frac{b_n^{\ell-1}}{(\ell-1)!}\sum_{j=1}^{q}\int_{[-1,1]} K(v_j)v_j^{\ell-1}\frac{\partial^{\ell-1}f(\boldsymbol{z},\boldsymbol{x}+\tau b_n\boldsymbol{v})}{\partial x_j^{\ell-1}}dv_j \\
&= \frac{b_n^{\ell-1}}{(\ell-1)!}\sum_{j=1}^{q}\int_{[-1,1]} K(v_j)v_j^{\ell-1}\bigg\{\frac{\partial^{\ell-1}f(\boldsymbol{z},\boldsymbol{x}+\tau b_n\boldsymbol{v})}{\partial x_j^{\ell-1}} - \frac{\partial^{\ell-1}f(\boldsymbol{z},\boldsymbol{x})}{\partial x_j^{\ell-1}}\bigg\}dv_j,
\end{aligned}$$

where the second equality holds because of K2. Using A1′, we get

$$\sup_{(\boldsymbol{z},\boldsymbol{x})\in\mathcal{S}}\big|\mathrm{E}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} - f(\boldsymbol{z},\boldsymbol{x})\big| \leq \frac{b_n^{\ell}L\tau}{(\ell-1)!}\sum_{j=1}^{q}\int_{[-1,1]}|K(v_j)||v_j|^{\ell}dv_j = O(b_n^{\ell}), \tag{7.10}$$

for a positive constant $L < \infty$. Furthermore,

$$\mathrm{E}\Big\{\big|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - \mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\}\big|^2\Big\} = \mathrm{Var}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} = O\big\{(nh_n^p b_n^q)^{-1}\big\},$$

as in Theorem 7.1. And since convergence in L^2 implies convergence in probability,

$$\big|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - \mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\}\big| = O_p\big\{(nh_n^p b_n^q)^{-1/2}\big\},$$

which, together with (7.10), proves (7.4).

Moreover, there is a positive constant $\overline{c}_1 < \infty$ such that almost surely

$$\lim_{n\to\infty}\sqrt{\frac{nh_n^p b_n^q}{\max\{\ln\ln n, \ln h_n^{-1}, \ln b_n^{-1}\}}}\sup_{\mathcal{S}}\big|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - \mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\}\big| \leq \overline{c}_1, \tag{7.11}$$

see Theorem 1 of Einmahl and Mason (2005). Combining (7.10) and (7.11) proves (7.5).

7.8.3 Proof of Theorem 7.3

Note that we can write

$$\mathrm{E}\big\{d^2(\widetilde{f},f)\big\} = \mathrm{E}\Big\{\sup_{\mathcal{S}'}\big|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - f(\boldsymbol{z},\boldsymbol{x})\big|^2\Big\},$$

where $\mathcal{S}' = \{(\boldsymbol{z}, \boldsymbol{x})\}$ for $d_{(\boldsymbol{z},\boldsymbol{x})}$ and $\mathcal{S}' = \mathcal{S}$ for $d_{\infty,\mathcal{S}}$. It holds

$$\begin{aligned}\frac{1}{2}\mathrm{E}\big\{d^2(\widetilde{f}, f)\big\} &\leq \sup_{\mathcal{S}'}\big|\mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\} - f(\boldsymbol{z},\boldsymbol{x})\big|^2 + \mathrm{E}\big[\sup_{\mathcal{S}'}\big|\widetilde{f}(\boldsymbol{z},\boldsymbol{x}) - \mathrm{E}\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\}\big|^2\big] \\ &= a_1 + a_2 \end{aligned} \tag{7.12}$$

Using arguments almost identical to (7.10), we obtain for some $\tau \in [0,1]$,

$$\sup_{(\boldsymbol{z},\boldsymbol{x})\in\mathcal{S}}\big|\mathrm{E}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} - f(\boldsymbol{z},\boldsymbol{x})\big| \leq \frac{b_n^\beta \lambda \tau^{\beta-r}}{r!}\sum_{j=1}^q \int_{[-1,1]} |K(v_j)||v_j|^\beta dv_j = b_n^\beta \overline{c}_2,$$

For bounding a_2, we need to consider the characteristics of scenarios (i)–(iv).

