Causal Inference and Machine Learning

PhD Thesis

Lasse Petersen

April 2021

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

Lasse Petersen
lassepetersen@protonmail.com

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen, Denmark

Principal supervisor:	Professor Niels Richard Hansen University of Copenhagen
Assessment committee:	Professor Alexandra Carpentier Otto-von-Guericke-Universität Magdeburg
	Professor Stéphane Gaïffas Université Paris Diderot
	Professor Jonas Martin Peters University of Copenhagen
Date of submission:	May 3, 2021
ISBN:	978-87-7125-042-8

Preface

The present thesis is the culmination of three years of PhD studies that I conducted from 2018 to 2021 under the supervision of Professor Niels Richard Hansen at the Department of Mathematical Sciences of University of Copenhagen. My PhD project was supported by a research grant (13358) from VILLUM FONDEN.

Conducting research as a PhD student can be compared to finding your way through a large and obscure maze. In the beginning you have a clear point of entry and perhaps even an idea of where the exit could be located. However, immediately as you step into the maze and start exploring, you realize that most of the paths lead to dead ends, and a majority of your turns are the result of mere randomness rather than informed choices. After months or years of exploring you start questioning your strategy and doubt the existence of the exit that was promised by your supervisor. At some point you realize that the maze has multiple points of entry, and you also start redefining your perception of an exit. Given persistence, resilience and a good amount of shear luck, you might find a way through in the end.

I would like to thank my colleagues and office mates throughout the last three years — Martin, Søren, Nikolaj, Laura, Rune, Philip, Sebastian, Angélica and Gherardo among others — for making the exploration fun, providing feedback to my research and always being up for a cup of coffee. I would like to thank the senior members of the Copenhagen Causality Lab — Steffen Lauritzen, Jonas Peters and Niklas Pfister — for always taking their time to give academic advice and showing interest in my research projects. Also, I am grateful to Anders Tolver for mentoring me during my teaching periods.

I am particularly grateful for the support of my friends and family, who always encouraged me throughout the years, even though they didn't understand the mazes that I was exploring or even why I was exploring them in the first place. I am the most grateful for the support of Cecilie, especially during the last year of my PhD studies, for encouraging me when I felt like giving up, and cheering for me during the last sprint to the finish line.

A last and special thanks to Niels for his supervision and academic mentoring during my PhD studies. In particular for never showing me the shortest path, but rather inspiring me to explore the dead ends, redefine my points of entry and exit, and most importantly finding my own way through.

Lasse Petersen
Copenhagen, April 2021

Abstract

This thesis is concerned with the problem of performing causal graphical structure learning. The unifying approach to the problems studied throughout the thesis is the use of nonparametric machine learning techniques in order to relax distributional and functional assumptions on the data generating processes under consideration. The contribution of the thesis are four distinct manuscripts that are each concerned with different aspects of structure learning, which can be divided into two overall themes. The first theme is structure learning of graphical models for multivariate time series. Here we consider detecting the edges of a graphical model by posing regression models of the time series and reading the graph structure off the fitted models. The second theme is nonparametric hypothesis tests for constraint-based structure learning. Here we develop novel tests for conditional independence and conditional local independence. Our test for conditional independence is based on a generalized correlation in the partial copula, where we estimate nonparametric residuals using quantile regression. Our test for conditional local independence is based on a stochastic integral, which is a zero-mean local martingale under the hypothesis, and where the test statistic process requires nonparametric estimation of an intensity function and a predictable projection process. For both tests we utilize techniques from double machine learning to perform inference on a test statistic of a dependence measure in the presence of infinite dimensional nuisance parameters.

Resumé

Emnet for denne afhandling er strukturlæring for kausale grafiske modeller. Den overordnede tilgangsvinkel til problemerne, som vi studerer i afhandlingen, er brugen af ikke-parametriske machine learning teknikker til at lempe på fordelingsantagelser og funktionelle antagelser vedrørende de involverede data genererende processer. Afhandlingen bidrager med fire manuskripter, som beskræftiger sig med forskellige aspekter af strukturlæring. Disse bidrag kan inddeles i to overordnede temaer. Det første tema er strukturlæring af grafiske modeller for tidsrækker. Her antager vi at tidsrækken er genereret af en regressionsmodel, og vi løser strukturlæringsproblemet ved at træne regressionsmodellen til data og derefter aflæse strukturen fra den trænede model. Det andet tema er ikke-parametriske hypotesetest til brug i strukturlæringsalgoritmer. Her udvikler vi nye test for betinget uafhængighed og lokal uafhængighed. Vores test for betinget uafhængighed er baseret på en generaliseret korrelation i det partielle copula, hvor vi estimerer ikke-parametriske residualer ved brug af fraktilregression. Vores test for lokal uafhængighed er baseret på et stokastisk integral, som er en centreret, lokal martingal under hypotesen, og hvor vores teststatistik process kræver ikke-parametrisk estimation af en intensitetsfunktion og en forudsigelig projektionsprocess. Til begge test bruger vi teknikker fra dobbelt machine learning til at drage inferens om en teststatistik af et afhængighedsmål, som kræver yderligere estimation af en uendeligdimensionel støjparameter.

Contents

Chapter 1

Introduction

Graphical models [Lauritzen, 1996] are statistical models that encode the conditional independencies among random variables into graphs in order to give a compact representation of their dependence structure. Given a collection of random variables $X = (X_1, \ldots, X_p)$ assumed to be Markov with respect to a directed acyclic graph $\mathcal{D} = (V, E)$, the joint density factorizes according to the graph as a series of successive regressions $f(x) = \prod_{j \in V} f(x_j \mid \text{pa}_{\mathcal{D}}(x_j))$. Hence, by restricting X to be Markov with respect to $\mathcal{D}$ we achieve a decomposition of the problem of conducting statistical inference on the joint distribution into a series of subproblems of (hopefully) lower complexity. However, aside from the inferential benefits, this factorization also provides a useful intuition about the data generating mechanism of X, in the sense that it provides a recipe for simulation. We initialize by simulating the random variables represented by the source nodes of $\mathcal{D}$, and then continue simulating random variables given the value of their parents until we have reached the sink nodes of $\mathcal{D}$. This intuition, where a random variable is generated as an effect of causes, has popularized graphical models, in particular directed acyclic graphical models, as a language for reasoning about causality.

In the field of causal inference [Pearl, 2009, Spirtes et al., 2000, Peters et al., 2017] this simulation scheme is not merely seen as a way of building a probability distribution by successive conditional distributions, but more explicitly as a series of ordered structural assignments that describe the causal generating mechanism in terms of causes and their effects. Most importantly, the structural assignments and the associated causal graph provides a language for discussing interventions in a system — namely by replacing a structural assignment according to the intervention of interest. If we assume that the random variable $X = (X_1, \ldots, X_p)$ is causally generated according to a graphical model, one is able to use the graph structure to, for example, identify interventional distributions, determine the presence of confounding, choose valid adjustment sets and discuss mediation.

However, as with most statistical methodology, causal inference is a two-step procedure. For some applications the causal graph can be determined by expert knowledge, but for many problems the graph is (at least partially) unknown. Hence, the first step of a causal analysis is determining the causal graph, and secondly carrying out the inference of interest. This can be seen as a causal model selection, known as causal structure learning or causal discovery.

The main inspiration for the work presented in this thesis are problems arising in causal structure learning. In particular, scrutinizing the assumptions underlying the existing methodology, where the vast majority focuses on parametric models, either in terms of distributional assumptions or the functional form of the structural assignments. Here our primary motivation has been how to utilize nonparametric methods from machine learning to relax distributional and functional assumptions. In other words —

and perhaps put a bit boldly — how can we automate the process of going from data to graph with as few assumptions and as little human decision-making as possible?

1.1 Contributions and organization

Let us outline the contributions of this thesis and how we intend to present them. The contributions can be divided into four distinct parts. Firstly, in Chapters 2 to 4 we present the three published papers:

- Lasse Petersen. Sparse Learning in Gaussian Chain Graphs for State Space Models. In *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pages 332–343, Prague, Czech Republic, 11–14 Sep 2018. PMLR.

- Sebastian Weichwald, Martin E. Jakobsen, Phillip B. Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 27–36, Vancouver, CA, 08–14 Dec 2020. PMLR.

- Lasse Petersen and Niels Richard Hansen. Testing Conditional Independence via Quantile Regression Based Partial Copulas. *Journal of Machine Learning Research*, 22(70):1–47, 2021b.

Each of Chapters 2 to 4 contains a paper followed by a discussion of its content and directions of further research. In addition to these papers representing finished work, Chapters 5 and 6 contains the current status of ongoing work. In Chapter 5 we present the manuscript:

- Lasse Petersen and Niels Richard Hansen. Nonparametric conditional local independence testing. 2021a.

Finally, in Chapter 6 we present a software implementation on intensity estimation with recurrent neural networks that is being developed in connection with this manuscript, which we believe is of independent interest.

The remainder of this introduction will be used to motivate the individual problems studied throughout the thesis. We will not spend time on giving a formal introduction to the fields of graphical models, causal inference nor any specific machine learning technique. Each of the manuscripts listed above are self-contained in the sense that they introduce the concepts that are needed to understand them. Instead, we will focus on giving the reader motivation for the problems studied in the thesis.

1.2 Model selection by regression

Consider a response $Y \in \mathbb{R}$ and a set of covariates $X \in \mathbb{R}^p$ such that

$$Y = X^T \beta + \varepsilon \tag{1.1}$$

where $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. A popular method for performing simultaneous model selection and parameter estimation in the model (1.1) given a set of i.i.d. observations $(Y_i, X_i)_{i=1}^N$ is the lasso estimator

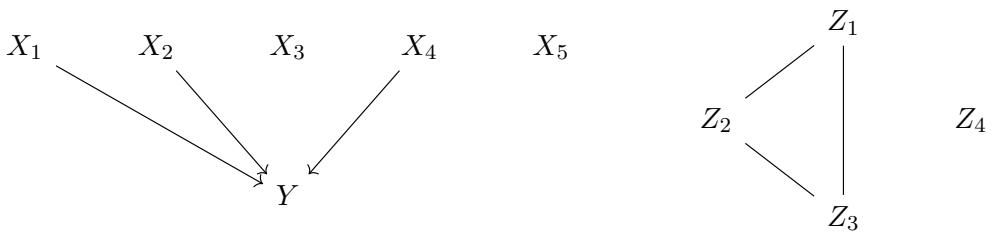

Figure 1.1: Left: An estimated directed acyclic graph, where the lasso estimator has chosen $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_4 \neq 0$ and $\hat{\beta}_3 = \hat{\beta}_5 = 0$. Right: An estimated concentration graph, where the graphical lasso estimator has chosen $\hat{\Theta}_{12}, \hat{\Theta}_{13}, \hat{\Theta}_{23} \neq 0$ and $\hat{\Theta}_{14} = \hat{\Theta}_{24} = \hat{\Theta}_{34} = 0$.

[Tibshirani, 1996], which performs maximum likelihood estimation with an $||\cdot||_1$-penalty by carrying out the optimization problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (Y_i - X_i^T \beta)^2 + \lambda ||\beta||_1 \right\} \tag{1.2}$$

where $\lambda \geq 0$ is a tuning parameter. The penalty function $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ induces a sparsity in the parameter vector β, which is responsible for the automatic model selection of the estimator, namely that $\hat{\beta}_j = 0$ is interpreted as the covariate X_j not being included in the model (1.1). This can be seen as a graphical model selection procedure in the directed acyclic graph, where Y is the sink node, $X_1, \ldots, X_p$ are source nodes, and there is a directed edge $X_j \to Y$ for each $j = 1, \ldots, p$ if and only if its regression parameter β_j is non-zero. See Figure 1.1 where we deliberately ignore a possible dependence between the covariates $X_1, \ldots, X_p$.

The idea of model selection by penalization of parameters was utilized for structure estimation of undirected Gaussian graphical models by Banerjee et al. [2008] and Friedman et al. [2008]. Here Z is assumed to follow a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where the conditional independencies are encoded in the concentration matrix $\Theta = \Sigma^{-1}$ such that $Z_i \perp\!\!\!\perp Z_j \mid Z_{\{i,j\}^c}$ if and only if $\Theta_{ij} = 0$. This means that Z is Markov with respect to its so-called concentration graph which has an undirected edge between Z_i and Z_j if and only if $\Theta_{ij} \neq 0$. This correspondence between conditional independence and zeros of the concentration matrix is exploited in the graphical lasso estimator, which performs maximum likelihood estimation of the concentration matrix Θ with $||\cdot||_1$-penalization given i.i.d. samples $Z_1, \ldots, Z_N$ by solving

$$\hat{\Theta} = \underset{\Theta \in \mathcal{S}_p^{++}}{\arg\min} \{ \operatorname{tr}(\Theta S) - \log\det\Theta + \lambda ||\Theta||_1 \} \tag{1.3}$$

where $S = \frac{1}{N} \sum_{i=1}^{N} Z_i Z_i^T$ is the empirical covariance matrix of the sample and $\mathcal{S}_p^{++}$ are the positive definite symmetric matrices. Again the penalty function $||\Theta||_1 = \sum_{i \neq j} |\Theta_{ij}|$ induces sparsity in the concentration matrix, which in turn gives a sparse estimated concentration graph. See Figure 1.1.

Now consider the situation where the response in model (1.1) is multidimensional $Y \in \mathbb{R}^q$ such that we can write the multivariate linear Gaussian model

$$Y = BX + \mathcal{E} \tag{1.4}$$

where $B \in \mathbb{R}^{q \times p}$ is a matrix of regression coefficients and $\mathcal{E} \sim \mathcal{N}(0, \Sigma)$. We could then be interested in examining the dependence structure between the components $Y_1, \ldots, Y_q$ by an undirected graphical

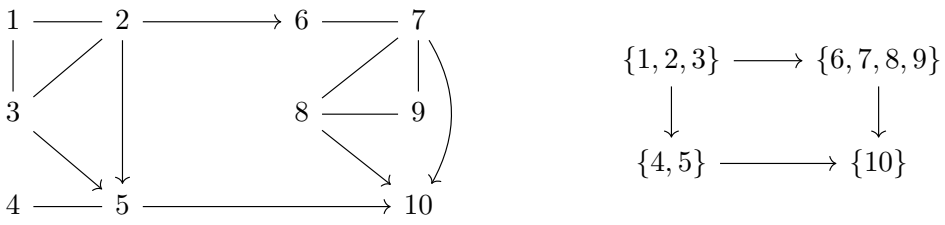

Figure 1.2: Left: a chain graph. Right: its directed acyclic graph of chain components.

model, since the symmetric nature of this graphical model type treats the variables as being "on equal footing", i.e., neither of the components are believed to be a response of the others. At the same time, we wish to be able to adjust this associative modeling of Y for the influence of the covariates X, and for this purpose a directed acyclic graphical model is natural due to the asymmetric nature where edges are directed from covariates to responses.

Graphs that contain both directed and undirected edges such that there are no semi-directed cycles, i.e., cycles where all directed edges point in the same directed, are called chain graphs. The chain components of a chain graph are the connected components after deleting all directed edges, and the chain components can themselves be considered nodes of a directed acyclic graph. See Figure 1.2. It turns out that there are multiple different Markov properties that one can associate with chain graphs, where the most common are the Lauritzen-Wermuth-Frydenberg (LWF) Markov property [Frydenberg, 1990, Lauritzen and Wermuth, 1989] and the Andersson-Madigan-Perlman (AMP) Markov property [Andersson et al., 2001].

Under a Gaussian chain graph model the two different Markov properties give different relations between sparsity of the model parameters and the presence of edges. The model (1.4) can be represented as a chain graph model, where the responses $Y = (Y_1, \dots, Y_q)$ and the covariates $X = (X_1, \dots, X_p)$ are chain components and the directed acyclic graph of chain components is given by $X \to Y$. Under the model (1.4) it holds that

$$Y \mid X = x \sim \mathcal{N}(Bx, \Theta^{-1})$$

and one can show that if (X, Y) is AMP Markov with respect to a chain graph, then the absence of directed edges from $X_1, \dots, X_p$ to $Y_1, \dots, Y_q$ gives zeros in the regression matrix B, while the absence of undirected edges among $Y_1, \dots, Y_q$ gives zeros of the concentration matrix $\Theta = \Sigma^{-1}$. One can also reparametrize the model in terms of its exponential family representation as

$$Y \mid X = x \sim \mathcal{N}(\Theta^{-1}\Lambda x, \Theta^{-1})$$

where $\Lambda = \Theta B$ and Θ are the canonical parameters. Under the LWF Markov property, one can show that the absence of directed edges from $X_1, \dots, X_p$ to $Y_1, \dots, Y_q$ implies zeros of Λ, and the absence of undirected edges between $Y_1, \dots, Y_q$ implies zeros of Θ (as with the AMP Markov property).

In the paper Petersen [2018], presented in Chapter 2, we adapt this relationship between chain graphical models and the multivariate linear Gaussian model to state space models. A state space model is a time series model, where there is an underlying continuous state space Markov chain (X_t) that is hidden, and we observe a noisy version (Y_t). The linear Gaussian state space model that we consider in the paper is given by

$$X_t \mid X_{t-1} = x_{t-1} \sim \mathcal{N}(Bx_{t-1}, \Sigma) \quad \text{and} \quad Y_t \mid X_t = x_t \sim \mathcal{N}(x_t, \rho^2 I)$$

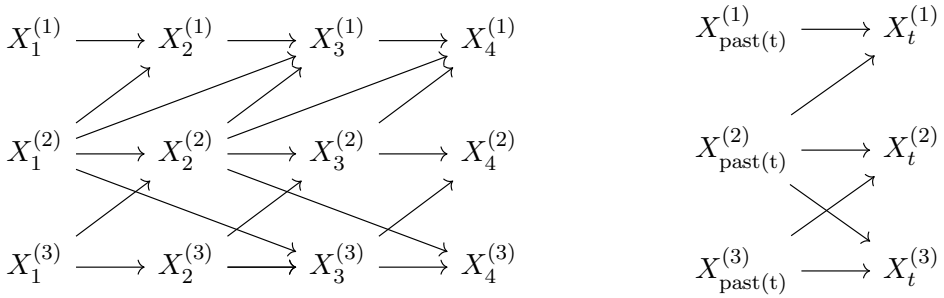

Figure 1.3: Left: a 3-dimensional time series. Right: its summary graph.

for $t = 1, \ldots, N$ where (X_t) is latent and (Y_t) is observed. The research questions underlying this paper are 1) how can we represent this model in a chain graphical framework under the AMP and LWF Markov properties, and 2) how can we perform simultaneous graphical model selection and parameter estimation by using penalized regression techniques. The contributions of the paper are two Expectation-Maximization-algorithms — one for each Markov interpretation of chain graphs — for estimating the parameters in the model, where penalization is applied in the M-step. Moreover, we demonstrate how these M-steps can be solved by using the lasso estimator (1.2) and the graphical lasso estimator (1.3).