(i) We proceed as in the proof of Theorem 7.1 to get

$$a_2 = \mathrm{Var}\big\{\widetilde{f}(\boldsymbol{z},\boldsymbol{x})\big\} = \frac{\kappa^{p+q}}{nh_n^p b_n^q} f(\boldsymbol{z},\boldsymbol{x}) + \frac{f^2(\boldsymbol{z},\boldsymbol{x})}{n} + o\bigg(\frac{1}{nb_n^q}\bigg).$$

For $q \geq 1$, choosing $b_n \sim n^{-1/(2\beta+q)}$ yields

$$\limsup_{n\to\infty} n^{2\beta/(2\beta+q)} a_2 \leq \frac{\kappa^{p+q}}{h_0^p} f(\boldsymbol{z},\boldsymbol{x}) = \overline{c}_3 < \infty.$$

If $q = 0$, it holds $f \leq 1$, and we get

$$\limsup_{n\to\infty} na_2 \leq \frac{\kappa^{p+q}}{h_0^p} f(\boldsymbol{z}) + f^2(\boldsymbol{z}) \leq \frac{\kappa^{p+q}}{h_0^p} + 1 = \overline{c}_4 < \infty. \tag{7.13}$$

(ii) With $h_n \sim 1$ and $b_n \sim (n/\ln n)^{-1/(2\beta+q)}$ in (7.11), we get

$$\limsup_{n\to\infty} (n/\ln n)^{2\beta/(2\beta+q)} a_2 \leq \overline{c}_1,$$

(iii) Using (7.13) yields

$$\limsup_{n\to\infty} na_2 \leq \sum_{\boldsymbol{z}\in\mathcal{S}'} \limsup_{n\to\infty} \mathrm{E}\big[\big|\widetilde{f}_{\boldsymbol{Z}}(\boldsymbol{z}) - \mathrm{E}\{\widetilde{f}_{\boldsymbol{Z}}(\boldsymbol{z})\}\big|^2\big] \leq |\mathcal{S}'|\overline{c}_4 = \overline{c}_5 < \infty.$$

(iv) With $h_n \sim 1$ and $b_n = 1$ in (7.11), we get

$$\limsup_{n\to\infty} (n/\ln\ln n) a_2 \leq \overline{c}_1.$$

Setting $\overline{c} = 2(\overline{c}_1 + \overline{c}_2 + \overline{c}_3 + \overline{c}_4 + \overline{c}_5)$ concludes the proof.

7.8.4 Proof of Theorem 7.4

We start with lower bounds for (i) and (iii). Fix $\boldsymbol{z}_0 \in \mathbb{Z}^p$ and define

$$\mathcal{G}_1(\beta, \lambda) = \big\{ f \in \mathcal{H}(\beta, \lambda) \colon f(\boldsymbol{z}', \boldsymbol{x}) = 0 \text{ for } \boldsymbol{z}' \neq \boldsymbol{z}_0 \big\}.$$

This set contains all probability densities in $\mathcal{H}(\beta, \lambda)$ that correspond to a random vector $(\boldsymbol{Z}, \boldsymbol{X})$ with $\boldsymbol{Z} = \boldsymbol{z}_0$ almost surely. This is equivalent to the case where all variables are continuous. By definition, $\mathcal{G}_1(\beta, \lambda) \subset \mathcal{H}(\beta, \lambda)$ and, thus, $\mathcal{R}_n^*\{\mathcal{G}_1(\beta, \lambda), d\} \leq \mathcal{R}_n^*\{\mathcal{H}(\beta, \lambda), d\}$. The two rates in Theorem 7.4 (i) and (iii) then follow from Theorem 9 in (Ibragimov and Khas' minskii, 1983).

For (ii) and (iv), we can simply consider a parametric family of densities $\mathcal{G}_2$. This yields the classical lower bound $n^{-1/2}$ for estimating a finite dimensional parameter (see, e.g., Tsybakov, 2008, Chapter 2).

"Every answer raises new questions."

— Gildamere

8 Conclusion

8.1 Summary

The main theme of this thesis was to deepen our understanding and to extend the applicability of nonparametric vine copula estimators.

We compared and tested existing methods in simulations and established theoretical results regarding their asymptotic behavior. In particular, it was shown that simplified vine copula models allow for nonparametric density estimation at a rate equivalent to a two-dimensional problem, irrespective of the actual number of variables. This property is uncommon in nonparametric function estimation, where the curse of dimensionality is prevalent. But it is powerful, especially in light of the masses of data being collected: we can benefit from the increasing number of samples, but do not suffer from the increasing number of variables.

We showed how copulas can be used to learn various types of regression functions based on estimating equations. The asymptotic behavior of this technique is intimately linked to the behavior of the copula density estimator. This suggests that the simplifying assumption is equally suited to lift the curse of dimensionality in a regression context. We further discussed the jittering trick to make these methods applicable to discrete and mixed data types. By adding noise to the discrete variables, the learning problem can be transformed into a purely continuous one. And if we choose the noise distribution wisely, solutions of the two problems become equivalent. The jittering kernel density estimator was analyzed in-depth with the somewhat surprising result that adding noise does not impair the estimator's efficiency.

8.2 Outlook

The convenience of the simplifying assumption for nonparametric estimation was a key motivation for this thesis. In regression problems, additivity is another popular structural assumption that has a similar effect on the convergence rate, but is more thoroughly understood. It is important to know when assumptions are appropriate or detrimental. Hence, a deeper understanding of the simplifying assumption will be vital. Figuring out the similarities and differences between the simplifying assumption and additivity could be a first step.

An inevitable problem with structural assumptions is that they are usually not true. In statistics, there is always a trade off between the difficulty of the question we ask and the accuracy of the answer we get. The simplifying assumptions reduces the difficulty of the question. However, we may want to balance this trade off more flexibly. One possibility was briefly mentioned in Remark 4.2: we can weaken the simplifying assumption by allowing the pair-copulas to depend on a small number of conditioning variables. A more clever variant requires that the pair copulas depend only on (unknown) projections of the conditioning variables in low-dimensional subspaces. Finding inference methods for such models and understanding their properties opens a whole new playground with much to discover.

Another recurring topic was the potential of nonparametric vine copula estimators for machine learning problems (e.g., in the applications given in Section 4.6 and Section 6.4). Vine copula models have recently left their footprints in the machine learning community (Chen, 2016, Carrera et al., 2016, Sun et al., 2017, Tekumalla et al., 2017), but only in their parametric version. The nonparametric version would fit more naturally in the state-of-the-art methods in this field (e.g., random forests, neural networks, and support vector machines are essentially nonparametric). Two issues were obstructive to the wider use of nonparametric vine copulas. The first is the requirement for continuous data, which can be resolved by jittering. The second is the estimators' computational feasibility on large data sets. A naive implementation does not scale well, typically having quadratic complexity in the sample size n and dimension d.

The `vinecopulib` project aims to address these issues. It centers around a C++ library that provides interfaces to both R and Python (Nagler and Vatter, 2018a,c, Arabas et al., 2017). More carefully crafted algorithms and data structures plus a few computational tricks allow to bring the complexity down to $O(nd^2)$ and $O(d^2)$ in time and memory, respectively. Even better: if the model is truncated after some tree level, the models scale almost linearly in the dimension. With these obstacles out of the way, the road towards big data applications is well-paved. And yet there are many open methodological problems, like tuning parameter and model selection. I look forward to see what nonparametric vine copulas can achieve and which problems they may solve.

Bibliography

Aas, K. (2016). Pair-copula constructions for financial applications: A review. *Econometrics*, 4(4):43.

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

Acar, E. F., Craiu, R. V., and Yao, F. (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, 7:2822–2850.

Acar, E. F., Genest, C., and Nešlehová, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110(0):74–90. Special Issue on Copula Modeling and Dependence.