This part of the thesis stand out from the rest, since we assume a completely parametric model in terms of distributional and functional assumptions. In return, we obtain a procedure for performing both the model selection *and* the parameter estimation simultaneously. Furthermore, the chain graph models give us a tool for assessing any conditional independence statement among the variables.

Sometimes such a fine-grained representation of the dependence structure of a time series is not required, but a more coarse representation is sufficient. Consider the time series $(X_t^{(1)}, X_t^{(2)}, X_t^{(3)})$ that is generated by the directed acyclic graph in Figure 1.3. In this multivariate time series there is dependence between several lags of the components, and such a graphical structure might be hard to learn from data without parametric assumptions. A more pragmatic approach to the problem is to assume time-homogeneity of the time series, and then represent the overall dependence structure between the time series in a so-called summary graph. For a p-dimensional time series, the summary graph is given by a $p \times p$ adjacency matrix A such that $A_{ij} = 1$ if and only if $X_t^{(i)}$ depends on $X_s^{(j)}$ for some $s < t$. See Figure 1.3. In other words, the summary graph describes the Granger-causality relations in the time series [Granger, 1969]. The summary graph is not a graphical model per se, since we cannot read of conditional independencies from the graph using a Markov property. Moreover, it contains strictly less information than the full directed acyclic graph.

In the paper Weichwald et al. [2020], presented in Chapter 3, we consider nonparametric estimation of summary graphs of time series. The origin of the paper was the Climate 4 Causality competition[1], in connection with the competition and demonstration track of the 2019 NeurIPS conference. The purpose was causal structure learning in time series, and the inferential target of the competition was summary graphs of time series. In the competition the teams were given training data from a number of different simulated time series, where the underlying data generating graphical structure was withheld. The teams could then perform queries to an online platform, which provided feedback on the performance of the prediction without revealing the true summary graph in order to avoid overfitting. In the end, the

[1] `https://causeme.uv.es/neurips2019`

teams were evaluated by the predictions on a test data set. We participated in the competition with a team of PhD students and post docs from the Copenhagen Causality Lab, and ended up winning the competition, and were given the opportunity to submit a paper on our findings.

Our approach to the problem of estimating a summary graph of a time series was the following. We assume that the time series (X_t) is time-homogeneous and generated from a differentiable function $F = (F_1, \ldots, F_p) : \mathbb{R}^{p \times L} \to \mathbb{R}^p$ such that

$$X_{t+1} = F(X_t, X_{t-1}, \ldots, X_{t-L+1}) + N_t$$

where (N_t) are i.i.d. zero-mean noise terms, and $L \geq 1$ is a pre-defined number of lags of the time series. For a fixed lag $1 \leq \ell \leq L$ we let

$$D_{ij}^{\ell}(x) = \partial_z F_{ji}(x_L, \ldots, x_{\ell+1}, z, x_{\ell-1}, \ldots, x_1)|_{z=x}$$

denote the partial derivative of the part of F that describes the functional dependency of $X_t^{(j)}$ on $X_{t-\ell}^{(i)}$. We then consider the parameter

$$\theta_{ij}^{\ell} = E|D_{ij}^{\ell}(X_{t-\ell}^{(i)})|$$

which quantifies the expected effect of $X_{t-\ell}^{(i)}$ on $X_t^{(j)}$ with respect to the distribution of $X_{t-\ell}^{(i)}$. If $X_t^{(j)}$ does not depend on $X_{t-\ell}^{(i)}$, then F_j is constant in the ℓ-lag of $X^{(i)}$, and therefore $D_{ij}^{\ell}(x)$ is zero. However, if there is a functional dependence, then we expect θ_{ij}^{ℓ} to be non-zero.

In Weichwald et al. [2020] we use this idea to approximate θ_{ij}^{ℓ} by using linear regression methods, to detect whether the (possibly non-linear) regression function F has regions where it is non-constant. If we detect a lag $1 \leq \ell \leq L$ such that θ_{ij}^{ℓ} is non-zero, then we let $A_{ij} = 1$ in the summary graph. Note that the parameters $(\theta_{ij}^{\ell})_{i,j=1,\ldots,p,\ell=1,\ldots,L}$ provide no information about the functional form of F, nor can it be used to predict $X^{(j)}$ from $X^{(i)}$. It is a purely exploratory model selection tool for determining the presence of dependence. However, we can use it as a justification to use linear regression methods to detect for the presence of non-linear functional relationships in time series.

The approaches to perform graphical structure learning of time series in Petersen [2018] and Weichwald et al. [2020] are quite different, but they are also achieving different goals under different assumptions. In Petersen [2018] we assume a fully parametric model and use penalized regression techniques to perform graphical model selection. In return, we estimate a proper chain graphical model that can be used to analyze the conditional independencies of the time series. The approach of Weichwald et al. [2020] is more heuristic, and the estimated summary graph does not contain as much information as the chain graph model, however, it does not make parametric nor distributional assumptions.

Note that there is nothing intrinsically causal about the models presented in this section nor the models presented in the rest of the thesis. In order to draw causal conclusions from statistical models, one needs to make causal assumptions such as the regressions of a directed acyclic graph being structural assignments, that there is no unmeasured confounding and that the system remains invariant under interventions. See Chapter 2 of Peters et al. [2017]. Nevertheless, the graphical structure learning is relevant during the statistical part of the analysis, while the causal conclusions based on this graph must be justified by making the relevant causal assumptions.

1.3 Constraint-based structure learning

Consider a target random variable Y and another set of random variables $X_1, \ldots, X_p$ that are potentially related to Y. In Section 1.2 we considered Y to be a response and $X_1, \ldots, X_p$ to be covariates such

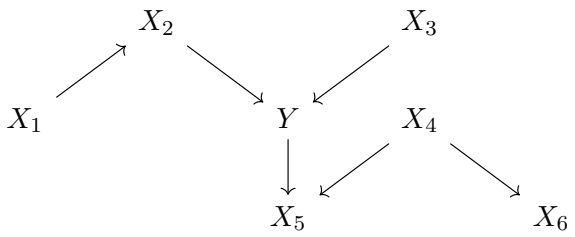

Figure 1.4: The Markov blanket of Y is given by $\text{mb}(Y) = \{X_2, X_3, X_4, X_5\}$.

that it was natural to consider $X_1, \ldots, X_p$ to be potential parents of Y in a directed acyclic graph or chain graph. However, in some situations it might be that $(Y, X_1, \ldots, X_p)$ is simply assumed Markov with respect to an acyclic directed graph, where Y might be the child of some variables and a parent of others.

The graph can then be used to guide a model selection on a regression model of Y on $X_1, \ldots, X_p$. In a graphical model the Markov blanket [Pearl, 1988] of a node Y is the minimal set of nodes M such that Y is independent of remaining variables given M. In the case of acyclic directed graphs, the Markov blanket of a node are the parents, the children and the parents of the children. See Figure 1.4. Consequently, we know that the conditional distribution of Y given $X_1, \ldots, X_p$ only depends on the Markov blanket, and so the graph structure gives a dimensionality reduction and model selection tool.

Conversely, the graph structure can be (partially) reconstructed from the conditional independencies of the distribution that it represents. Recall that a collection of random variables $X = (X_1, \ldots, X_p)$ are said to satisfy a global Markov property with respect to a graph $\mathcal{G}$ if

$$A \perp_{\mathcal{G}} B \mid C \quad \Longrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_C \tag{1.5}$$

where $A, B, C \subset \{1, \ldots, p\}$ and $\perp_{\mathcal{G}}$ denotes separation relative to the type of graph $\mathcal{G}$ [Lauritzen, 1996]. If $\mathcal{G}$ is a directed acyclic graph, then we usually consider d-separation [Pearl, 2009]. That X is globally Markov with respect to a graph can be quite an empty statement, since X is always globally Markov with respect to a fully connected undirected graph, where there are no separations. Thus, we need a requirement saying that the conditional independencies of X should be present in the graph $\mathcal{G}$ as separations. This reverse statement of the global Markov condition (1.5) is called faithfulness:

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad \Longrightarrow \quad A \perp_{\mathcal{G}} B \mid C. \tag{1.6}$$

The faithfulness assumption is the backbone of the causal structure learning paradigm known as constraint-based structure learning [Spirtes et al., 2000]. The basic intuition is that under faithfulness each separation corresponds to a conditional independence statement, and that the graph can be reconstructed from its separations. Therefore, we can learn the graph by making conditional independence queries to the distribution of interest. However, this is only partially true, since many graphs can entail the same separations.

Two graphs are called Markov equivalent, if they induce the same conditional independencies under the global Markov property. In the case of acyclic directed graphs, two graphs are Markov equivalent if they have the same skeleton and the same v-structures, i.e., subgraphs of the form $X \to Z \leftarrow Y$, where there is no edge between X and Y [Verma and Pearl, 1990]. The Markov equivalence class $\mathcal{M}(\mathcal{D})$ of a

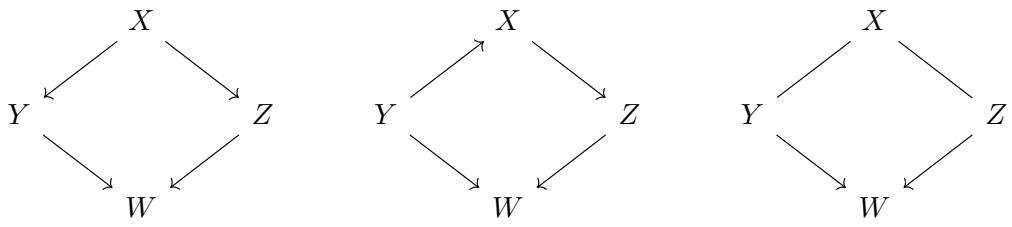

Figure 1.5: Left and middle: two Markov equivalent directed acyclic graphs. They have the same skeleton, and the same v-structures, namely $Y \to W \leftarrow Z$. Right: the CPDAG representing their Markov equivalence class. Note that we cannot orient $Y \to X$ and $Z \to X$ at the same time, since this would create a new v-structure.

directed acyclic graph $\mathcal{D}$ can be represented by a so-called completed partially directed acyclic graph $\mathcal{C}$ (CPDAG). The CPDAG $\mathcal{C}$ representing $\mathcal{M}(\mathcal{D})$ has a directed edge $X \to Y$ between two nodes, if this is present in all members of $\mathcal{M}(\mathcal{D})$, and it has an undirected edge $X - Y$ between two nodes, if there is a member of $\mathcal{M}(\mathcal{D})$ where $X \to Y$ and another where $X \leftarrow Y$. The undirected edge can be interpreted as an uncertainty about the orientation of that edge, since it can be oriented in either direction and still the graph has the same separations. Here one should be aware of not creating new v-structures when orienting undirected edges of the CPDAG. See Figure 1.5.

One of the most popular constraint-based structure learning algorithms is the PC-algorithm [Spirtes et al., 2000], which assumes that the distribution of X is Markov and faithful to a directed acyclic graph $\mathcal{D}$, and furthermore that causal sufficiency is satisfied, i.e., that there are no latent confounders. The algorithm is a two-step procedure. In the first step, the skeleton is reconstructed by making conditional independence queries to the distribution of X, which produces an undirected graph. In the second step, as many edges of the skeleton as possible are oriented using a set of orientation rules [Meek, 1995], and the final output is a CPDAG representing the Markov equivalence class of the true graph $\mathcal{D}$. For a full description of the PC-algorithm and the orientation rules see, e.g., Kalisch and Bühlmann [2007], who consider learning high-dimensional sparse directed acyclic graphs under a Gaussian assumption. Even though the PC-algorithm only outputs a CPDAG this can still be used to carry out causal inference by, e.g., the IDA algorithm [Maathuis et al., 2009], which provides bounds on interventional effects based on a CPDAG under a Gaussian assumption.

The PC-algorithm has several variations and extensions. For example the Fast-Causal-Inference (FCI) algorithm allows for latent and selection variables [Spirtes et al., 2000, Colombo et al., 2012], and estimates a so-called partial ancestral graph (PAG) that represent the Markov equivalence class of a directed acyclic graph with latent and selection variables. We will not go into further details about how these algorithms work, since this thesis does not contribute directly to this part of the literature. However, the motivation for the paper Petersen and Hansen [2021b], presented in Chapter 4, is the statistical inference part of constraint-based structure learning algorithms, where the skeleton estimation requires a hypothesis test for conditional independence.

Let $X, Y \in \mathbb{R}$ be random variables and $Z \in \mathbb{R}^p$ a random vector, and assume for simplicity that they have a joint density function with respect to Lebesgue measure on $\mathbb{R}^{p+2}$. In what follows we will use f to denote a generic density function. Recall that we say that X and Y are conditionally independent given Z if the density factorizes as

$$f(x, y \mid z) = f(x \mid z) f(y \mid z) \tag{1.7}$$

for almost all $x, y \in \mathbb{R}$ and $z \in \mathbb{R}^p$ with $f(z) > 0$. An equivalent characterization is that

$$f(y \mid x, z) = f(y \mid z) \tag{1.8}$$

for almost all $x, y \in \mathbb{R}$ and $z \in \mathbb{R}^p$ with $f(z) > 0$. These two definitions gives rise to (at least) three different strategies for testing conditional independence.

If we were to assume that Z was categorical, then the hypothesis (1.7) can be tested by stratification according to Z, i.e., testing the unconditional independence $X \perp\!\!\!\perp Y$ for each stratum of Z. This strategy can also be used when Z is continuous by applying a binning or unsupervised clustering to Z. However, in order to combine the tests we need to adjust the significance level for multiple testing, and if Z has many categories, then this will lead to a severe loss of power of the conditional independence test. Moreover, if Z is high-dimensional, then each stratum contains few or no samples, which makes the individual independence tests infeasible.

Another approach is to use the characterization (1.8), which treats Y as a response and X and Z as covariates, where the interpretation is that X is irrelevant for predicting Y in the presence of Z under conditional independence $X \perp\!\!\!\perp Y \mid Z$. Here we could propose conditional mean models $g_1(x, z) = E(Y \mid X = x, Z = z)$ and $g_2(z) = E(Y \mid Z = z)$ and test the fitted model $\hat{g}_1$ against $\hat{g}_2$ for the significance of X. However, in order to employ this strategy for testing conditional independence, we would need to propose a parametric model for the part of g_1 that depends on X in order to test the significance of parameters. Moreover, it is not given that the dependence of Y on X lies in the conditional mean. Here we could choose a more nonparametric approach by performing density estimation $\hat{f}(y \mid x, z)$ and compare this to $\hat{f}(y \mid z)$, but this is a difficult problem under high-dimensionality of Z due to the curse of dimensionality. A third complicating feature of this approach is the asymmetry in X and Y, since we are not guaranteed that using this test strategy will yield the same result with X and Y switched.

The third strategy is residualization. To explain the intuition of this approach, assume first that (X, Y, Z) is multivariate Gaussian and write

$$X = Z^T\beta_1 + \varepsilon_1 \quad \text{and} \quad Y = Z^T\beta_2 + \varepsilon_2$$

where $\beta_1, \beta_2 \in \mathbb{R}^p$ and $\varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $\varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2)$. Then the partial correlation of X and Y given Z can be defined as

$$\rho_{XY|Z} = \text{Corr}(\varepsilon_1, \varepsilon_2) = \text{Corr}(X - Z^T\beta_1, Y - Z^T\beta_2)$$

and $\rho_{XY|Z} = 0$ if and only if $X \perp\!\!\!\perp Y \mid Z$. To carry out the test we estimate the residuals $\hat{\varepsilon}_{1,i} = Z_i^T\hat{\beta}_1$ and $\hat{\varepsilon}_{2,i} = Z_i^T\hat{\beta}_2$ for a sample $(X_i, Y_i, Z_i)_{i=1}^n$, and then test for vanishing correlation in $(\hat{\varepsilon}_{1,i}, \hat{\varepsilon}_{2,i})_{i=1}^n$. The basic idea is that we perform a residualization by removing the marginal dependence of X on Z and Y on Z such that $\varepsilon_1 \perp\!\!\!\perp Z$ and $\varepsilon_2 \perp\!\!\!\perp Z$, and then look for remaining dependence between the residuals.

This idea was generalized by Shah and Peters [2020] who proposed to perform nonparametric conditional mean regressions of

$$f(z) = E(X \mid Z = z) \quad \text{and} \quad g(z) = E(Y \mid Z = z) \tag{1.9}$$

and test for vanishing correlation between the residuals $R_{1,i} = X_i - \hat{f}(Z_i)$ and $R_{2,i} = Y_i - \hat{f}(Z_i)$. They call their test the Generalised Covariance Measure (GCM), and their test is nonparametric in the sense that the conditional means (1.9) can be estimated using any machine learning technique, as long as they are consistently estimated with sufficiently fast rates [Shah and Peters, 2020, Theorem 6]. Note

that without a Gaussian assumption we only know that the residuals R_1 and R_2 are uncorrelated under conditional independence, but not necessarily independent.

Our motivation in Petersen and Hansen [2021b] was to develop a nonparametric conditional independence test similar to the GCM by using a different residualization approach. Let $F_{X|Z}(\cdot \mid z)$ and $F_{Y|Z}(\cdot \mid z)$ denote the conditional distribution functions of $X \mid Z = z$ and $Y \mid Z = z$ respectively. Then we can consider the transformations

$$U_1 = F_{X|Z}(X \mid Z) \quad \text{and} \quad U_2 = F_{Y|Z}(Y \mid Z).$$

The joint distribution of (U_1, U_2) is called the partial copula, and it was suggested as a device for testing conditional independence in Bergsma [2004, 2011], which comes from the fact that $U_1 \perp\!\!\!\perp U_2$ when $X \perp\!\!\!\perp Y \mid Z$ without any functional nor distributional assumptions. This can be considered a residualization approach since the variables U_1 and U_2 always satisfy $U_1 \perp\!\!\!\perp Z$ and $U_2 \perp\!\!\!\perp Z$ [Rosenblatt, 1952], and U_1 and U_2 were termed nonparametric residuals by Patra et al. [2016]. Analogously to the GCM, a partial copula based conditional independence test is a two-step procedure:

(1) Estimate the conditional distribution functions $\hat{F}_{X|Z}$ and $\hat{F}_{Y|Z}$ given a sample $(X_i, Y_i, Z_i)_{i=1}^n$.