Ahmad, I. A. and Cerrito, P. B. (1994). Nonparametric estimation of joint discrete-continuous probability densities with applications. *Journal of Statistical Planning and Inference*, 41(3):349–364.

Aitchison, J. and Aitken, C. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420.

Aitken, C. G. G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):109–122.

Andrews, D. W. (1991). An empirical process central limit theorem for dependent non-identically distributed random variables. *Journal of Multivariate Analysis*, 38(2):187–203.

Antal, B. and Hajdu, A. (2014). An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems*, 60:20–27.

Arabas, S., Nagler, T., and Vatter, T. (2017). pyvinecopulib: High Performance Algorithms for Vine Copula Modeling in Python. Python library, URL: `https://github.com/vinecopulib/pyvinecopulib`.

Athey, S., Tibshirani, J., and Wager, S. (2017). Generalized Random Forests. *arXiv:1610.01271*.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268.

Bedford, T. and Cooke, R. M. (2002). Vines — a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068.

Bergsma, W. (2011). Nonparametric testing of conditional independence by means of the partial copula. *arXiv:1101.4607*.

Bock, R., Chilingarian, A., Gaug, M., Hakl, F., Hengstebeck, T., Jiina, M., Klaschka, J., Kotr, E., Savick, P., Towers, S., Vaiciulis, A., and Wittek, W. (2004). Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516:511–528.

Bolancé, C., Guillén, M., and Nielsen, J. P. (2008). Inverse beta transformation in kernel density estimation. *Statistics & Probability Letters*, 78(13):1757–1764.

Bouezmarni, T. and Rombouts, J. V. (2010). Nonparametric density estimation for multivariate bounded data. *Journal of Statistical Planning and Inference*, 140(1):139–152.

Bouyé, E. and Salmon, M. (2009). Dynamic copula quantile regressions and tail area dynamic dependence in Forex markets. *The European Journal of Finance*, 15(7-8):721–750.

Brechmann, E. C. and Schepsmeier, U. (2013). Modeling dependence with C- and D-vine copulas: The R package CDVine. *Journal of Statistical Software*, 52(3):1–27.

Cantelli, F. (1933). Sulla determinazione empirica della leggi di probabilità. *Giornale dell 'Istituto Italiano degli Attuari*, 4:92–99.

Carrera, D., Santana, R., and Lozano, J. A. (2016). Vine copula classifiers for the mind reading problem. *Progress in Artificial Intelligence*, 5(4):289–305.

Chacón, J. and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *TEST*, 19(2):375–398.

Charpentier, A., Fermanian, J.-D., and Scaillet, O. (2006). The estimation of copulas: Theory and practice. In Rank, J., editor, *Copulas: From theory to application in finance*. Risk Books.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2):131–145.

Chen, X., Koenker, R., and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *The Econometrics Journal*, 12(1):50–67.

Chen, Y. (2016). A copula-based supervised learning classification for continuous and discrete data. *Journal of Data Science*, 14(4):769–782.

Cooke, R. M., Joe, H., and Chang, B. (2015). Vine regression. *Resources for the Future - Discussion Paper*, 15(52):15–52.

Czado, C. (2010). Pair-copula constructions of multivariate copulas. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, Lecture Notes in Statistics, pages 93–109. Springer Berlin Heidelberg.

Czado, C., Jeske, S., and Hofmann, M. (2013). Selection strategies for regular vine copulae. *Journal of the French Statistical Society*, 154(1):74–191.

Darolles, S., Fan, Y., Florens, J. P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

De Backer, M., El Ghouch, A., and Van Keilegom, I. (2017). Semiparametric copula quantile regression for complete or censored data. *Electronic Journal of Statistics*, 11(1):1660–1698.

Dehling, H., Mikosch, T., and Sörensen, M. (2002). *Empirical Process Techniques for Dependent Data*. Springer Science & Business Media.

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1):40–57.

Dette, H., Van Hecke, R., and Volgushev, S. (2014). Some Comments on Copula-Based Regression. *Journal of the American Statistical Association*, 109(507):1319–1324.

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59(0):52 – 69.

Donsker, M. D. (1952). Justification and extension of doob's heuristic approach to the kolmogorov- smirnov theorems. *Ann. Math. Statist.*, 23(2):277–281.

Duong, T. (2014). *ks: Kernel smoothing*. R package version 1.9.3, URL: `http://CRAN.R-project.org/package=ks`.

Efromovich, S. (2010). Orthogonal series density estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):467–476.

Efromovich, S. (2011). Nonparametric estimation of the anisotropic probability density of mixed variables. *Journal of Multivariate Analysis*, 102(3):468 – 481.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 11(2):89–121. With comments and a rejoinder by the authors.

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403.

Elidan, G. (2013). Copulas in machine learning. In *Copulae in mathematical and quantitative finance*, pages 39–60. Springer.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Few, S. (2008). Solutions to the problem of over-plotting in graphs. *Visual Business Intelligence Newsletter.*

Fischer, M., Köck, C., Schlüter, S., and Weigert, F. (2009). An empirical analysis of multivariate copula models. *Quantitative Finance*, 9(7):839–854.

Gawronski, W. (1985). Strong laws for density estimators of bernstein type. *Periodica Mathematica Hungarica*, 16(1):23–43.

Geenens, G., Charpentier, A., and Paindaveine, D. (2017). Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli*, 23(3):1848–1873.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Genest, C., Masiello, E., and Tribouley, K. (2009). Estimating copula densities through wavelets. *Insurance: Mathematics and Economics*, 44:170–181.

Genest, C. and Neslehova, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(02):475–515.

Genest, C., Nelehov, J. G., and Rmillard, B. (2017). Asymptotic behavior of the empirical multilinear copula process under broad conditions. *Journal of Multivariate Analysis*, 159:82–110.

Gijbels, I. and Mielniczuk, J. (1990). Estimating the density of a copula function. *Communications in Statistics - Theory and Methods*, 19(2):445–464.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2015a). Estimation of a copula when a covariate affects only marginal distributions. *Scandinavian Journal of Statistics*, 42(4):1109–1126.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2015b). Partial and average copulas and association measures. *Electronic Journal of Statistics*, 9(2):2420–2474.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38:907–921.

Glivenko, V. (1933). Sulla determinazione empirica della leggi di probabilità. *Giornale dell 'Istituto Italiano degli Attuari*, 4:421–424.

Godambe, V. P. (1991). *Estimating Functions*, volume 7 of *Oxford Statistical Science Series*. Clarendon Press.

Hall, P. et al. (1983). Orthogonal series methods for both qualitative and quantitative data. *The Annals of Statistics*, 11(3):1004–1007.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026.

Han, P., Wang, L., and Song, P. X.-K. (2016). Doubly robust and locally efficient estimation with missing outcomes. *Statistica Sinica*, 26(2):691–719.

Han, Y., Jiao, J., and Weissman, T. (2015). Minimax estimation of discrete distributions under loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354.

Hansen, B. E. (2004). Nonparametric estimation of smooth conditional distributions. Technical report, Department of Economics, University of Wisconsin.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748.

Härdle, W., Janssen, P., and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *The Annals of Statistics*, 16:1428–1449.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Hobæk Haff, I., Aas, K., and Frigessi, A. (2010). On the simplified pair-copula construction — simply useful or too simplistic? *Journal of Multivariate Analysis*, 101(5):1296–1310.

Hobæk Haff, I. and Segers, J. (2015). Nonparametric estimation of pair-copula constructions with the empirical pair-copula. *Computational Statistics & Data Analysis*, 84:1–13.

Hoeffding, W. (1940). Massstabinvariante Korrelationstheorie. *Schriften des Mathematischen Seminars und des Instituts für Angewandte Mathematik der Universität Berlin*, 5:181–233.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based Boosting 2.0. *Journal of Machine Learning Research*, 11(Aug):2109–2113.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2016). *mboost: Model-Based Boosting*. R package version 2.8-1, URL: https://cran.r-project.org/package=mboost.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293.