(2) Compute estimated nonparametric residuals

$$\hat{U}_{1,i} = \hat{F}_{X|Z}(X_i \mid Z_i) \quad \text{and} \quad \hat{U}_{2,i} = \hat{F}_{Y|Z}(Y_i \mid Z_i)$$

for $i = 1, \ldots, n$ and test for independence in $(\hat{U}_{1,i}, \hat{U}_{2,i})_{i=1}^n$.

Our research goals were the following. Firstly, to propose a nonparametric estimator of the conditional distribution functions that can utilize nonparametric techniques from machine learning, and quantifying the rate of convergence of the estimator. Secondly, to control the nested estimation uncertainty involved with first estimating the nonparametric residuals and thereafter plugging them into a test for independence. In particular, to find a class of test statistics for independence such that we can transfer the convergence rates of the conditional distribution function estimator into asymptotic level and power properties of the conditional independence test.

Both the GCM and a partial copula based conditional independence test can be seen as examples of double machine learning procedures [Chernozhukov et al., 2018]. The target parameter of interest is a dependence measure between the residuals R_1 and R_2 or nonparametric residuals U_1 and U_2. However, in order to estimate this parameter, we first need to estimate an infinite dimensional nuisance parameter, which for the GCM are the conditionals means f and g, while for a partial copula based test it is the conditional distribution functions $F_{X|Z}$ and $F_{Y|Z}$. In the paradigm of double machine learning these nuisances are estimated using nonparametric machine learning techniques, and then the estimates are plugged back into an estimator of the target parameter. The main challenge is to control the nested estimation uncertainty error, which is further complicated by the fact that the nuisance parameters might be estimated at a slow rate due to the nonparametric estimation. In Chernozhukov et al. [2018] this problem is solved by the use of Neyman-orthogonal score functions, sample splitting and cross-fitting. While we do not formally cast our hypothesis test in Petersen and Hansen [2021b] in the language of score functions and Neyman-orthogonalization, we were still heavily inspired by the ideas and techniques of double machine learning.

1.4 Testing dependence in dynamical systems

Until this point we have considered the situation, where we have a real valued (and possibly multidimensional) target variable Y and a set of potential real valued covariates $X_1, \dots, X_p$. We will now shift our focus to the setup, where the response is a positive random variable τ, which is interpreted as an event time, e.g., time until death or failure of a machine. We can represent this response by a counting process $N_t = 1(\tau \leq t)$, which also gives the opportunity to effortlessly incorporate recurrent events, but for the sake of this introduction, we will stick with a single event. Instead of having real valued covariates, we will now consider a set of caglad $(X_t^d)_{d=1,\dots,p}$ covariate processes, i.e. processes that are left continuous with right limits.

In this framework, we would like to be able to describe the dependence structure between the stochastic processes $(N_t, X_t^d)_{d=1,\dots,p}$. One possibility is to choose a discretization $0 \leq t_1 < t_2 < \cdots$ and consider modeling the time series $(N_{t_j}, X_{t_j}^d)_{d=1,\dots,p, j\in\mathbb{N}}$ by a directed acyclic graph. However, this solution is unsatisfactory, since it depends on the discretization, and it does not truly capture the infinitesimal dependence structure in a continuous time dynamical system.

In order to study the dependence structure of stochastic processes, Schweder [1970] introduced the concept of conditional local independence. Let us describe it in the current context. Let $C \subset \{1, \dots, p\}$ and $b \in \{1, \dots, p\}$ with $b \notin C$. Denote by $\mathcal{F}$ the filtration generated by N and $X^C = (X^c)_{c\in C}$, and let $\mathcal{G}$ be the filtration generated by N and (X^b, X^C). Then we say that N is conditionally locally independent of X^b given X^C if the $\mathcal{F}$-compensator of N is also a $\mathcal{G}$-compensator of N. The intuition is that the history of X^b is irrelevant for predicting N in the presence of the history of X^C, and one can think of conditional local independence of stochastic processes as a continuous time version of Granger non-causality [Granger, 1969].

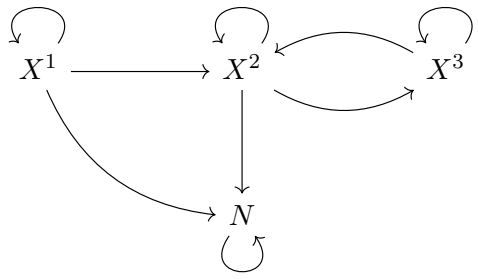

Figure 1.6: Example of local independence graph. If the system $(N_t, X_t^d)_{d\in\{1,2,3\}}$ satisfies the global Markov property defined by δ-separation with respect to this local independence graph, then X^1 is conditionally locally independent of X^2 without conditioning on additional processes, and (N_t) is conditionally locally independent of (X_t^3) given (X_t^1, X_t^2). The loops represent feedback from the individual stochastic processes with itself.

Didelez [2006, 2008] considered representing the conditional local independence relations using directed graphs, where she introduced a global Markov property based on δ-separation, which is a generalization of d-separation in directed acyclic graphs. Such graphs are called local independence graphs. See Figure 1.6. This concept was generalized by Mogensen et al. [2018] and Mogensen and Hansen [2020] who considered partially observed systems, where there can be processes acting as unmeasured confounders. They introduced a graphical representation by directed mixed graphs that have both directed and bi-

directed edges, and they introduced a global Markov property by μ-separation, which is a generalization of m-separation in directed acyclic mixed graphs. Mogensen et al. [2018] also developed a constraint-based structure learning algorithm of local independence graphs, where the skeleton reconstruction uses a conditional local independence oracle.

The motivation for the manuscript Petersen and Hansen [2021a], presented in Chapter 5, was to contribute to the statistical inference part of the structure learning of local independence graphs, by developing a nonparametric test for the hypothesis of conditional local independence. The manuscript presented here is the current status of ongoing work.

The setup of the manuscript is as follows. We still consider a counting process N, which we assume is adapted to a filtration $\mathcal{F}$, which may contain additional information on covariate processes. We let Z be an auxiliary caglad stochastic process, and we let $\mathcal{G}$ be the filtration generated by $\mathcal{F}$ and Z. Our goal is then to test whether N is conditionally locally independent of Z given $\mathcal{F}$. Let λ denote the $\mathcal{F}$-intensity of N, such that with $\Lambda_t = \int_0^t \lambda_s \mathrm{d}s$ being the $\mathcal{F}$-compensator of N, then the process M defined by

$$M_t := N_t - \Lambda_t,$$

is a zero-mean $\mathcal{F}$-martingale. Then the hypothesis can be equivalently stated by saying that N is conditionally locally independent of Z given $\mathcal{F}$ if the $\mathcal{F}$-martingale M is also a $\mathcal{G}$-martingale, i.e., the additional information augmented to $\mathcal{F}$ by the process Z is indeed irrelevant for N.

The initial research question of the manuscript Petersen and Hansen [2021a] was how to construct a hypothesis test for conditional local independence from the property of M being a martingale with respect to $\mathcal{G}$ under the hypothesis. Furthermore, a requirement of the proposed test was to make it nonparametric in the sense that it does not assume that the stochastic processes involved belong to a certain family of processes other than N being a counting process.

This last point is crucial since many model classes of stochastic processes are non-collapsable, i.e., the model classes are not closed under marginalization. For example, a multivariate Hawkes process is no longer a Hawkes process if we marginalize over one of the coordinates. Thus, we cannot consider the full dynamical system to be a multivariate Hawkes process, and then consider subsets of the coordinates without misspecifying the model. In Section 1.1 of the manuscript we give a more detailed example in terms of a Cox model with time varying covariates.

Our basic construction is as follows. First we let $\Pi_t = E(Z_t \mid F_{t-})$ be the predictable projection process of Z onto $\mathcal{F}$, such that the difference $Z - \Pi$ is $\mathcal{G}$-predictable since Z is assumed caglad. We then consider the process I defined as the stochastic integral

$$I_t = \int_0^t (Z_s - \Pi_s)\mathrm{d}M_s = \int_0^t (Z_s - \Pi_s)\mathrm{d}(N_s - \Lambda_s).$$

Under the hypothesis of conditional local independence, M is a zero-mean $\mathcal{G}$-martingale, so the process I is also a zero-mean $\mathcal{G}$-martingale. On the contrary, if conditional local independence is not satisfied, then M is *not* a $\mathcal{G}$-martingale, so I is *not* (necessarily) a $\mathcal{G}$-martingale, and we expect the process I to have a drift. Consequently, the function $t \mapsto \gamma_t = E(I_t)$ is identically zero under the hypothesis, and we expect it to be different from zero under the alternative. Our approach to perform the test is to use ideas from double machine learning [Chernozhukov et al., 2018], which was also a theme in Petersen and Hansen [2021b]. We consider nonparametric estimators of the nuisance parameters λ and Π, which we plug back into an estimator $\hat{I}^{(n)}$ of I. A novelty in our setup, compared to the usual double machine learning setup, is that both our nuisance parameters λ and Π and our target parameter $t \mapsto \gamma_t$ are infinite dimensional. Hence, in order to develop asymptotic theory we need the entire weak limit of our test statistic $\hat{I}^{(n)}$ as a stochastic process.

As part of our current work on this hypothesis test, we are also developing new nonparametric estimators of the intensity λ and the predictable projection Π. In Chapter 6 we present a software implementation for estimation of intensity functions based on recurrent neural networks, in the case where the filtration $\mathcal{F}$ is entirely generated by counting processes. We include this since we believe that the implementation is of independent interest and state-of-the-art.

Chapter 2

Sparse Learning in Chain Graphs

Lasse Petersen. Sparse Learning in Gaussian Chain Graphs for State Space Models. In *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pages 332–343, Prague, Czech Republic, 11–14 Sep 2018. PMLR.

Proceedings of Machine Learning Research vol 72, 332-343, 2018 PGM 2018

Sparse Learning in Gaussian Chain Graphs for State Space Models

Lasse Petersen LP@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen, Denmark

Abstract

The graphical lasso is a popular method for estimating the structure of undirected Gaussian graphical models from data by penalized maximum likelihood. This paper extends the idea of structure estimation of graphical models by penalized maximum likelihood to Gaussian chain graph models for state space models. First we show how the class of linear Gaussian state space models can be interpreted in the chain graph set-up under both the LWF and AMP Markov properties, and we demonstrate how sparsity of the chain graph structure relates to sparsity of the model parameters. Exploiting this relation we propose two different penalized maximum likelihood estimators for recovering the chain graph structure from data depending on the Markov interpretation at hand. We frame the penalized maximum likelihood problem in a missing data set-up and carry out estimation in each of the two cases using the EM algorithm. The common E-step is solved by smoothing, and we solve the two different M-steps by utilizing existing methods from high dimensional statistics and convex optimization.

Keywords: state space models; chain graph models; high dimensional statistics; sparse learning; EM algorithm; convex optimization.

1. Introduction

The *graphical lasso* (Banerjee et al., 2008; Friedman et al., 2008) produces a sparse estimate of the concentration matrix $\Theta = \Sigma^{-1}$ of a regular multivariate Gaussian distribution by penalized maximum likelihood from independent samples $x_1, \dots, x_N \sim \mathcal{N}(0, \Sigma)$. The estimator is given by

$$\hat{\Theta} = \underset{\Theta \in \mathcal{S}_p^{++}}{\arg\min} \{\text{tr}(\Theta S) - \log\det\Theta + ||W \circ \Theta||_1\} \tag{1}$$

where $\mathcal{S}_p^{++}$ are the real, symmetric, positive definite $p \times p$ matrices, $S = \frac{1}{N}\sum_{i=1}^N x_i x_i^T$ is the empirical covariance matrix, $||A||_1 = \sum_{ij}|A_{ij}|$ for a matrix A, $\circ$ denotes elementwise multiplication and W is a matrix of non-negative tuning parameters, e.g. $W = \lambda 1_{p\times p}$ for $\lambda \geq 0$. This is related to undirected Gaussian graphical models through the fact that if $X = (X_v)_{v\in V} \sim \mathcal{N}(0, \Sigma)$, then X is Markov w.r.t. its *concentration graph* $\mathcal{G} = (V, E)$ with edges $E = \{(u, v) \mid u \neq v, \Theta_{uv} \neq 0\}$. See Lauritzen (1996). Hence a sparse estimated concentation matrix $\hat{\Theta}$ gives rise to a sparse associated concentration graph $\hat{\mathcal{G}}$, which gives simple model interpretations.

This paper is concerned with exploiting the principle of penalized maximum likelihood for structure estimation in Gaussian chain graph models. This has previously been studied in a multivariate regression framework, $Y = BX + \varepsilon$ with $Y \in \mathbb{R}^d$, $X \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}(0, \Theta^{-1})$, which corresponds to a chain graph with two chain components — one for covariates and one for responses. Rothman et al. (2010) and Lin et al. (2016) consider sparse estimation of B and Θ in this set-up, which results in an estimated chain graph in the Andersson-Madigan-Perlman (AMP) Markov interpretation (Andersson et al., 2001). However, this estimator gives rise to a non-convex optimization problem for which there are no guarenties of convergence to a global optimum.

Lee and Liu (2012) and McCarter and Kim (2014) also consider multivariate regression, but in an exponential family parametrization with sparsity inducing penalties on the canonical parameters $\Theta = \Sigma^{-1}$ and $\Lambda = \Theta B$, which gives an estimated chain graph in the Lauritzen-Wermuth-Frydenberg (LWF) Markov interpretation (Frydenberg, 1990; Lauritzen and Wermuth, 1989). Here the estimator gives rise to a convex optimization problem as a result of the exponential family parametrization.

The purpose of this paper is to extend the existing methodology to Gaussian chain graphs for state space models. This extends the usage from multivariate regressions to time series data and allows for the case where the observations are corrupted by additive noise. State space models with sparsity inducing penalties has previously been considered in Noor et al. (2012) and Hasegawa et al. (2014), and our inference approach is similar to what is employed in their work. However, we further give the problem a graphical modeling framework, and we relate the penalization strategy to the chain graph Markov interpretation at hand. The main contributions of this paper are two EM algorithms for performing simultaneous parameter estimation and structure learning of a state space model and its associated chain graph under both the LWF and AMP Markov interpretation.

The paper is organized as follows. First we introduce linear Gaussian state space models and motivate the necessity of chain graphs for giving a detailed description of conditional independence for this model class. Next we give a brief introduction to chain graphs and their Markov properties, and we demonstrate how state space models can be viewed in a chain graph framework. We then develop an E-step and two different M-steps according to the Markov interpretation at hand.

2. Model Formulation

Let us begin by introducing our model class of interest.

Definition 1 *We define a **linear Gaussian state space model** (LGSSM) to be a pair of discrete time stochastic processes (X_t, Y_t) with X_t and Y_t both taking values in $\mathbb{R}^p$ such that*

$$X_t \mid X_{t-1} = x_{t-1} \sim \mathcal{N}(Bx_{t-1}, \Sigma) \quad \textit{and} \quad Y_t \mid X_t = x_t \sim \mathcal{N}(x_t, \rho^2 I_p) \tag{2}$$

for $t = 1, \ldots, N$ where X_0 is degenerate at $x_0 \in \mathbb{R}^p$. Here $\Sigma \in \mathcal{S}_p^{++}$ is a covariance matrix, $B \in \mathbb{R}^{p \times p}$ is a matrix of regression coefficients and $\rho^2 \geq 0$. The process (X_t) is assumed to be latent, while the process (Y_t) is observable.

From the distributional specification (2) we have the following factorization of the density of $(X_1, Y_1, \ldots, X_N, Y_N)$ conditional on the initial value X_0 of the latent process:

$$f(x_1, y_1, \ldots, x_N, y_N \mid x_0) = \prod_{t=1}^{N} f(x_t \mid x_{t-1}) \prod_{t=1}^{N} f(y_t \mid x_t). \tag{3}$$

Therefore the conditional independence structure of the process can be described by a directed acyclic graphical model as in Figure 1. From this DAG we can read of conditional independencies between the variables $X_1, Y_1, \ldots, X_N, Y_N$ by using, e.g., d-separation. However, the DAG does not give information about conditional independencies between single coordinates of the processes, e.g., whether there are conditional independencies among $X_{t,1}, \ldots, X_{t,p}$ when conditioning on $X_{t-1,1}, \ldots, X_{t-1,p}$. In order to provide such a detailed description of the conditional independence structure of the model, we will describe the model in a chain graph setting.

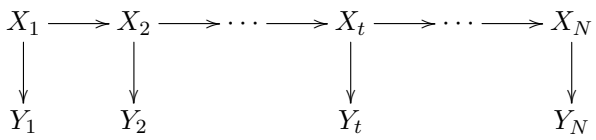

Figure 1: Directed acyclic graphical model for $(X_1, Y_1, \ldots, X_N, Y_N) \mid X_0 = x_0$.

3. Chain Graph Models

We now introduce the basic definition of a chain graph, the two different Markov properties that are usually associated with such graphs and their parametric restrictions in the Gaussian case.

Definition 2 *Let $\mathcal{G} = (V, E)$ be a graph where E is allowed to contain both undirected and directed edges. If $\mathcal{G}$ has no semi-directed cycles, i.e., cycles where all directed edges point in the same direction, then we call $\mathcal{G}$ a **chain graph**. Associated with a chain graph $\mathcal{G}$ we form the directed graph $\mathcal{D} = (\mathcal{T}, \mathcal{E})$, where $\mathcal{T}$ are the connected components of $\mathcal{G}$ after deleting all directed edges, and $\tau \to \tau' \in \mathcal{E}$ for $\tau, \tau' \in \mathcal{T}$ if there exists $u \in \tau$ and $u' \in \tau'$ such that $u \to u' \in E$. We call $\mathcal{D}$ the associated **graph of chain components** of $\mathcal{G}$ and note that the absence of semi-directed cycles in $\mathcal{G}$ ensures that $\mathcal{D}$ is a DAG.*

Chain graphs can be endowed with (at least) two different Markov interpretations, namely the AMP and LWF interpretation, which we will now describe. Let $Z = (Z_v)_{v \in V}$ be a collection of random variables indexed by the vertices of a chain graph $\mathcal{G} = (V, E)$. For a subset of vertices $A \subset V$, we denote by $\mathrm{pa}_{\mathcal{G}}(A)$ and $\mathrm{nb}_{\mathcal{G}}(A)$ the parents and neighbors of A relative to the graph $\mathcal{G}$. Consider the following four properties that Z can potentially fulfill w.r.t. $\mathcal{G}$:

C1) The distribution of Z satisfies the directed local Markov property w.r.t. $\mathcal{D}$.