Ibragimov, I. and Khas' minskii, R. (1983). Estimation of distribution density. *Journal of Soviet Mathematics*, 21(1):40–57.

Janssen, P., Swanepoel, J., and Veraverbeke, N. (2014). A note on the asymptotic behavior of the Bernstein estimator of the copula density. *Journal of Multivariate Analysis*, 124:480–487.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In Rschendorf, L., Schweizer, B., and Taylor, M. D., editors, *Distributions with fixed marginals and related topics*, volume 28 of *Lecture Notes — Monograph Series*, pages 120–141. Institute of Mathematical Statistics, Hayward, CA.

Joe, H. (2014a). *Dependence modeling with copulas.* CRC Press.

Joe, H. (2014b). *Dependence Modeling with Copulas.* Chapman & Hall/CRC.

Kauermann, G. and Schellhase, C. (2014). Flexible pair-copula estimation in D-vines using bivariate penalized splines. *Statistics and Computing*, 24(6):1081–1100.

Kauermann, G., Schellhase, C., and Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical b-splines. *Scandinavian Journal of Statistics*, 40:685–705.

Kie, J. G., Matthiopoulos, J., Fieberg, J., Powell, R. A., Cagnacci, F., Mitchell, M. S., Gaillard, J.-M., and Moorcroft, P. R. (2010). The home-range concept: are traditional estimators still relevant with modern telemetry technology? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1550):2221–2231.

Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51(6):2836–2850.

Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs.* Cambridge university press.

Kolev, N. and Paiva, D. (2009). Copula-based regression models: A survey. *Journal of Statistical Planning and Inference*, 139(11):3847–3856.

Kosorok, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis*, 84(2):299–318.

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer Science & Business Media.

Kraus, D. and Czado, C. (2017). D-vine copula based quantile regression. *Computational Statistics & Data Analysis*, 110:1–18.

Kunihiro, B., Ritei, S., and Masaaki, S. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664.

Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2):266–292.

Li, Q. and Racine, J. (2004). Cross-Validated Local Linear Nonparametric Regression. *Statistica Sinica*, 14(2):485–512.

Lichman, M. (2013). UCI machine learning repository, url: http://archive.ics.uci.edu/ml.

Loader, C. (1999). *Local regression and likelihood*. Springer New York.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618.

Lopez-Paz, D., Hernández-lobato, J. M., and Schölkopf, B. (2012). Semi-supervised domain adaptation with non-parametric copulas. In *Advances in neural information processing systems*, pages 665–673.

Lorentz, G. (1953). Bernstein polynomials. *Mathematical Expositions, no. 8, Univ. of Toronto Press, Toronto.*

Marron, J. S. (1994). Visual understanding of higher-order kernels. *Journal of Computational and Graphical Statistics*, 3(4):447–458.

Monahan, J. F. (2011). *Numerical methods of statistics*. Cambridge University Press.

Morales-Nápoles, O., Cooke, R., and Kurowicka, D. (2011). About the number of vines and regular vines on n nodes. In Kurowicka, D. and Joe, H., editors, *DEPENDENCE MODELING: Vine Copula Handbook*, number 7699 in World Scientific Books. World Scientific Publishing Co. Pte. Ltd.

Müller, D. and Czado, C. (2017a). Dependence modeling in ultra high dimensions with vine copulas and the graphical lasso. *arXiv:1709.05119*.

Müller, D. and Czado, C. (2017b). Representing sparse gaussian dags as sparse r-vines allowing for non-gaussian dependence. *Journal of Computational and Graphical Statistics*, to appear.

Müller, D. and Czado, C. (2018). Selection of sparse vine copulas in high dimensions with the lasso. *Statistics and Computing*, to appear.

Nagler, T. (2014). Kernel methods for vine copula estimation. Master's thesis, Technische Universität München.

Nagler, T. (2017). *kdevine: Multivariate Kernel Density Estimation with Vine Copulas.* R package version 0.4.1, URL: https://github.com/tnagler/kdevine.

Nagler, T. (2018a). Asymptotic analysis of the jittering kernel density estimator. *Mathematical Methods of Statistics*, 27:32–46.