C2) For each $\tau \in \mathcal{T}$, the distribution of $Z_\tau \mid Z_{\mathrm{pa}_{\mathcal{D}}(\tau)} = z_{\mathrm{pa}_{\mathcal{D}}(\tau)}$ is globally Markov w.r.t. $\mathcal{G}_\tau$.

C3) For each $\tau \in \mathcal{T}$ and $\sigma \subset \tau$ we have $\sigma \perp\!\!\!\perp (\mathrm{pa}_{\mathcal{D}}(\tau) \setminus \mathrm{pa}_{\mathcal{G}}(\sigma)) \mid \mathrm{pa}_{\mathcal{G}}(\sigma) \cup \mathrm{nb}_{\mathcal{G}}(\sigma)$.

C4) For each $\tau \in \mathcal{T}$ and $\sigma \subset \tau$ we have $\sigma \perp\!\!\!\perp (\mathrm{pa}_{\mathcal{D}}(\tau) \setminus \mathrm{pa}_{\mathcal{G}}(\sigma)) \mid \mathrm{pa}_{\mathcal{G}}(\sigma)$.

Here $A \perp\!\!\!\perp B \mid C$ is shorthand for $Z_A \perp\!\!\!\perp Z_B \mid Z_C$ for disjoint $A, B, C \subset V$. From these conditions, we can formulate the two Markov properties that we will associate with chain graphs.

Definition 3 *Let $Z = (Z_v)_{v \in V}$ and $\mathcal{G} = (V, E)$ be as above. If Z satisfies C1, C2 and C3, then we say it has the **LWF Markov property** w.r.t. $\mathcal{G}$. If Z satisfies C1, C2 and C4, we say it has the **AMP Markov property** w.r.t. $\mathcal{G}$.*

As with undirected Gaussian graphical models, the LWF and AMP Markov properties impose certain parametric restrictions for the Gaussian distribution. To describe these, assume further that Z follows a regular multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^V$. If Z satisfies C1, which is common to the LWF and AMP Markov property, then the distribution of Z is determined by the

conditional distributions of $Z_\tau \mid Z_{\mathrm{pa}_\mathcal{D}(\tau)} = z_{\mathrm{pa}_\mathcal{D}(\tau)}$ for $\tau \in \mathcal{T}$, because the density of Z factorizes according to $\mathcal{D}$. We can write each of these conditional distributions as a multivariate regression

$$Z_\tau \mid Z_{\mathrm{pa}_\mathcal{D}(\tau)} = z_{\mathrm{pa}_\mathcal{D}(\tau)} \sim \mathcal{N}\left(B_\tau z_{\mathrm{pa}_\mathcal{D}(\tau)}, \Sigma_\tau\right) \tag{4}$$

where B_τ is a matrix of regression coefficients. Alternatively we can introduce the parameters $K_\tau = \Sigma_\tau^{-1}$ and $\Lambda_\tau = K_\tau B_\tau$, which are the canonical parameters in an exponential family representation of the distribution, such that

$$Z_\tau \mid Z_{\mathrm{pa}_\mathcal{D}(\tau)} = z_{\mathrm{pa}_\mathcal{D}(\tau)} \sim \mathcal{N}\left(K_\tau^{-1}\Lambda_\tau z_{\mathrm{pa}_\mathcal{D}(\tau)}, K_\tau^{-1}\right). \tag{5}$$

We then have the following description of the parametric restriction implied by the LWF and AMP Markov properties. See Andersson et al. (2001) for details.

Proposition 4 *Let $Z = (Z_v)_{v\in V}$ and $\mathcal{G} = (V, E)$ be as above with Z satisfying C1. Then it holds for any chain component $\tau \in \mathcal{T}$ that*

i) if Z satisfies C2, then $(K_\tau)_{uv} = 0$ for all $u, v \in \tau$ with $u - v \notin E$,

ii) if Z satisfies C3, then $(\Lambda_\tau)_{uv} = 0$ for $u \in \tau$ and $v \in pa_\mathcal{D}(\tau) \setminus pa_\mathcal{G}(u)$,

iii) if Z satisfies C4, then $(B_\tau)_{uv} = 0$ for $u \in \tau$ and $v \in pa_\mathcal{D}(\tau) \setminus pa_\mathcal{G}(u)$.

In conclusion, the AMP Markov property encodes zeros in B_τ and K_τ of the regressions (4), while the LWF Markov property implies zeros in the canonical parameters Λ_τ and K_τ corresponding to the parametrization (5).

4. Chain Graphs for State Space Models

Let us now describe how LGSSMs can be viewed in the chain graph model setting. Naturally, we associate the vertices V of our chain graph $\mathcal{G} = (V, E)$ with the random variables in our LGSSM, and the chain components of our chain graph are the variables $X_1, Y_1, \ldots, X_N, Y_N$. Due to the property C1 and the factorization (3), the graph in Figure 1 must necessarily be the DAG of chain components of the chain graph. Now introduce the canonical parameters $\Theta = \Sigma^{-1}$ and $\Lambda = \Theta B$ such that we can reparametrize our model as an exponential family:

$$X_t \mid X_{t-1} = x_{t-1} \sim \mathcal{N}(\Theta^{-1}\Lambda x_{t-1}, \Theta^{-1}) \quad \text{and} \quad Y_t \mid X_t = x_t \sim \mathcal{N}(\rho^{-2}x_t, (\rho^{-2}I_p)^{-1}). \tag{6}$$

With inspiration from concentration graphs for undirected Gaussian graphical model and the properties i), ii) and iii) we can define the following chain graphs to associate with a LGSSM.

Definition 5 *Let (X_t, Y_t) follow a LGSSM with parameters Λ (B resp.) $\in \mathbb{R}^{p\times p}$, $\Theta \in \mathcal{S}_p^{++}$ and $\rho^2 \geq 0$. Then we define the **LWF (AMP resp.) concentration graph** $\mathcal{G} = (V, E)$ associated with these parameters to have $\mathcal{D}$ in Figure 1 as its associated DAG of chain components and edges E as follows. If $\Theta_{uv} \neq 0$, then we include the undirected edge $X_{t,u} - X_{t,v} \in E$ for each $t = 1, \ldots, N$. If $\rho^2 > 0$, then we let $X_{t,u} \to Y_{t,u} \in E$ for each $t = 1, \ldots, N$ and $u = 1, \ldots, p$. Lastly, if Λ_{uv} (B_{uv} resp.) $\neq 0$, then we include the directed edge $X_{t-1,v} \to X_{t,u} \in E$ for each $t = 2, \ldots, N$.*

Example 1 *Consider a LGSSM in the LWF interpretation with parameters*

$$\Lambda = \begin{pmatrix} * & 0 & 0 \\ * & 0 & * \\ 0 & * & 0 \end{pmatrix}, \quad \Theta = \begin{pmatrix} * & * & * \\ * & * & 0 \\ * & 0 & * \end{pmatrix} \quad \textit{and} \quad \rho^2 > 0 \tag{7}$$

where $*$ refers to some non-zero value. The subgraph of the LWF concentration graph $\mathcal{G}$ containing the latent process (X_t) can be seen in Figure 2. The undirected graphical structure within each

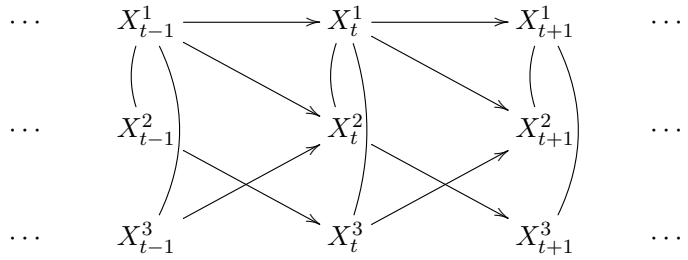

Figure 2: LWF concentration graph $\mathcal{G}$ associated with the parameters (7) restricted to (X_t).

chain component is constructed from Θ analogously with undirected Gaussian graphical models, while the directed edges between chain components are drawn using the zero-pattern of Λ. The full chain graph $\mathcal{G}$ simply has a directed edge from each coordinate of X_t to the corresponding coordinate of Y_t, since there is a non-trivial noise term in this particular example.

It is not hard to show that if (X_t, Y_t) follows a LGSSM with parameters Λ (B resp.) $\in \mathbb{R}^{p\times p}, \Theta \in \mathcal{S}_p^{++}$ and $\rho^2 \geq 0$, then $(X_1, Y_1, \ldots, X_N, Y_N) \mid X_0 = x_0$ will be LWF (AMP resp.) Markov with respect to its associated LWF (AMP resp.) concentration graph.

In conclusion, sparse parameters give rise to sparse chain graphs, which, in turn, give simple model interpretations through the properties C1-C4. In practise the parameters — and thus the graph structure — are unknown and must be estimated from data. However, we cannot expect to estimate entries of the parameters to be exactly zero, and so the need for sparse estimation procedures arise.

5. Sparse Learning via EM Algorithm

Given data $x_0, y_1, \ldots, y_N$ from a LGSSM we will carry out estimation by penalized maximum likelihood with sparsity inducing ℓ_1-penalties on Λ and Θ or B and Θ depending on the chain graph interpretation at hand.

We frame the estimation problem in a missing data set-up and perform inference using the EM algorithm. In this context, the complete data is $x_0, x_1, y_1, \ldots, x_N, y_N$, the missing data is $x_1, \ldots, x_N$ while the observed data is $x_0, y_1, \ldots, y_N$. We will here use a penalized version of the EM algorithm, where penalization is applied in the M-step (Green, 1990). First we derive the E-step, which involves computing the conditional expectation of the complete data log-likelihood given data and current EM estimate.

Proposition 6 *Let $x_0, y_1, \ldots, y_N$ be data from a LGSSM and $\theta^{(k)} = (\Lambda^{(k)}, \Theta^{(k)}, (\rho^2)^{(k)})$ the parameter estimate in the current EM iteration. Then the expected complete data log-likelihood given data and current parameter estimate is given by*

$$Q(\theta \mid \theta^{(k)}) = \log\det\Theta - \text{tr}(\Theta M_1) + 2\text{tr}(\Lambda M_2) - \text{tr}(\Lambda^T\Theta^{-1}\Lambda M_3) - p\log\rho^2 - \frac{1}{\rho^2}M_4$$

where $M_4 \in \mathbb{R}$ and $M_1, M_2, M_3 \in \mathbb{R}^{p\times p}$ depends on data and $\theta^{(k)}$ and are given by

$$\begin{aligned}
M_1 &= \frac{1}{N}\sum_{t=1}^{N} E(X_t X_t^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),\\
M_2 &= \frac{1}{N}\sum_{t=1}^{N} E(X_{t-1} X_t^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),\\
M_3 &= \frac{1}{N}\sum_{t=1}^{N} E(X_{t-1} X_{t-1}^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),\\
M_4 &= \frac{1}{N}\sum_{t=1}^{N} y_t y_t^T - 2y_t^T E(X_t \mid Y_{1:N} = y_{1:N}, \theta^{(k)}) + E(X_t^T X_t \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),
\end{aligned}$$

where we use the shorthand notation $Y_{1:N} = (Y_1, \ldots, Y_N)$.

Proof Due to the factorization (3) we can write the complete data log-likelihood as

$$\begin{aligned}
\ell(\Lambda, \Theta, \rho^2 \mid x, y) =& \frac{N}{2}\log\det\Theta - \frac{1}{2}\sum_{t=1}^{N}(x_t - \Theta^{-1}\Lambda x_{t-1})^T\Theta(x_t - \Theta^{-1}\Lambda x_{t-1})\\
&- \frac{Np}{2}\log\rho^2 - \frac{1}{2\rho^2}\sum_{t=1}^{N}(y_t - x_t)^T(y_t - x_t)
\end{aligned}$$

where we have ignored additive constants. By rescaling with $2/N$ and using the cyclic property of the matrix trace, i.e. $\text{tr}(AB) = \text{tr}(BA)$ for conformable matrices, we obtain

$$\begin{aligned}
\ell(\Lambda, \Theta, \rho^2 \mid x, y) \propto& \log\det\Theta - \frac{1}{N}\sum_{t=1}^{N}\text{tr}(\Theta x_t x_t^T) - 2\text{tr}(\Lambda x_{t-1}x_t^T) + \text{tr}(\Lambda^T\Theta^{-1}\Lambda x_{t-1}x_{t-1}^T)\\
&- p\log\rho^2 - \frac{1}{\rho^2}\frac{1}{N}\sum_{t=1}^{N} y_t^T y_t - 2y_t^T x_t + x_t^T x_t.
\end{aligned}$$

Taking conditional expectation with respect to data and current EM estimate and using linearity of the trace we obtain the wanted result. ■

Note that we have formulated the E-step in terms of the canonical parameters, but it is always possible to re-parametrize using $\Lambda = \Theta B$, and θ is simply a placeholder for the parameters under consideration in what follows. The conditional expectations that are needed when computing the

quantities $M_1, \ldots, M_4$ are the topic of smoothing in hidden Markov models, and one can use, e.g., the Rauch-Tung-Striebel smoother. See, e.g., Särkkä (2013) for details.

We now turn to the M-step. In the LWF interpretation we parametrize the expected complete data log-likelihood using canonical parameters and put sparsity inducing ℓ_1-penalties on Λ and Θ. The next EM iteration is produced by carrying out the optimization:

$$\hat{\theta}^{(k+1)} = \underset{(\Lambda,\Theta,\rho^2)\in\mathbb{R}^{p\times p}\times\mathcal{S}_p^{++}\times\mathbb{R}_+}{\arg\min} \left\{ f_1(\Lambda,\Theta) + f_3(\rho^2) + \lambda_1||\Lambda||_{1,\text{off}} + \lambda_2||\Theta||_{1,\text{off}} \right\}. \tag{8}$$

Here the function f_1 is the part of the negative expected complete data log-likelihood that depends on the parameters Λ and Θ,

$$f_1(\Lambda,\Theta) = \text{tr}(\Theta M_1) - 2\text{tr}(\Lambda M_2) + \text{tr}(\Lambda^T\Theta^{-1}\Lambda M_3) - \log\det\Theta,$$

and $f_3(\rho^2) = p\log\rho^2 + M_4/\rho^2$ is the part that depends on ρ^2. We let $||\Lambda||_{1,\text{off}} = \sum_{i\neq j}|\Lambda_{ij}|$, i.e. we choose not to penalize the diagonal. The numbers $\lambda_1, \lambda_2 \geq 0$ are tuning parameters determining the sparsity level of the estimates.

In the AMP interpretation we parametrize the expected complete data log-likelihood using the regression matrix and put ℓ_1-penalties on B and Θ. The M-step is then the optimization:

$$\hat{\theta}^{(k+1)} = \underset{(B,\Theta,\rho^2)\in\mathbb{R}^{p\times p}\times\mathcal{S}_p^{++}\times\mathbb{R}_+}{\arg\min} \left\{ f_2(B,\Theta) + f_3(\rho^2) + \lambda_1||B||_{1,\text{off}} + \lambda_2||\Theta||_{1,\text{off}} \right\}. \tag{9}$$

Here f_2 is the part that depends on the parameters B and Θ,

$$f_2(B,\Theta) = \text{tr}(\Theta M_1) - 2\text{tr}(\Theta B M_2) + \text{tr}(B^T\Theta B M_3) - \log\det\Theta,$$

and $f_3(\rho^2) = p\log\rho^2 + M_4/\rho^2$ as before.

Note that in both (8) and (9) there is variation independence between ρ^2 and the remaining parameters. Hence the optimization regarding ρ^2 can be performed separately, and is given by the conditional expectation of the empirical residual variance of the regression from X_t to Y_t:

$$(\hat{\rho}^2)^{(k+1)} = \frac{1}{Np}\sum_{t=1}^{N} E\left((Y_t - X_t)^T(Y_t - X_t) \mid Y_{1:N} = y_{1:N}, \theta^{(k)}\right) = \frac{1}{p}M_4. \tag{10}$$

Next we turn to the problem of performing the optimization in (8) regarding Λ and Θ. As we shall see, this task can be re-formulated into an equivalent optimization problem. Let the $2p \times 2p$ matrices $\mathbf{T}(\Lambda,\Theta)$ and $\mathbf{M}$ be given by

$$\mathbf{T}(\Lambda,\Theta) = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{12}^T & \mathbf{T}_{22} \end{pmatrix} = \begin{pmatrix} \Theta & -\Lambda \\ -\Lambda^T & I_p + \Lambda^T\Theta^{-1}\Lambda \end{pmatrix} \quad \text{and} \quad \mathbf{M} = \begin{pmatrix} M_1 & M_2^T \\ M_2 & M_3 \end{pmatrix}.$$

Consider the optimization problem

$$\min_{(\Lambda,\Theta)\in\mathbb{R}^{p\times p}\times\mathcal{S}_p^{++}} \{\text{tr}(\mathbf{T}(\Lambda,\Theta)\mathbf{M}) - \log\det\mathbf{T}(\Lambda,\Theta) + ||W\circ\mathbf{T}(\Lambda,\Theta)||_1\} \tag{11}$$

where W is the $2p \times 2p$ matrix

$$W = \begin{pmatrix} \lambda_2 E & \frac{1}{2}\lambda_1 E \\ \frac{1}{2}\lambda_1 E & 0_{p\times p} \end{pmatrix}$$

where E is the $p \times p$ matrix given by $E_{kk} = 0$ for $k = 1, \ldots, p$ and $E_{ij} = 1$ for $i \neq j$.

Proposition 7 *Performing the optimization in* (8) *regarding* Λ *and* Θ *is equivalent with solving the optimization problem* (11)*. Furthermore,* (11) *can be solved by applying graphical lasso* (1) *with optimization variable* $\boldsymbol{T}$*, empirical covariance matrix* $S = \boldsymbol{M}$ *and penalization matrix* W*, and afterwards extracting* $\hat{\Theta}^{(k+1)} = \hat{\boldsymbol{T}}_{11}$ *and* $\hat{\Lambda}^{(k+1)} = -\hat{\boldsymbol{T}}_{12}$*.*

Proof We show the equivalence by simply writing out the objective function of (11) and compare with (8). First we see that the trace term of (11) can be written

$$\begin{aligned}\operatorname{tr}(\mathbf{T}(\Lambda,\Theta)\mathbf{M}) &= \operatorname{tr}(\Theta M_1 - \Lambda M_2 - \Lambda^T M_2^T + M_3 + M_3\Lambda^T\Theta^{-1}\Lambda) \\ &= \operatorname{tr}(\Theta M_1) - 2\operatorname{tr}(\Lambda M_2) + \operatorname{tr}(\Lambda^T\Theta^{-1}\Lambda M_3) + \operatorname{tr}(M_3),\end{aligned}$$

and the determinant of $\mathbf{T}(\Lambda,\Theta)$ is equal to

$$\det \mathbf{T}(\Lambda,\Theta) = \det\Theta \det(I_p + \Lambda^T\Theta^{-1}\Lambda - (-\Lambda^T)\Theta^{-1}(-\Lambda)) = \det\Theta \det I_p = \det\Theta.$$

Lastly, we clearly have $||W \circ \mathbf{T}(\Lambda,\Theta)||_1 = \lambda_1||\Lambda||_{1,\text{off}} + \lambda_2||\Theta||_{1,\text{off}}$. Comparing to the part of the objective function of (8) concerning Λ and Θ, we see that the two objective function are equal up to the additive constant $\operatorname{tr}(M_3)$. This constant is computed using the current EM estimate $\theta^{(k)}$, but does not depend on the optimization variable θ, so it does not affect the optimization.

Let us argue that (11) can be solved by graphical lasso. First we note that the objective function has the correct functional form when comparing to (1). Secondly, we have $\mathbf{T}(\Lambda,\Theta) \in \mathcal{S}_p^{++}$ if and only if $\Theta \in \mathcal{S}_p^{++}$ so that the optimization domains are in fact equal. This is realized by using the Schur complement characterization of positive definiteness, i.e. that $\mathbf{T} \in \mathcal{S}_p^{++}$ if and only if $\mathbf{T}_{11} \in \mathcal{S}_p^{++}$ and $\mathbf{T}/\mathbf{T}_{11} \in \mathcal{S}_p^{++}$. Since $\mathbf{T}_{11} = \Theta$ and $\mathbf{T}/\mathbf{T}_{11} = I_p$ we conclude the wanted. ■

Input: Data $x_0, y_1, \ldots, y_N$, initial parameter values $\theta^{(0)}$ and $\lambda_1, \lambda_2 \geq 0$.
Output: Sparse parameter estimates $\hat{\Lambda}, \hat{\Theta}$ and $\hat{\rho}^2$.
begin
  $k \leftarrow 0$;
  repeat
    Compute $M_1, \ldots, M_4$ by smoothing using data and current estimate $\theta^{(k)}$;
    Update ρ^2 using (10);
    Solve the optimization (11) using graphical lasso to obtain $\hat{\mathbf{T}}$;
    Update Λ and Θ using estimate $\hat{\mathbf{T}}$ as described in Proposition 7;
    $k \leftarrow k + 1$;
  until *convergence criterion is met*;
  return $(\Lambda^{(k)}, \Theta^{(k)}, (\rho^2)^{(k)})$;
end

Algorithm 1: EM algorithm for sparse estimation of Λ, Θ and ρ^2 in a LGSSM.

We now turn to the optimization (9) regarding the parameters B and Θ. First note that the optimization problem (8) is convex, which is due to f_1 being the (expected) negative log-likelihood of an exponential family and that the ℓ_1-penalty is convex. However, the function f_2 is not jointly convex in B and Θ, but it is bi-convex, i.e. $B \mapsto f_2(B, \Theta_0)$ and $\Theta \mapsto f_2(B_0, \Theta)$ are convex for fixed $\Theta_0 \in \mathcal{S}_p^{++}$ and $B_0 \in \mathbb{R}^{p\times p}$ respectively. See Lee and Liu (2012) for a discussion. Therefore,

we will perform the optimization regarding B and Θ using an alternating convex search. More specifically, set $B_*^{(0)} := B^{(k)}$ and $\Theta_*^{(0)} := \Theta^{(k)}$, and then perform the optimizations

$$\Theta_*^{(i+1)} = \underset{\Theta \in \mathcal{S}_p^{++}}{\arg\min} \left\{ f_2(B_*^{(i)}, \Theta) + \lambda_2 ||\Theta||_{1,\text{off}} \right\}, \tag{12}$$

$$B_*^{(i+1)} = \underset{B \in \mathbb{R}^{p \times p}}{\arg\min} \left\{ f_2(B, \Theta_*^{(i+1)}) + \lambda_1 ||B||_{1,\text{off}} \right\} \tag{13}$$

for $i = 0, 1, \ldots$ until convergence. Then set $B^{(k+1)} := B_*^{(\infty)}$ and $\Theta^{(k+1)} := \Theta_*^{(\infty)}$ at convergence. The following proposition gives a way of solving (12) using existing methods.

Proposition 8 *The optimization* (12) *can be solved with graphical lasso with empirical covariance matrix* $S = M_1 - 2B_*^{(i)} M_2 + B_*^{(i)} M_3 (B_*^{(i)})^T$ *and penalty matrix* $W = \lambda_2 E$ *where* E *is as before.*

Proof We observe that

$$\begin{aligned} f_2(B_*^{(i)}, \Theta) &= \text{tr}(\Theta M_1) - 2\text{tr}(\Theta B_*^{(i)} M_2) + \text{tr}((B_*^{(i)})^T \Theta B_*^{(i)} M_3) - \log\det\Theta \\ &= \text{tr}(\Theta(M_1 - 2B_*^{(i)} M_2 + B_*^{(i)} M_3 (B_*^{(i)})^T)) - \log\det\Theta \end{aligned}$$

such that the objective function of (12) matches that of the graphical lasso problem (1) with $S = M_1 - 2B_*^{(i)} M_2 + B_*^{(i)} M_3 (B_*^{(i)})^T$. Note that this is a valid empirical covariance matrix since in fact

$$S = E\left(\frac{1}{N} \sum_{t=1}^{N} (X_t - B_*^{(i)} X_{t-1})(X_t - B_*^{(i)} X_{t-1})^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)} \right),$$

i.e. S is the conditional expectation of the empirical covariance of the fitted residuals for the regression from X_{t-1} to X_t given data and current EM estimate. ■

Just as (12) turned out to be solvable by applying graphical lasso, also (13) can be solved by existing methods, namely the lasso estimator (Tibshirani, 1996). The lasso estimates $\beta \in \mathbb{R}^p$ in the general linear model $y = A\beta + \varepsilon$, where A is a $N \times p$ design matrix and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$, by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2N} (y - A\beta)^T (y - A\beta) + ||\lambda \circ \beta||_1 \right\} \tag{14}$$

with $\lambda \in \mathbb{R}^p$ a vector of non-negative tuning parameters.

Proposition 9 *The optimization* (13) *can be solved by applying lasso regression in the following way. Let* $\hat{\beta}$ *be the result of a lasso regression with design matrix* A *and response vector* y *given by*

$$A = \sqrt{2N}(M_3 \otimes \Theta_*^{(i+1)})^{1/2} \quad \textit{and} \quad y = \sqrt{2N}(M_3 \otimes \Theta_*^{(i+1)})^{1/2} \textit{vec}(M_2^T M_3^{-1}),$$

and tuning parameter $\lambda = \lambda_1 R$ *where* $R \in \mathbb{R}^{p^2}$ *with* $R_1 = R_{p+2} = R_{2p+3} = \cdots = R_{p^2} = 0$ *and all other entries are* 1. *Here* $\otimes$ *denotes the Kronecker product, vec denotes vectorization of matrices and* $C^{1/2}$ *denotes the square root of a matrix* C. *Then* $B_*^{(i+1)}$ *is given by setting* $\textit{vec}(B_*^{(i+1)}) := \hat{\beta}$.

Proof Writing out the lasso objective function yields

$$\frac{1}{2N}(y - A\beta)^T(y - A\beta) + \lambda_1||\beta||_1 = \text{const.} + \frac{1}{2N}\beta^T A^T A\beta - \frac{1}{N}y^T A\beta + \lambda||\beta||_1, \qquad (15)$$

while writing out the objective function of (13) gives

$$f_2(B, \Theta_*^{(i+1)}) + \lambda_1||B||_{1,\text{off}} = \text{const.} + \text{tr}(B^T\Theta_*^{(i+1)}BM_3) - 2\text{tr}(\Theta_*^{(i+1)}BM_2) + \lambda_1||B||_{1,\text{off}}.$$

First note the useful relation between the matrix trace and the kronecker product and vectorization of matrices, $\text{tr}(ABCD) = \text{vec}(A^T)^T(D^T \otimes B)\text{vec}(C)$, where A, B, C and D are conformable matrices. Using this relation we can write

$$\text{tr}(B^T\Theta_*^{(i)}BM_3) = \text{vec}(B)^T(M_3 \otimes \Theta_*^{(i+1)})\text{vec}(B)$$

and also

$$2\text{tr}(\Theta_*^{(i)}BM_2) = 2\text{tr}(M_3^{-1}M_2\Theta_*^{(i+1)}BM_3) = 2\text{vec}(M_2^T M_3^{-1})^T(M_3 \otimes \Theta_*^{(i+1)})\text{vec}(B).$$

Letting A and y be as in the proposition and pluggin into the lasso objective function (15) we recover the objective function of (13) written in terms of vec and $\otimes$ as proposed. ■

Input: Data $x_0, y_1, \ldots, y_N$, initial parameter values $\theta^{(0)}$ and $\lambda_1, \lambda_2 \geq 0$.
Output: Sparse parameter estimates $\hat{B}, \hat{\Theta}$ and $\hat{\rho}^2$.
begin
  $k \leftarrow 0$;
  repeat
    Compute $M_1, \ldots, M_4$ by smoothing using data and current estimate $\theta^{(k)}$;
    Update ρ^2 using (10);
    Set $B_*^{(0)} \leftarrow B^{(k)}$, $\Theta_*^{(0)} \leftarrow B^{(k)}$ and $i \leftarrow 0$;
    repeat
      Update $\Theta_*^{(i+1)}$ using Proposition 8;
      Update $B_*^{(i+1)}$ using Proposition 9;
      $i \leftarrow i + 1$;
    until *convergence criterion is met*;
    Set $B^{(k+1)} \leftarrow B_*^{(\infty)}$, $\Theta^{(k+1)} \leftarrow \Theta_*^{(\infty)}$ and $k \leftarrow k + 1$;
  until *convergence criterion is met*;
  return $(B^{(k)}, \Theta^{(k)}, (\rho^2)^{(k)})$;
end

Algorithm 2: EM algorithm for sparse estimation of B, Θ and ρ^2 in a LGSSM.

6. Simulations

In this section we evaluate convergence of our proposed algorithms by means of simulation. For the case $p = 40$ and $N = 200$ we simulate valid true parameters for the LWF and AMP model such

that each of the matrices Θ, Λ and B has 75% zero entries, and ρ^2 is chosen to be 0.1 times the average of the diagonal of $\Sigma = \Theta^{-1}$. We then simulate 100 independent data set from each of the two LGSSMs and perform estimation using Algorithm 1 for LWF-data and Algorithm 2 for AMP-data. For tuning parameters λ_1, λ_2 we choose values ad hoc that do not produce neither completely sparse nor completely dense solutions. For each variable, say Θ, we track the relative difference from one EM iteration to the next by computing $\mathrm{RD}_\Theta(k) := ||\Theta^{(k)} - \Theta^{(k-1)}||_F \cdot ||\Theta^{(k-1)}||_F^{-1}$, where $||\cdot||_F$ is the Frobenious norm. The results can be seen in Figure 3. We observe that the relative

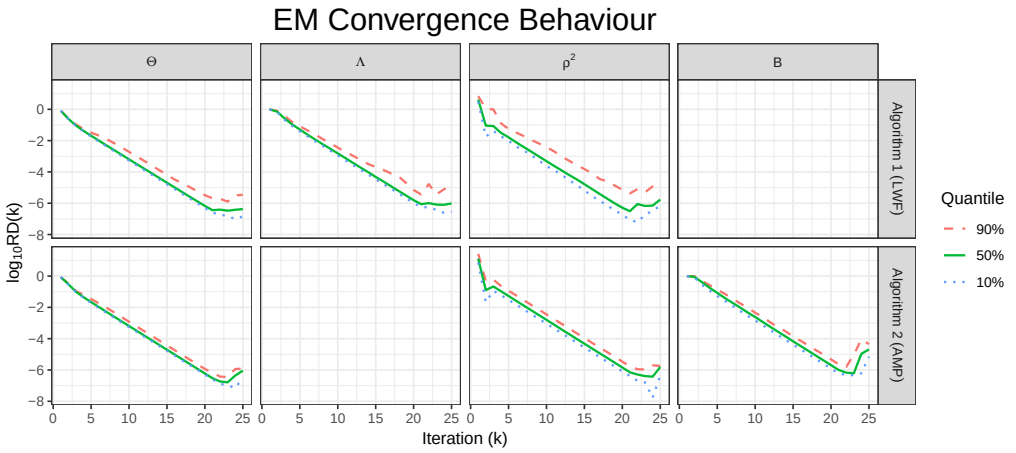

Figure 3: Selected quantiles of $\log_{10} \mathrm{RD}(k)$ computed for each variable at each iteration k based on 100 runs of Algorithm 1 and Algorithm 2 on simulated data with fixed true parameters. The algorithms were terminated when each relative difference dropped below 10^{-6}.

difference decreases approximately linearly on $\log_{10}$-scale. Moreover, the convergence behaviour is stable over the 100 runs. On average Algorithm 1 took 22.23 iterations before convergence, while Algorithms 2 needed 22.61 iterations on average before convergence. On average Algorithm 2 took 11.23 times longer than Algorithm 1 before convergence.

7. Conclusion

The purpose of this paper was to give a chain graph model framework for linear Gaussian state space model and develop algorithms for performing parameter estimation and structure learning from empirical data. We have proposed two different EM algorithms for performing this task depending on the chain graph interpretation (LWF or AMP) at hand, and we have justified convergence of the algorithms empirically through simulation. Next steps include developing methods for choosing the tuning parameters of the algorithms. This will enable us to consider edge-recovery properties of the algorithms and, moreover, make the algorithms useful in real world applications of the models.

Acknowledgments

This work was supported by a research grant (13358) from VILLUM FONDEN.

References

S. A. Andersson, D. Madigan, and M. D. Perlman. Alternative Markov properties for chain graphs. *Scand. J. Statist.*, 28(1):33–85, 2001. ISSN 0303-6898.

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

M. Frydenberg. The chain graph Markov property. *Scand. J. Statist.*, 17(4):333–353, 1990. ISSN 0303-6898.

P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B*, 52(3):443–452, 1990. ISSN 0035-9246.

T. Hasegawa, R. Yamaguchi, M. Nagasaki, S. Miyano, and S. Imoto. Inference of gene regulatory networks incorporating multi-source biological knowledge via a state space model with l1 regularization. *PLoS One*, 9(8):e105942, 2014.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1):31–57, 1989. ISSN 0090-5364.

W. Lee and Y. Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of multivariate analysis*, 111:241–255, 2012.

J. Lin, S. Basu, M. Banerjee, and G. Michailidis. Penalized maximum likelihood estimation of multi-layered gaussian graphical models. *J. Mach. Learn. Res.*, 17(1):5097–5147, Jan. 2016. ISSN 1532-4435.

C. McCarter and S. Kim. On sparse gaussian chain graph models. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2014.

A. Noor, E. Serpedin, M. Nounou, and H. Nounou. Inferring gene regulatory networks via nonlinear state-space models and exploiting sparsity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):1203–1211, 2012.

A. J. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

S. Särkkä. *Bayesian filtering and smoothing*, volume 3 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, Cambridge, 2013. ISBN 978-1-107-61928-9.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

2.1 Causal interpretation of the models

Let us discuss the interpretation of chain graph models in terms of the data generating distributions that they represent. The directed edges of a chain graph model can be interpreted as direct causal links, however, it is less obvious how to interpret the undirected edges. A heuristic argument is to say that they represent non-causal association, however, it remains unclear what the source of this association is. Lauritzen and Richardson [2002] argues that the undirected edges of LWF chain graphs represent feedback in continuous time dynamical systems that are observed discretely.

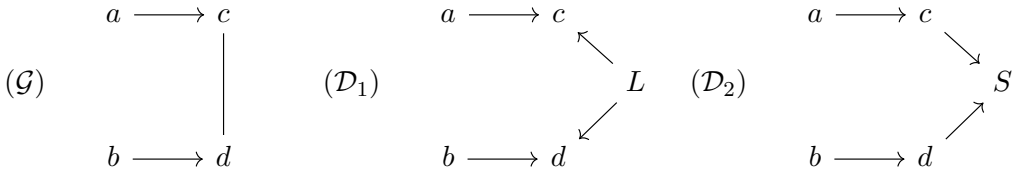

Figure 2.1: Left: A chain graph $\mathcal{G}$. Middle: Directed acyclic graph $\mathcal{D}_1$ where L is an latent confounder of c and d. Right: Directed acyclic graph $\mathcal{D}_2$ where S is a selection variable that is implicitly being conditioned on.

Consider the chain graph $\mathcal{G}$ and the two acyclic directed graphs $\mathcal{D}_1$ and $\mathcal{D}_2$ in Figure 2.1. Under the LWF Markov property, $\mathcal{G}$ encodes the conditional independencies:

$$a \perp\!\!\!\perp b, \quad a \perp\!\!\!\perp d \mid \{b, c\} \quad \text{and} \quad b \perp\!\!\!\perp c \mid \{a, d\}. \tag{2.1}$$

Now let us examine whether we can explain the association between c and d in $\mathcal{G}$ by a directed acyclic graph. In $\mathcal{D}_1$ we attempt to attribute the association to the presence of a latent confounder L causing c and d. However, this opens a d-connecting path from a to d when conditioning on the collider c, such that $a \not\perp\!\!\!\perp d \mid \{b, c\}$ in $\mathcal{D}_1$, which is not compatible with the conditional independencies (2.1) of $\mathcal{G}$ under the LWF Markov property. In $\mathcal{D}_2$ we attempt to explain the association between c and d by a selection variable S, which is implicitly being conditioned upon. This indeed opens a d-connecting path, which gives the association between c and d. However, it also opens a d-connecting path from a to b, such that $a \not\perp\!\!\!\perp b$ in $\mathcal{D}_2$, which is again not compatible with (2.1). Thus, neither $\mathcal{D}_1$ nor $\mathcal{D}_2$ are causal explanations for the association between c and d.

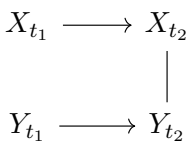

Figure 2.2: LWF chain graph for stochastic processes (X_t) and (Y_t) observed discretely.