Nagler, T. (2018b). A generic approach to nonparametric function estimation with mixed data. *Statistics & Probability Letters*, 137:326–330.

Nagler, T. (2018c). kdecopula: An R package for the kernel estimation of bivariate copula densities. *Journal of Statistical Software*, 84(7):1–22.

Nagler, T., Bumann, C., and Czado, C. (2018). Model selection in sparse high-dimensional vine copula models with application to portfolio risk. *arXiv:1801.09739*.

Nagler, T. and Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89.

Nagler, T., Schellhase, C., and Czado, C. (2017). Nonparametric estimation of simplified vine copula models: comparison of methods. *Dependence Modeling*, 5(1):99–120.

Nagler, T. and Vatter, T. (2018a). *rvinecopulib: High Performance Algorithms for Vine Copula Modeling in R.* R package version 0.2.8.1.0, URL: https://github.com/vinecopulib/rvinecopulib.

Nagler, T. and Vatter, T. (2018b). Solving estimating equations with copulas. *arXiv:1801.10576*.

Nagler, T. and Vatter, T. (2018c). vinecopulib: High Performance C++ Algorithms for Vine Copula Modeling. C++ library version 0.2.8, URL: https://github.com/vinecopulib/vinecopulib.

Nelsen, R. B. (2006). *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag New York, Inc.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Nikoloulopoulos, A. K. (2013). On the estimation of normal copula discrete regression models using the continuous extension and simulated likelihood. *Journal of Statistical Planning and Inference*, 143(11):1923–1937.

Noh, H., Ghouch, A. E., and Bouezmarni, T. (2013). Copula-Based Regression Estimation and Inference. *Journal of the American Statistical Association*, 108(502):676–688.

Noh, H., Ghouch, A. E., and Van Keilegom, I. (2015). Semiparametric Conditional Quantile Estimation Through Copula-Based Multivariate Models. *Journal of Business and Economic Statistics*, 33(2):167–178.

Oakes, D. and Ritz, J. (2000). Regression in a Bivariate Copula Model. *Biometrika*, 87(2):345–352.

Otneim, H. and Tjøstheim, D. (2017). The locally gaussian density estimator for multivariate data. *Statistics and Computing*, 27(6):1595–1616.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

Panagiotelis, A., Czado, C., Joe, H., and Stoeber, J. (2017). Model selection for discrete regular vine copulas. *Computational Statistics & Data Analysis*, 106:138 – 152.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient Bayesian Inference for Gaussian Copula Regression Models. *Biometrika*, 93(3):537–554.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130.

Racine, J. S. and Nie, Z. (2018). *crs: Categorical Regression Splines*. R package version 0.15-30, URL: https://CRAN.R-project.org/package=crs.

Rémillard, B., Nasri, B., and Bouezmarni, T. (2017). On copula-based conditional quantile estimators. *Statistics & Probability Letters*, 128:14 – 20.

Ripley, B. D. (1987). *Stochastic Simulation*. John Wiley & Sons, Inc.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rose, D. (2015). *Modeling and estimating multivariate dependence structures with the Bernstein copula*. PhD thesis, Ludwig-Maximilians Universität München.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Sain, S. R. and Scott, D. W. (1996). On locally adaptive density estimation. *Journal of the American Statistical Association*, 91(436):1525–1534.

Salvadori, G. and De Michele, C. (2007). On the use of copulas in hydrology: theory and practice. *Journal of Hydrologic Engineering*, 12(4):369–380.

Sancetta, A. and Satchell, S. (2004). The bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory*, pages 535–562.

Schallhorn, N., Kraus, D., Nagler, T., and Czado, C. (2017). D-vine quantile regression with discrete variables. *arXiv preprint arXiv:1705.08310*.

Scheffer, M. and Weiß, G. N. F. (2017). Smooth nonparametric Bernstein vine copulas. *Quantitative Finance*, 17(1):139–156.