The potential causal explanation for the association given in Section 6.3 of Lauritzen and Richardson [2002] is the following. Consider the situation where the nodes a, b, c, d represent measurements of two continuous time stochastic processes (X_t) and (Y_t) at discrete time points $t_1 < t_2$. See Figure 2.2. Under the LWF Markov property the density of $(X_{t_1}, X_{t_2}, Y_{t_1}, Y_{t_2})$ factorizes as

$$f(x_{t_1}, x_{t_2}, y_{t_1}, y_{t_2}) = f(x_{t_1}) f(y_{t_1}) f(x_{t_2}, y_{t_2} \mid x_{t_1}, x_{t_1})$$

which suggests that (X_{t_2}, Y_{t_2}) can be realized by a structural assignment

$$(X_{t_2}, Y_{t_2}) \leftarrow G(X_{t_1}, Y_{t_1}).$$

We can then choose to represent the data generating mechanism G as a Gibbs sampler as follows. The inputs are the realizations (x_{t_1}, y_{t_1}) of (X_{t_1}, Y_{t_1}) and some initial points $x^{(0)}, y^{(0)} \in \mathbb{R}$. Then for $k \geq 1$ we update the values $x^{(k)}$ and $y^{(k)}$ by drawing realizations from the conditional distributions

$$x^{(k)} \sim f_{X_{t_2}|Y_{t_2},X_{t_1},Y_{t_1}}(x_{t_2} \mid y^{(k-1)}, x_{t_1}, y_{t_1})$$
$$y^{(k)} \sim f_{Y_{t_2}|X_{t_2},X_{t_1},Y_{t_1}}(y_{t_2} \mid x^{(k)}, x_{t_1}, y_{t_1})$$

until the Gibbs sampler has converged after k' steps, and then we set $X_{t_2} = x^{(k')}$ and $Y_{t_2} = y^{(k')}$. Proposition 6 of Lauritzen and Richardson [2002] show that the distribution realized by this scheme has exactly the conditional independencies (2.1). The interpretation is that we observe the dynamical system (X_t, Y_t) at t_1, and then the stochastic processes interact through feedback until they reach an equilibrium distribution, which we observe at t_2. This can heuristically be interpreted as the infinite directed acyclic graph in Figure 2.3, where (X_{t_2}, Y_{t_2}) are realized as the stationary distribution in the limit $k \to \infty$.

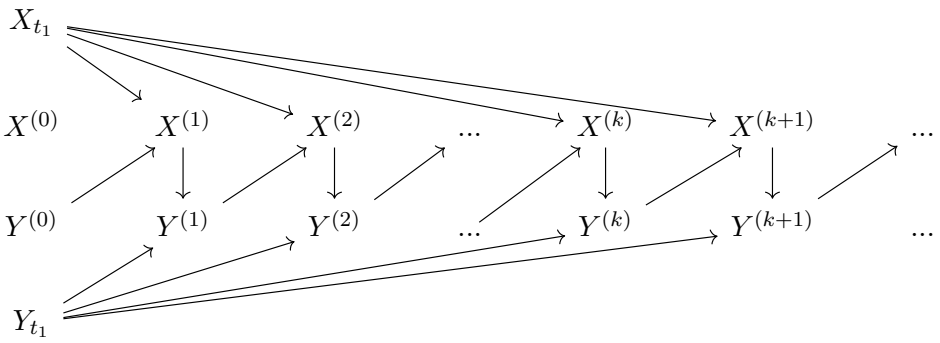

Figure 2.3: Infinite directed acyclic graph depicting the causal generating mechanism by a feedback between the stochastic processes (X_t) and (Y_t).

In conclusion, the canonical interpretation of LWF chain graph models is that they represent dynamical systems with feedback, where we observe the equilibrium distribution at discrete time points. It is less obvious how to interpret AMP chain graph models. The conditional independencies implied by $\mathcal{G}$ in Figure 2.1 under the AMP Markov properties are the following:

$$a \perp\!\!\!\perp b, \quad a \perp\!\!\!\perp d \mid b \quad \text{and} \quad c \perp\!\!\!\perp b \mid a.$$

These conditional independencies are consistent with $\mathcal{D}_1$ where L is a latent confounder. However, not all AMP chain graphs are Markov equivalent to a directed acyclic graph, and to the best of our knowledge there is no data generating mechanism analogous to the Gibbs sampler presented here, which in general explains the meaning of the undirected edges in AMP chain graphs. Nonetheless, AMP chain graph models are still valid statistical models that can represent a set of conditional independencies regardless of whether they can be given a causal interpretation.

Chapter 3

Learning Summary Graphs of Time Series

Sebastian Weichwald, Martin E. Jakobsen, Phillip B. Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 27–36, Vancouver, CA, 08–14 Dec 2020. PMLR.

Proceedings of Machine Learning Research 123:27–36, 2020 NeurIPS 2019 Competition and Demonstration Track

Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values

Sebastian Weichwald SWEICHWALD@MATH.KU.DK
Martin E Jakobsen M.JAKOBSEN@MATH.KU.DK
Phillip B Mogensen PBM@MATH.KU.DK
Lasse Petersen LP@MATH.KU.DK
Nikolaj Thams THAMS@MATH.KU.DK
Gherardo Varando GHERARDO.VARANDO@MATH.KU.DK
Copenhagen Causality Lab, Department of Mathematical Sciences, University of Copenhagen

Editors: Hugo Jair Escalante and Raia Hadsell

Abstract

In this article, we describe the algorithms for causal structure learning from time series data that won the Causality 4 Climate competition at the Conference on Neural Information Processing Systems 2019 (NeurIPS). We examine how our combination of established ideas achieves competitive performance on semi-realistic and realistic time series data exhibiting common challenges in real-world Earth sciences data. In particular, we discuss a) a rationale for leveraging linear methods to identify causal links in non-linear systems, b) a simulation-backed explanation as to why large regression coefficients may predict causal links better in practice than small p-values and thus why normalising the data may sometimes hinder causal structure learning. For benchmark usage, we detail the algorithms here and provide implementations at github.com/sweichwald/tidybench. We propose the presented competition-proven methods for baseline benchmark comparisons to guide the development of novel algorithms for structure learning from time series.

Keywords: Causal discovery, structure learning, time series, scaling.

1. Introduction

Inferring causal relationships from large-scale observational studies is an essential aspect of modern climate science (Runge et al., 2019a,b). However, randomised studies and controlled interventions cannot be carried out, due to both ethical and practical reasons. Instead, simulation studies based on climate models are state-of-the-art to study the complex patterns present in Earth climate systems (IPCC, 2013).

Causal inference methodology can integrate and validate current climate models and can be used to probe cause-effect relationships between observed variables. The Causality 4 Climate (C4C) NeurIPS competition (Runge et al., 2020) aimed to further the understanding and development of methods for structure learning from time series data exhibiting common challenges in and properties of realistic weather and climate data.

Structure of this work Section 2 introduces the structure learning task considered. In Section 3, we describe our winning algorithms. With a combination of established ideas, our algorithms achieved competitive performance on semi-realistic data across all 34 challenges in the C4C competition track. Furthermore, at the time of writing, our algorithms lead the rankings for all hybrid and realistic data set categories available on the CauseMe.net benchmark platform which also offers additional synthetic data categories (Runge et al., 2019a). These algorithms—which can be implemented in a few lines of code—are built on simple methods, are computationally efficient, and exhibit solid performance across a variety of different data sets. We therefore encourage the use of these algorithms as baseline benchmarks and guidance of future algorithmic and methodological developments for structure learning from time series.

Beyond the description of our algorithms, we aim at providing intuition that can explain the phenomena we have observed throughout solving the competition task. First, if we *only* ask whether a causal link exists in some non-linear time series system, then we may sidestep the extra complexity of explicit non-linear model extensions (cf. Section 4). Second, when data has a meaningful natural scale, it may—somewhat unexpectedly—be advisable to forego data normalisation and to use raw (vector auto)-regression coefficients instead of p-values to assess whether a causal link exists or not (cf. Section 5).

2. Causal structure learning from time-discrete observations

The task of inferring the causal structure from observational data is often referred to as 'causal discovery' and was pioneered by Pearl (2009) and Spirtes et al. (2001). Much of the causal inference literature is concerned with structure learning from independent and identically distributed (iid) observations. Here, we briefly review some aspects and common assumptions for causally modelling time-evolving systems. More detailed and comprehensive information can be found in the provided references.

Time-discrete observations We may view the discrete-time observations as arising from an underlying continuous-time causal system (Peters et al., 2020). While difficult to conceptualise, the correspondence between structural causal models and differential equation models can be made formally precise (Mooij et al., 2013; Rubenstein et al., 2018; Bongers and Mooij, 2018). Taken together, this yields some justification for modelling dynamical systems by discrete-time causal models.

Summary graph as inferential target It is common to assume a time-homogeneous causal structure such that the dynamics of the observation vector X are governed by $X^t := F(X^{\mathrm{past}(t)}, N^t)$ where the function F determines the next observation based on past values $X^{\mathrm{past}(t)}$ and the noise innovation N^t. Here, structure learning amounts to identifying the summary graph with adjacency matrix A that summarises the causal structure in the following sense: the $(i,j)^{\mathrm{th}}$ entry of the matrix A is 1 if $X_i^{\mathrm{past}(t)}$ enters the structural equation of X_i^t via the i^{th} component of F and 0 otherwise. If $A_{ij} = 1$, we say that "X_i causes X_j". While summary graphs can capture the existence and non-existence of cause-effect relationships, they do in general not correspond to a time-agnostic structural causal model that admits a causal semantics consistent with the underlying time-resolved structural causal model (Rubenstein et al., 2017; Janzing et al., 2018).

Time structure may be helpful for discovery In contrast to the iid setting, the Markov equivalence class of the summary graph induced by the structural equations of a dynamical system is a singleton when assuming causal sufficiency and no instantaneous effects (Peters et al., 2017; Mogensen and Hansen, 2020). This essentially yields a justification and a constraint-based causal inference perspective on Wiener-Granger-causality (Wiener, 1956; Granger, 1969; Peters et al., 2017).

Challenges for causal structure learning from time series data Structure learning from time series is a challenging task hurdled by further problems such as time-aggregation, time-delays, and time-subsampling. All these challenges were considered in the C4C competition and are topics of active research (Danks and Plis, 2013; Hyttinen et al., 2016).

3. The time series discovery benchmark (`tidybench`): Winning algorithms

We developed four simple algorithms,

`SLARAC`	Subsampled Linear Auto-Regression Absolute Coefficients (cf. Alg. 1)
`QRBS`	Quantiles of Ridge regressed Bootstrap Samples (cf. Alg. 2)
`LASAR`	LASso Auto-Regression
`SELVAR`	Selective auto-regressive model

which came in first in 18 and close second in 13 out of the 34 C4C competition categories and won the overall competition (Runge et al., 2020). Here, we provide detailed descriptions of the `SLARAC` and `QRBS` algorithms. Analogous descriptions for the latter two algorithms and implementations of all four algorithms are available at github.com/sweichwald/tidybench.

All of our algorithms output an edge score matrix that contains for each variable pair (X_i, X_j) a score that reflects how likely it is that the edge $X_i \to X_j$ exists. Higher scores correspond to edges that are inferred to be more likely to exist than edges with lower scores, based on the observed data. That is, we rank edges relative to one another but do not perform hypothesis tests for the existence of individual edges. A binary decision can be obtained by choosing a cut-off value for the obtained edge scores. In the C4C competition, submissions were compared to the ground-truth cause-effect adjacency matrix and assessed based on the achieved ROC-AUC when predicting which causal links exist.

The idea behind our algorithms is the following: regress present on past values and inspect the regression coefficients to decide whether one variable is a Granger-cause of another. `SLARAC` fits a VAR model on bootstrap samples of the data each time choosing a random number of lags to include; `QRBS` considers bootstrap samples of the data and Ridge-regresses time-deltas $X(t) - X(t-1)$ on the preceding values $X(t-1)$; `LASAR` considers bootstrap samples of the data and iteratively—up to a maximum lag—LASSO-regresses the residuals of the preceding step onto values one step further in the past and keeps track of the variable selection at each lag to fit an OLS regression in the end with only the selected variables at selected lags included; and `SELVAR` selects edges employing a hill-climbing procedure based on the leave-one-out residual sum of squares and finally scores the selected edges with the absolute values of the regression coefficients. In the absence of instantaneous effects and hidden confounders, Granger-causes are equivalent to a variable's causal parents (Peters et al., 2017, Theorem 10.3). In Section 5, we argue that the size of

the regression coefficients may in certain scenarios be more informative about the existence of a causal link than standard test statistics for the hypothesis of a coefficient being zero. It is argued that for additive noise models, information about the causal ordering may be contained in the raw marginal variances. In test statistics such as the F- and T-statistics, this information is lost when normalising by the marginal variances.

4. Capturing non-linear cause-effect links by linear methods

We explain the rationale behind our graph reconstruction algorithms and how they may capture non-linear dynamics despite being based on linearly regressing present on past values. For simplicity we will outline the idea in a multivariate regression setting with additive noise, but it extends to the time series setting by assuming time homogeneity.

Let $N, X(t_1), X(t_2) \in \mathbb{R}^d$ be random variables such that $X(t_2) := F(X(t_1)) + N$ for some differentiable function $F = (F_1, \ldots, F_d) : \mathbb{R}^d \to \mathbb{R}^d$. Assume that N has mean zero, that it is independent from $X(t_1)$, and that it has mutually independent components. For each $i, j = 1, \ldots, d$ we define the quantity of interest

$$\theta_{ij} = \mathbb{E}\left|\partial_i F_j\left(X(t_1)\right)\right|,$$

such that θ_{ij} measures the expected effect from $X_i(t_1)$ to $X_j(t_2)$. We take the matrix $\Theta = \left(\mathbf{1}_{\theta_{ij}>0}\right)$ as the adjacency matrix of the summary graph between $X(t_1)$ and $X(t_2)$.

In order to detect regions with non-zero gradients of F we create bootstrap samples $\mathcal{D}_1, \ldots, \mathcal{D}_B$. On each bootstrap sample $\mathcal{D}_b$ we obtain the regression coefficients $\widehat{A}_b$ as estimate of the directional derivatives by a (possibly penalised) linear regression technique. Intuitively, if θ_{ij} were zero, then on any bootstrap sample we would obtain a small non-zero contribution. Conversely, if θ_{ij} were non-zero, then we may for some bootstrap samples obtain a linear fit of $X_j(t_2)$ with large absolute regression coefficient for $X_i(t_1)$. The values obtained on each bootstrap sample are then aggregated by, for example, taking the average of the absolute regression coefficients $\widehat{\theta}_{ij} = \frac{1}{B}\sum_{b=1}^{B}\left|(\widehat{A}_b)_{ij}\right|$.

This amounts to searching the predictor space for an effect from $X_i(t_1)$ to $X_j(t_2)$, which is approximated linearly. It is important to aggregate the absolute values of the coefficients to avoid cancellation of positive and negative coefficients. The score $\widehat{\theta}_{ij}$ as such contains no information about whether the effect from $X_i(t_1)$ to $X_j(t_2)$ is positive or negative and it cannot be used to predict $X_j(t_2)$ from $X_i(t_1)$. It serves as a score for the existence of a link between the two variables. This rationale explains how linear methods may be employed for edge detection in non-linear settings without requiring extensions of Granger-type methods that explicitly model the non-linear dynamics and hence come with additional sample complexity (Marinazzo et al., 2008, 2011; Stramaglia et al., 2012, 2014).

5. Large regression coefficients may predict causal links better in practice than small p-values

This section aims at providing intuition behind two phenomena: We observed a considerable drop in the accuracy of our edge predictions whenever 1) we normalised the data or 2) used the T-statistics corresponding to testing the hypothesis of regression coefficients being zero to score edges instead of the coefficients' absolute magnitude. While one could

LARGE REGRESSION COEFFICIENTS MAY PREDICT CAUSAL LINKS BETTER THAN SMALL P-VALUES

Algorithm 1: Subsampled Linear Auto-Regression Absolute Coefficients (`SLARAC`)

Input : Data $\boldsymbol{X}$ with T time samples $\boldsymbol{X}(1), \ldots, \boldsymbol{X}(T)$ over d variables.
Parameters : Max number of lags, $L \in \mathbb{N}$.
Number of bootstrap samples, $B \in \mathbb{N}$.
Individual bootstrap sample sizes, $\{v_1, \ldots, v_B\}$.
Output : A $d \times d$ real-valued score matrix, $\widehat{A}$.

Initialise A_{full} as a $d \times dL$ matrix of zeros and $\widehat{A}$ as an empty $d \times d$ matrix;
for $b = 1, \ldots, B$ **do**
  lags $\leftarrow$ random integer in $\{1, \ldots, L\}$;
  Draw a bootstrap sample $\{t_1, \ldots, t_{v_b}\}$ from $\{\text{lags}+1, \ldots, T\}$ with replacement;
  $\boldsymbol{Y}^{(b)} \leftarrow (\boldsymbol{X}(t_1), \ldots \boldsymbol{X}(t_{v_b}))$;
  $\boldsymbol{X}^{(b)}_{\text{past}} \leftarrow \begin{pmatrix} \boldsymbol{X}(t_1 - 1) & \cdots & \boldsymbol{X}(t_1 - \text{lags}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}(t_{v_b} - 1) & \cdots & \boldsymbol{X}(t_{v_b} - \text{lags}) \end{pmatrix}$;
  Fit OLS estimate β of regressing $\boldsymbol{Y}^{(b)}$ onto $\boldsymbol{X}^{(b)}_{\text{past}}$;
  Zero-pad β such that $\dim \beta = d \times dL$;
  $A_{\text{full}} \leftarrow A_{\text{full}} + |\beta|$;
end
Aggregate $(\widehat{A})_{i,j} \leftarrow \max((A_{\text{full}})_{i,j+0\cdot d}, \ldots, (A_{\text{full}})_{i,j+L\cdot d})$ for every i, j;
Return: Score matrix $\widehat{A}$.

Algorithm 2: Quantiles of Ridge regressed Bootstrap Samples (`QRBS`)

Input : Data $\boldsymbol{X}$ with T time samples $\boldsymbol{X}(1), \ldots, \boldsymbol{X}(T)$ over d variables.
Parameters : Number of bootstrap samples, $B \in \mathbb{N}$.
Size of bootstrap samples, $v \in \mathbb{N}$.
Ridge regression penalty, $\kappa \geq 0$.
Quantile for aggregating scores, $q \in [0, 1]$.
Output : A $d \times d$ real-valued score matrix, $\widehat{A}$.

for $b = 1, \ldots, B$ **do**
  Draw a bootstrap sample $\{t_1, \ldots, t_v\}$ from $\{2, \ldots, T\}$ with replacement;
  $\boldsymbol{Y}^{(b)} \leftarrow (\boldsymbol{X}(t_1) - \boldsymbol{X}(t_1 - 1), \ldots, \boldsymbol{X}(t_v) - \boldsymbol{X}(t_v - 1))$;
  $\boldsymbol{X}^{(b)} \leftarrow (\boldsymbol{X}(t_1 - 1), \ldots, \boldsymbol{X}(t_v - 1))$;
  Fit a ridge regression of $\boldsymbol{Y}^{(b)}$ onto $\boldsymbol{X}^{(b)}$: $\widehat{A}_b = \arg\min_A \|\boldsymbol{Y}^{(b)} - A\boldsymbol{X}^{(b)}\| + \kappa\|A\|$;
end
Aggregate $\widehat{A} \leftarrow q^{th}$ element-wise quantile of $\{|\widehat{A}_1|, \ldots, |\widehat{A}_B|\}$;
Return Score matrix $\widehat{A}$.

try to attribute these phenomena to some undesired artefact in the competition setup, it is instructive to instead try to understand when exactly one would expect such behaviour.

We illustrate a possible explanation behind these phenomena and do so in an iid setting in favour of a clear exposition, while the intuition extends to settings of time series observations and our proposed algorithms. The key remark is, that under comparable noise variances, the variables' marginal variances tend to increase along the causal ordering. If data are observed at comparable scales—say sea level pressure in different locations measured in the same units—or at scales that are in some sense naturally relative to the true data generating mechanism, then absolute regression coefficients may be preferable to T-test statistics. Effect variables tend to have larger marginal variance than their causal ancestors. This helpful signal in the data is diminished by normalising the data or the rescaling when computing the T-statistics corresponding to testing the regression coefficients for being zero. This rationale is closely linked to the identifiability of Gaussian structural equation models under equal error variances Peters and Bühlmann (2014). Without any prior knowledge about what physical quantities the variables correspond to and their natural scales, normalisation remains a reasonable first step. We are not advocating that one should use the raw coefficients and not normalise data, but these are two possible alterations of existing structure learning procedures that may or may not, depending on the concrete application at hand, be worthwhile exploring. Our algorithms do not perform data normalisation, so the choice is up to the user whether to feed normalised or raw data, and one could easily change to using p-values or T-statistics instead of raw coefficients for edge scoring.

5.1. Instructive iid case simulation illustrates scaling effects

We consider data simulated from a standard acyclic linear Gaussian model. Let $N \sim \mathcal{N}\left(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)\right)$ be a d-dimensional random variable and let $\boldsymbol{B}$ be a $d \times d$ strictly lower-triangular matrix. Further, let X be a d-valued random variable constructed according to the structural equation $X = \boldsymbol{B}X + N$, which induces a distribution over X via $X = (I - \boldsymbol{B})^{-1}N$. We have assumed, without loss of generality, that the causal order is aligned such that X_i is further up in the causal order than X_j whenever $i < j$. We ran 100 repetitions of the experiment, each time sampling a random lower triangular 50×50-matrix $\boldsymbol{B}$ where each entry in the lower triangle is drawn from a standard Gaussian with probability $1/4$ and set to zero otherwise. For each such obtained $\boldsymbol{B}$ we sample $n = 200$ observations from $X = \boldsymbol{B}X + N$ which we arrange in a data matrix $\boldsymbol{X} \in \mathbb{R}^{200\times 50}$ of zero-centred columns denoted by $\boldsymbol{X}_j$.

We regress each X_j onto all remaining variables $X_{\neg j}$ and compare scoring edges $X_i \to X_j$ by the absolute values of a) the regression coefficients $|\widehat{b}_{i\to j}|$, versus b) the T-statistics $|\widehat{t}_{i\to j}|$ corresponding to testing the hypothesis that the regression coefficient $\widehat{b}_{i\to j}$ is zero. That is, we consider

$$|\widehat{b}_{i\to j}| = \left|(\boldsymbol{X}_{\neg j}^\top \boldsymbol{X}_{\neg j})^{-1}\boldsymbol{X}_{\neg j}^\top \boldsymbol{X}_j\right|_i$$

versus

$$|\widehat{t}_{i\to j}| = |\widehat{b}_{i\to j}|\sqrt{\frac{\widehat{\mathrm{var}}(X_i|X_{\neg i})}{\widehat{\mathrm{var}}(X_j|X_{\neg j})}}\sqrt{\frac{(n-d)}{\left(1 - \widehat{\mathrm{corr}}^2(X_i, X_j|X_{\neg\{i,j\}})\right)}} \tag{1}$$

LARGE REGRESSION COEFFICIENTS MAY PREDICT CAUSAL LINKS BETTER THAN SMALL P-VALUES

where $\widehat{\text{var}}(X_j|X_{\neg j})$ is the residual variance after regressing X_j onto the other variables $X_{\neg j}$, and $\widehat{\text{corr}}(X_i, X_j|X_{\neg\{i,j\}})$ is the residual correlation between X_i and X_j after regressing both onto the remaining variables.

We now compare, across three settings, the AUC obtained by either using the absolute value of the regression coefficients $|\widehat{b}_{i\to j}|$ or the absolute value of the corresponding T-statistics $|\widehat{t}_{i\to j}|$ for edge scoring. Results are shown in the left, middle, and right panel of Figure 1, respectively.

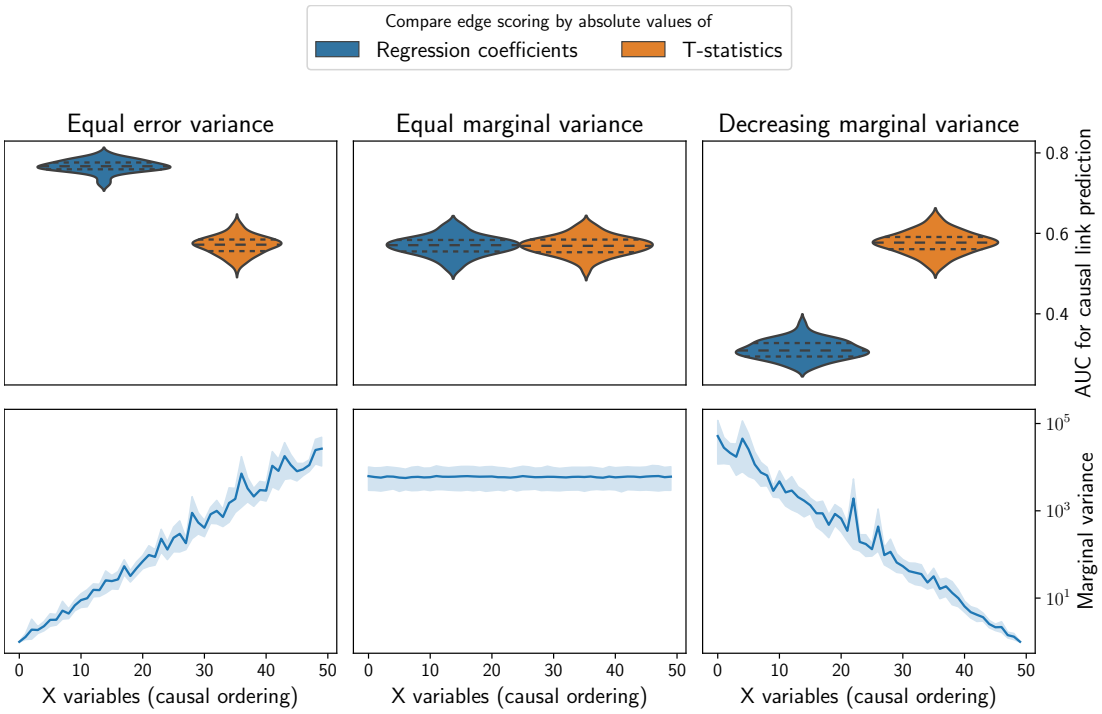

Figure 1: Results of the simulation experiment described in Section 5.1. Data is generated from an acyclic linear Gaussian model, in turn each variable is regressed onto all remaining variables and either the raw regression coefficient $|\widehat{b}_{i\to j}|$ or the corresponding T-statistics $|\widehat{t}_{i\to j}|$ is used to score the existence of an edge $i \to j$. The top row shows the obtained AUC for causal link prediction and the bottom row the marginal variance of the variables along the causal ordering. The left panel shows naturally increasing marginal variance for equal error variances, for the middle and right panel the model parameters and error variances are rescaled to enforce equal and decreasing marginal variance, respectively.

In the setting with equal error variances $\sigma_i^2 = \sigma_j^2 \ \forall i, j$**,** we observe that i) the absolute regression coefficients beat the T-statistics for edge predictions in terms of AUC, and ii) the marginal variances naturally turn out to increase along the causal ordering.

When moving from $|\widehat{b}_{i\to j}|$ to $|\widehat{t}_{i\to j}|$ for scoring edges, we multiply by a term that compares the relative residual variance of X_i and X_j. If X_i is before X_j in the causal ordering it tends to have both smaller marginal and—in our simulation set-up—residual variance than X_j as it becomes increasingly more difficult to predict variables further down the causal ordering. In this case, the fraction of residual variances will tend to be smaller than one and consequently the raw regression coefficients $|\widehat{b}_{i\to j}|$ will be shrunk when moving to $|\widehat{t}_{i\to j}|$. This can explain the worse performance of the T-statistics compared to the raw regression coefficients for edge scoring as scores will tend to be shrunk when in fact $X_i \to X_j$.

Enforcing equal marginal variances by rescaling the rows of B and the σ_i^2's, we indeed observe that regression coefficients and T-statistics achieve comparable performance in edge prediction in this somewhat artificial scenario. Here, neither the marginal variances nor the residual variances appear to contain information about the causal ordering any more and the relative ordering between regression coefficients and T-statistics is preserved when multiplying by the factor highlighted in Equation 1.

Enforcing decreasing marginal variances by rescaling the rows of B and the σ_i^2's, we can, in line with our above reasoning, indeed obtain an artificial scenario in which the T-statistics will outperform the regression coefficients in edge prediction, as now, the factors we multiply by will work in favour of the T-statistics.

6. Conclusion and Future Work

We believe competitions like the Causality 4 Climate competition (Runge et al., 2020) and causal discovery benchmark platforms like CauseMe.net (Runge et al., 2019a) are important for bundling and informing the community's joint research efforts into methodology that is readily applicable to tackle real-world data. In practice, there are fundamental limitations to causal structure learning that ultimately require us to employ untestable causal assumptions to proceed towards applications at all. Yet, both these limitations and assumptions are increasingly well understood and characterised by methodological research and time and again need to be challenged and examined through the application to real-world data.

Beyond the algorithms presented here and proposed for baseline benchmarks, different methodology as well as different benchmarks may be of interest. For example, our methods detect causal links and are viable benchmarks for the structure learning task but they do not per se enable predictions about the interventional distributions.

Acknowledgments

The authors thank Niels Richard Hansen, Steffen Lauritzen, and Jonas Peters for insightful discussions. Thanks to the organisers for a challenging and insightful Causality 4 Climate NeurIPS competition. NT was supported by a research grant (18968) from VILLUM FONDEN. LP and GV were supported by a research grant (13358) from VILLUM FONDEN. MEJ and SW were supported by the Carlsberg Foundation.

References

S. Bongers and J. M. Mooij. From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*, 2018.

D. Danks and S. Plis. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, 2013.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

A. Hyttinen, S. Plis, M. Järvisalo, F. Eberhardt, and D. Danks. Causal Discovery from Subsampled Time Series Data by Constraint Optimization. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, 2016.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2013.

D. Janzing, P. K. Rubenstein, and B. Schölkopf. Structural causal models for macro-variables in time-series. *arXiv preprint arXiv:1804.03911*, 2018.

D. Marinazzo, M. Pellicoro, and S. Stramaglia. Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77(5):056215, 2008.

D. Marinazzo, W. Liao, H. Chen, and S. Stramaglia. Nonlinear connectivity by Granger causality. *NeuroImage*, 58(2):330 – 338, 2011.

S. W. Mogensen and N. R. Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.

J. M. Mooij, D. Janzing, and B. Schölkopf. From Ordinary Differential Equations to Structural Causal Models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2013.

J. Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

J. Peters, S. Bauer, and N. Pfister. Causal models for dynamical systems. *arXiv preprint arXiv:2001.06208*, 2020.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2017.

P. K. Rubenstein, S. Bongers, J. M. Mooij, and B. Schölkopf. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2018.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, E. H. Muñoz-Marí, J. andand van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):2553, 2019a.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), 2019b.

J. Runge, X.-A. Tibau, M. Bruhns, J. Muñoz-Marí, and G. Camps-Valls. The causality for climate competition. In Hugo Jair Escalante and Raia Hadsell, editors, *PMLR NeurIPS Competition & Demonstration Track Postproceedings*, Proceedings of Machine Learning Research. PMLR, 2020. URL https://causeme.uv.es/. Forthcoming.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2 edition, 2001.

S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo. Expanding the transfer entropy to identify information circuits in complex systems. *Physical Review E*, 86(6):066211, 2012.

S. Stramaglia, J. M. Cortes, and D. Marinazzo. Synergy and redundancy in the Granger causal analysis of dynamical networks. *New Journal of Physics*, 16(10):105003, 2014.

N. Wiener. The theory of prediction. *Modern Mathematics for Engineers*, 1956.

3.1 Score matrix as a causal estimand

In this section we will briefly explain the evaluation method of the Causality 4 Climate competition[1], and discuss how to meaningfully evaluate causal structure learning methodology.

Given data from a multivariate time series, the inferential target of interest in the competition was the summary graph of the underlying time series represented by an adjacency matrix A such that $A_{ij} = 1$ if and only if there is a directed edge $X_s^{(i)} \to X_t^{(j)}$ for some $s < t$ in the graph representing the entire time series. However, instead of reporting an adjacency matrix A, the task for the competition was to provide a score matrix S, such that a large value of S_{ij} corresponds to high confidence that $A_{ij} = 1$ in the summary graph.

The evaluation method given a score matrix S based on data from the times series was as follows. First the entries are scaled to range from 0 to 1 such that each score is interpreted as a probability score. The summary graph reconstruction problem is then treated as a binary classification problem: each edge in the summary graph is represented by a 0 or 1 in the adjacency matrix A, and the entries of the (scaled) score matrix S are the probability predictions of that classification problem. The evaluation metric of a score matrix S is then the AUC score of those predictions. More precisely, a threshold t is varied from 0 to 1 converting the scaled score matrix into 0–1 predictions of the presence of an edge in the summary graph, $\hat{A}_{ij}(t) = 1(S_{ij}/\max S \geq t)$, and the false positive rate (FPR) and true positive rate (TPR) of the predictions are computed as a functions of the threshold. The AUC is then the area under the curve $\{(\mathrm{FPR}(t), \mathrm{TPR}(t)) \mid t \in [0, 1]\}$, where $\mathrm{AUC} \in [0, 1]$ and a score close to 1 is good. In the case of, e.g., a 5-dimensional time series, the AUC is based on $5^2 - 5 = 20$ predictions (there is assumed always to be dependence within each coordinate on itself).

However, there are several issues with using this as a metric to evaluate causal structure learning. Firstly, reporting a score matrix is insufficient for a real world application of structure learning, since one must make a binary decision on whether to include an edge in the summary graph. Moreover, a high AUC value only ensures the existence of a threshold with a low FPR and a high TPR, but this threshold cannot be determined without the true summary graph. In a normal classification problem, this threshold can be determined by choosing it according to the performance on a test data set with available labels. However, outside a simulation setup, we can never obtain a "data set" consisting of a true adjacency matrix of a summary graph of a causal problem.

In this sense, the score matrix is problematic, since there is no principled way to choose the threshold. However, this is a valid critique of constraint-based causal structure learning algorithms in general. In the skeleton estimation of, e.g., the PC algorithm, one must choose a significance level of the conditional independence tests. The significance level for each individual test has a meaningful interpretation in terms of type I error control, but when performing a sequence of dependent conditional independence tests, where edge inclusions or exclusions determines which subsequent tests are performed, the significance level looses its interpretation as a type I error control in relation to the estimated skeleton. Therefore, the significance level of a constraint-based structure learning algorithm is a threshold parameter, in the same way as the threshold of the score matrix of the summary graph in this application.

In general, it can be argued that an estimator should be evaluated according to how it is going to be used in practice. If the goal is simply graph reconstruction, then a measure such as the structural Hamming distance, which measures the distance between graphs in terms of edge insertions or deletions and flips, is a more appropriate metric. In the case of causal inference, the goal is typically to predict the effect of interventions. Here the structural intervention distance [Peters and Bühlmann, 2015], which

[1] `https://causeme.uv.es/neurips2019`

measures the distance from one directed acyclic graph $\mathcal{H}$ to another $\mathcal{G}$ by the number of incorrectly inferred interventional distribution there are in $\mathcal{H}$ relative to $\mathcal{G}$, is another choice which directly reflects the use of the graph for causal inference.

In the sense of real world usability, the score matrix can be regarded as an exploratory tool. The summary graph cannot be used to predict the effect of interventions, since $A_{ij} = 1$ only ensures the existence of a directed edge $X_i^s \to X_j^t$ for some $s < t$. Therefore, its role as a causal estimand is rather as an exploratory tool for causal hypothesis generation, which can afterwards be further investigated by a practitioner.

Chapter 4

Conditional Independence Testing

Lasse Petersen and Niels Richard Hansen. Testing Conditional Independence via Quantile Regression Based Partial Copulas. *Journal of Machine Learning Research*, 22(70):1–47, 2021b.

Journal of Machine Learning Research 22 (2021) 1-47 Submitted 9/20; Revised 1/21; Published 3/21

Testing Conditional Independence via Quantile Regression Based Partial Copulas

Lasse Petersen LP@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5, 2100 Copenhagen, Denmark

Niels Richard Hansen NIELS.R.HANSEN@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5, 2100 Copenhagen, Denmark

Editor: Peter Spirtes

Abstract

The partial copula provides a method for describing the dependence between two random variables X and Y conditional on a third random vector Z in terms of nonparametric residuals U_1 and U_2. This paper develops a nonparametric test for conditional independence by combining the partial copula with a quantile regression based method for estimating the nonparametric residuals. We consider a test statistic based on generalized correlation between U_1 and U_2 and derive its large sample properties under consistency assumptions on the quantile regression procedure. We demonstrate through a simulation study that the resulting test is sound under complicated data generating distributions. Moreover, in the examples considered the test is competitive to other state-of-the-art conditional independence tests in terms of level and power, and it has superior power in cases with conditional variance heterogeneity of X and Y given Z.