Schellhase, C. (2016). *penRvine: Pair-Copula Estimation in R-Vines using Bivariate Penalized Splines*. R package version 0.2, URL: https://cran.r-project.org/web/packages/penRvine/.

Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., and Erhardt, T. (2018). *VineCopula: Statistical Inference of Vine Copulas*. R package version 2.1.4, URL: https://github.com/tnagler/VineCopula.

Schwartz, S. C. (1967). Estimation of probability density by an orthogonal series. *The Annals of Mathematical Statistics*, 38(4):1261–1265.

Scott, D. W. (2008). The curse of dimensionality and dimension reduction. In *Multivariate Density Estimation: Theory, Practice, and Visualization*, pages 195–217. John Wiley & Sons, Inc.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B (Methodological)*, 53(3):683–690.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1):177–184.

Simar, L., Zelenyuk, V., et al. (2011). To smooth or not to smooth? the case of discrete variables in nonparametric regressions. *Centre for Efficiency and Productivity Analysis: Working Paper Series.*

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistique de L'Université de Paris*, 8:229–231.

Song, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320.

Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, 65(1):60–68.

Spanhel, F. and Kurz, M. S. (2016). The partial copula: Properties and associated dependence measures. *Statistics & Probability Letters*, 119:76 – 83.

Spanhel, F. and Kurz, M. S. (2017). The partial vine copula: A dependence measure and approximation based on the simplifying assumption. *arXiv:1510.06971v2.*

Stöber, J. and Czado, C. (2012). Sampling pair copula constructions with applications to mathematical finance. In Mai, J.-F. and Scherer, M., editors, *Simulating Copulas: Stochastic Models, Sampling Algorithms, and Applications.* World Scientific Publishing Co, Singapore.

Stöber, J., Hong, H. G., Czado, C., and Ghosh, P. (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics & Data Analysis*, 88:28–39.

Stöber, J., Joe, H., and Czado, C. (2013). Simplified pair copula constructions — limitations and extensions. *Journal of Multivariate Analysis*, 119(0):101–118.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Stone, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. *Recent advances in statistics*, 5.

Sun, M., Konstantelos, I., and Strbac, G. (2017). C-vine copula mixture model for clustering of residential electrical load pattern data. *IEEE Transactions on Power Systems*, 32(3):2382–2393.

Tekumalla, L. S., Rajan, V., and Bhattacharyya, C. (2017). Vine copulas for mixed data : multi-view clustering for mixed data beyond meta-gaussian dependencies. *Machine Learning*, 106(9):1331–1357.

Tibshirani, J., Athey, S., Wager, S., and Wright, M. (2018). *grf: Generalized Random Forests (Beta).* R package version 0.9.5, URL: https://CRAN.R-project.org/package=grf.

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series in Statistics. Springer Science & Business Media.

Vatter, T. and Nagler, T. (2018). Generalized Additive Models for Pair-Copula Constructions. *Journal of Computational and Graphical Statistics, published online.*

Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics.

Walter, G. G. and Shen, X. (2000). *Wavelets and other orthogonal systems.* CRC press.

Wand, M. (1992). Error analysis for general multtvariate kernel estimators. *Journal of Nonparametric Statistics*, 2(1):1–15.

Watson, G. S. (1969). Density estimation by orthogonal series. *The Annals of Mathematical Statistics*, 40(4):1496–1498.

Weingessel, A. (2013). *quadprog: Functions to solve Quadratic Programming Problems.* R package version 1.5-5, URL: https://CRAN.R-project.org/package=quadprog.

Wood, S. N. (2006). *Generalized additive models.* Chapman and Hall/CRC.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301.

Yin, G. and Yuan, Y. (2009). Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(2):211–224.

Zur, R., Jiang, Y., and Metz, C. (2004). Comparison of two methods of adding jitter to artificial neural network training. *International Congress Series*, 1268:886 – 889. {CARS} 2004 - Computer Assisted Radiology and Surgery. Proceedings of the 18th International Congress and Exhibition.

Zwingmann, T. and Holzmann, H. (2017). Weak convergence of quantile and expectile processes under general assumptions. *arXiv:1706.04668*.