Keywords: Conditional independence testing, nonparametric testing, partial copula, conditional distribution function, quantile regression

1. Introduction

This paper introduces a new class of nonparametric tests of conditional independence between real-valued random variables, $X \perp\!\!\!\perp Y \mid Z$, based on quantile regression. Conditional independence is an important concept in many statistical fields such an graphical models and causal inference (Lauritzen, 1996; Spirtes et al., 2000; Pearl, 2009). However, Shah and Peters (2020) proved that conditional independence is an untestable hypothesis when the distribution of (X, Y, Z) is only assumed to be absolutely continuous with respect to Lebesgue measure.

More precisely, let $\mathcal{P}$ denote the set of distributions of (X, Y, Z) that are absolutely continuous with respect to Lebesgue measure. Let $\mathcal{H} \subset \mathcal{P}$ be those distributions for which conditional independence holds. Then Shah and Peters (2020) showed that if ψ_n is a

hypothesis test for conditional independence with uniformly valid level $\alpha \in (0,1)$ over $\mathcal{H}$,

$$\sup_{P \in \mathcal{H}} E_P(\psi_n) \leq \alpha,$$

then the test cannot have power greater than α against any alternative $P \in \mathcal{Q} := \mathcal{P} \setminus \mathcal{H}$. This is true even when restricting the distribution of (X, Y, Z) to have bounded support. The purpose of this paper is to identify a subset $\mathcal{P}_0 \subset \mathcal{P}$ of distributions and a test ψ_n that has asymptotic (uniform) level over $\mathcal{P}_0 \cap \mathcal{H}$ and power against a large set of alternatives within $\mathcal{P}_0 \setminus \mathcal{H}$.

Our starting point is the so-called partial copula construction. Letting $F_{X|Z}$ and $F_{Y|Z}$ denote the conditional distribution functions of X given Z and Y given Z, respectively, we define random variables U_1 and U_2 by

$$U_1 := F_{X|Z}(X \mid Z) \quad \text{and} \quad U_2 := F_{Y|Z}(Y \mid Z).$$

Then the joint distribution of U_1 and U_2 is called the partial copula and it can be shown that $X \perp\!\!\!\perp Y \mid Z$ implies $U_1 \perp\!\!\!\perp U_2$. Thus the question about conditional independence can be transformed into a question about independence. The main challenge with this approach is that the conditional distribution functions are unknown and must be estimated.

In Section 3 we propose an estimator of conditional distribution functions based on quantile regression. More specifically, we let $\mathcal{T} = [\tau_{\min}, \tau_{\max}]$ be a range of quantile levels for $0 < \tau_{\min} < \tau_{\max} < 1$, and let $Q(\mathcal{T} \mid z)$ denote the range of conditional $\mathcal{T}$-quantiles in the distribution $X \mid Z = z$. To estimate a conditional distribution function F given a sample $(X_i, Z_i)_{i=1}^n$ we propose to perform quantile regressions $\hat{q}_{k,z} = \hat{Q}^{(n)}(\tau_k \mid z)$ along an equidistant grid of quantile levels $(\tau_k)_{k=1}^m$ in $\mathcal{T}$, and then construct the estimator $\hat{F}^{(m,n)}$ by linear interpolation of the points $(\hat{q}_{k,z}, \tau_k)_{k=1}^m$. The main result of the first part of the paper is Theorem 5, which states that we can achieve the following bound on the estimation error

$$\|F - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty} := \sup_z \sup_{t \in Q(\mathcal{T}|z)} |F(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \in \mathcal{O}_P(g_P(n))$$

where g_P is a rate function describing the $\mathcal{O}_P$-consistency of the quantile regression procedure over the conditional $\mathcal{T}$-quantiles for P in a specified set of distributions $\mathcal{P}_0 \subset \mathcal{P}$. This result demonstrates how pointwise consistency of a quantile regression procedure over $\mathcal{P}_0$ can be transferred to the estimator $\hat{F}^{(m,n)}$, and we discuss how this can be extended to uniform consistency over $\mathcal{P}_0$. We conclude the section by reviewing a flexible model class from quantile regression where such consistency results are available.

In Section 4 we describe a generic method for testing conditional independence based on estimated conditional distribution functions, $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$, obtained from a sample $(X_i, Y_i, Z_i)_{i=1}^n$. From these estimates we compute

$$\hat{U}^{(n)}_{1,i} := \hat{F}^{(n)}_{X|Z}(X_i \mid Z_i) \quad \text{and} \quad \hat{U}^{(n)}_{2,i} := \hat{F}^{(n)}_{Y|Z}(Y_i \mid Z_i),$$

for $i = 1, \ldots, n$, which can then be plugged into a bivariate independence test. If $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$ are consistent with a sufficiently fast rate of convergence, properties of the bivariate test, in terms of level and power, can be transferred to the test of conditional independence.

The details of this transfer of properties depend on the specific test statistic. The main contribution of the second part of the paper is a detailed treatment of a test given in terms of a generalized correlation, estimated as

$$\hat{\rho}_n := \frac{1}{n}\sum_{i=1}^{n} \varphi\left(\hat{U}_{1,i}^{(n)}\right)\varphi\left(\hat{U}_{2,i}^{(n)}\right)^T$$

for a function $\varphi = (\varphi_1, \ldots, \varphi_q) : [0,1] \to \mathbb{R}^q$ satisfying certain regularity conditions. A main result is Theorem 14, which states that $\sqrt{n}\hat{\rho}_n$ converges in distribution toward $\mathcal{N}(0, \Sigma \otimes \Sigma)$ under the hypothesis of conditional independence whenever $\hat{F}_{X|Z}^{(n)}$ and $\hat{F}_{Y|Z}^{(n)}$ are $\mathcal{O}_P$-consistent with rates g_P and h_P satisfying $\sqrt{n}g_P(n)h_P(n) \to 0$. The covariance matrix Σ depends only on φ. We use this to show asymptotic pointwise level of the test when restricting to the set of distributions $\mathcal{P}_0$ where the required consistency can be obtained. We then proceed to show in Theorem 18 that $\sqrt{n}\hat{\rho}_n$ diverges in probability under a set of alternatives of conditional dependence when we have $\mathcal{O}_P$-consistency of the conditional distribution function estimators. This we use to show asymptotic pointwise power of the test. We also show how asymptotic uniform level and power can be achieved when $\hat{F}_{X|Z}^{(n)}$ and $\hat{F}_{Y|Z}^{(n)}$ are uniformly consistent over $\mathcal{P}_0$. Lastly, we provide an out-of-the-box procedure for conditional independence testing in conjunction with our quantile regression based conditional distribution function estimator $\hat{F}^{(m,n)}$ from Section 3.

In Section 5 we examine the proposed test through a simulation study where we assess the level and power properties of the test and benchmark it against existing nonparametric conditional independence tests. All proofs are collected in Appendix A.

2. Related Work

The partial copula and its application for conditional independence testing was initially introduced by Bergsma (2004) and further explored by Bergsma (2011). Its use for conditional independence testing has also been explored by Song (2009), Patra et al. (2016) and Liu et al. (2018). Moreover, properties of the partial copula was studied by Gijbels et al. (2015) and Spanhel and Kurz (2016) among others. A related but different approach for testing conditional independence via the factorization of the joint copula of (X, Y, Z) is given by Bouezmarni et al. (2012). Common for the existing approaches to using the partial copula for conditional independence testing is that the conditional distribution functions $F_{X|Z}$ and $F_{Y|Z}$ are estimated using a kernel smoothing procedure (Stute et al., 1986; Einmahl and Mason, 2005). The advantage of the approach is that the estimator is nonparametric, however, it does not scale well with the dimension of the conditioning variable Z. This is partly remedied by Haff and Segers (2015) who suggest a nonparametric pair-copula estimator whose convergence rate is independent of the dimension of Z. This estimator requires the simplifying assumption, which is a strong assumption not required for the validity of our approach. Moreover, it is not obvious how to incorporate parametric assumptions, such as a certain functional dependence between response and covariates, using kernel smoothing estimators, since there is only the choice of a kernel and a bandwidth. Furthermore, a treatment of the relationship between level and power properties of a partial copula based conditional independence test, and consistency of the conditional distribution function esti-

mator is lacking in the existing literature. In this work we take a novel approach to testing conditional independence using the partial copula by using quantile regression for estimating the conditional distribution functions. This allows for a distribution free modeling of the conditional distributions $X \mid Z = z$ and $Y \mid Z = z$ that can handle high-dimensionality of Z through penalization, and complicated response-predictor relationships by basis expansions. We also make the requirements on consistency of the conditional distribution function estimator that are needed to obtain level and power of the test explicit. A similar recent approach to testing conditional independence using regression methods is given by Shah and Peters (2020), who propose to test for vanishing correlation between the residuals after nonparametric conditional mean regression of X on Z and Y on Z. See also Ramsey (2014) and Fan et al. (2020). This approach captures dependence between X and Y given Z that lies in the conditional correlation. However, as is demonstrated through a simulation study in Section 5.5, it does not adequately account for conditional variance heterogeneity between X and Y given Z, while our partial copula based test captures the dependence more efficiently.

3. Estimation of Conditional Distribution Functions

Throughout the paper we restrict ourselves to the set of distributions $\mathcal{P}$ over the hypercube $[0,1]^{2+d}$ that are absolutely continuous with respect to Lebesgue measure. Let $(X, Y, Z) \sim P \in \mathcal{P}$ such that $X, Y \in [0,1]$ and $Z \in [0,1]^d$. When we speak of the distribution of X given Z relative to P we mean the conditional distribution that is induced when $(X, Y, Z) \sim P$. In this section we consider estimation of the conditional distribution function $F_{X|Z}$ of X given Z using quantile regression. Estimation of $F_{Y|Z}$ can be carried out analogously.

3.1 Conditional distribution and quantile functions

Given $z \in [0,1]^d$ we denote by

$$F_{X|Z}(t \mid z) := P(X \leq t \mid Z = z)$$

the conditional distribution function of $X \mid Z = z$ for $t \in [0,1]$. We denote by

$$Q_{X|Z}(\tau \mid z) := \inf\{t \in [0,1] \mid F_{X|Z}(t \mid z) \geq \tau\}$$

the conditional quantile function of the conditional distribution $X \mid Z = z$ for $\tau \in [0,1]$ and $z \in [0,1]^d$. We will omit the subscript in $F_{X|Z}$ and $Q_{X|Z}$ when the conditional distribution of interest is clear from the context.

In quantile regression one models the function $z \mapsto Q(\tau \mid z)$ for fixed $\tau \in [0,1]$. Estimation of the quantile regression function is carried out by solving the empirical risk minimization problem

$$\hat{Q}(\tau \mid \cdot) \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^{n} L_\tau \left(X_i - f(Z_i)\right)$$

where the loss function $L_\tau(u) = u(\tau - 1(u < 0))$ is the so-called check function and $\mathcal{F}$ is some function class. For $\tau = 1/2$ the loss function is $L_{1/2}(u) = |u|$, and we recover median regression as a special case. One can also choose to add regularization as with conditional mean regression. See Koenker (2005) and Koenker et al. (2017) for an overview of the field.

3.2 Quantile regression based estimator

Based on the conditional quantile function Q we define an approximation $\tilde{F}^{(m)}$ of the conditional distribution function F as follows. We let $\tau_{\min}$ and $\tau_{\max}$ denote fixed quantile levels satisfying $0 < \tau_{\min} < \tau_{\max} < 1$, and we let $q_{\min,z} := Q(\tau_{\min} \mid z) > 0$ and $q_{\max,z} := Q(\tau_{\max} \mid z) < 1$ denote the corresponding conditional quantiles.

Let $\mathcal{T} = [\tau_{\min}, \tau_{\max}]$ denote the set of potential quantile levels. A grid in $\mathcal{T}$ is a sequence $(\tau_k)_{k=1}^m$ such that $\tau_{\min} = \tau_1 < \cdots < \tau_m = \tau_{\max}$ for $m \geq 2$. An equidistant grid is a grid $(\tau_k)_{k=1}^m$ for which $\tau_{k+1} - \tau_k$ is constant for $k = 1, \ldots, m-1$. Also let $\tau_0 = 0$ and $\tau_{m+1} = 1$ be fixed.

Given a grid $(\tau_k)_{k=1}^m$ we let $q_{k,z} := Q(\tau_k \mid z)$ for $k = 1, \ldots, m$ and define $q_{0,z} := 0$ and $q_{m+1,z} := 1$. For each $z \in [0,1]^d$ we define a function $\tilde{F}^{(m)}(\cdot \mid z) : [0,1] \to [0,1]$ by linear interpolation of the points $(q_{k,z}, \tau_k)_{k=0}^{m+1}$:

$$\tilde{F}^{(m)}(t \mid z) := \sum_{k=0}^{m} \left(\tau_k + (\tau_{k+1} - \tau_k) \frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} \right) 1_{(q_{k,z}, q_{k+1,z}]}(t). \tag{1}$$

Let $Q(\mathcal{T} \mid z) = [q_{\min,z}, q_{\max,z}]$ be the range of conditional $\mathcal{T}$-quantiles in the conditional distribution $X \mid Z = z$ for $z \in [0,1]^d$, and define the supremum norm

$$\|f\|_{\mathcal{T},\infty} = \sup_{z \in [0,1]^d} \sup_{t \in Q(\mathcal{T}|z)} |f(t,z)|$$

for a function $f : [0,1] \times [0,1]^d \to \mathbb{R}$. Note that this is a norm on the set of bounded functions on $\{(t,z) \mid z \in [0,1]^d, t \in Q(\mathcal{T} \mid z)\}$. Then we have the following approximation result.

Proposition 1 *Denote by $\tilde{F}^{(m)}$ the function (1) defined from a grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$. Then it holds that*

$$||F - \tilde{F}^{(m)}||_{\mathcal{T},\infty} \leq \kappa_m$$

where $\kappa_m := \max_{k=1,\ldots,m-1}(\tau_{k+1} - \tau_k)$ is the coarseness of the grid.

Choosing a finer and finer grid yields $\kappa_m \to 0$, which implies that $\tilde{F}^{(m)} \to F$ in the norm $\|\cdot\|_{\mathcal{T},\infty}$ for $m \to \infty$.

By an estimator of the conditional distribution function F we mean a mapping from a sample $(X_i, Z_i)_{i=1}^n$ to a function $\hat{F}^{(n)}(\cdot \mid z) : [0,1] \to [0,1]$ such that for every $z \in [0,1]^d$ it holds that $t \mapsto \hat{F}^{(n)}(t \mid z)$ is continuous and increasing with

$$\hat{F}^{(n)}(0 \mid z) = 0 \quad \text{and} \quad \hat{F}^{(n)}(1 \mid z) = 1.$$

Motivated by (1) we define the following estimator of the conditional distribution function.

Definition 2 *Let $(\tau_k)_{k=1}^m$ be a grid in $\mathcal{T}$. Define $\hat{q}_{0,z}^{(n)} := 0$ and $\hat{q}_{m+1,z}^{(n)} := 1$, and let $\hat{q}_{k,z}^{(n)} := \hat{Q}^{(n)}(\tau_k \mid z)$ for $k = 1, \ldots, m$ be the predictions of a quantile regression model obtained from an i.i.d. sample $(X_i, Z_i)_{i=1}^n$. We define the estimator $\hat{F}^{(m,n)}(\cdot \mid z) : [0,1] \to [0,1]$ by*

$$\hat{F}^{(m,n)}(t \mid z) := \sum_{k=0}^{m} \left(\tau_k + (\tau_{k+1} - \tau_k) \frac{t - \hat{q}_{k,z}^{(n)}}{\hat{q}_{k+1,z}^{(n)} - \hat{q}_{k,z}^{(n)}} \right) 1_{\left(\hat{q}_{k,z}^{(n)}, \hat{q}_{k+1,z}^{(n)}\right]}(t) \tag{2}$$

for each $z \in [0,1]^d$.

Note that the estimator is not monotone in the presence of quantile crossing (He, 1997). In this case we perform a re-arrangement of the estimated conditional quantiles in order to obtain monotonicity for finite sample size (Chernozhukov et al., 2010). However, the estimated conditional quantiles will be ordered correctly under the consistency assumptions that we will introduce in Assumption 1, that is, the re-arrangement becomes unnecessary, and the estimator becomes monotone with high probability as $n \to \infty$ for any grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$.

3.3 Pointwise consistency of $\hat{F}^{(m,n)}$

We will now demonstrate how pointwise consistency of the proposed estimator over a set of distributions $\mathcal{P}_0 \subset \mathcal{P}$ can be obtained under the assumption that the quantile regression procedure is pointwise consistent over $\mathcal{P}_0$.

We will evaluate the consistency of $\hat{F}^{(m,n)}$ according to the supremum norm $||\cdot||_{\mathcal{T},\infty}$ introduced in Section 3.2, that is, we restrict the supremum to be over $t \in Q(\mathcal{T} \mid z)$ and not the entire interval $[0,1]$. We do so because quantile regression generally does not give uniform consistency of all extreme quantiles, and in Section 4 we show how consistency of $\hat{F}^{(m,n)}$ between the conditional $\tau_{\min}$- and $\tau_{\max}$-quantiles is sufficient for conditional independence testing.

First, we have the following key corollary of Proposition 1, which is a simple application of the triangle inequality.

Corollary 3 *Let $\tilde{F}^{(m)}$ and $\hat{F}^{(m,n)}$ be given by* (1) *and* (2)*, respectively. Then*

$$\|F - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty} \leq \kappa_m + \|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty}$$

for all grids $(\tau_k)_{k=1}^m$ in $\mathcal{T}$.

The random part of the right hand side of the inequality is the term $\|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty}$, while κ_m is deterministic and only depends on the choice of grid $(\tau_k)_{k=1}^m$. Controlling the term $\|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty}$ is an easier task than controlling $\|F - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty}$ directly because $\tilde{F}^{(m)}$ and $\hat{F}^{(m,n)}$ are piecewise linear, while F is only assumed to be continuous and increasing.

Consistency assumptions on the quantile regression procedure will allow us to show consistency of the estimator $\hat{F}^{(m,n)}$. Let the random variable

$$\mathcal{D}_{\mathcal{T}}^{(n)} := \sup_{z \in [0,1]^d} \sup_{\tau \in \mathcal{T}} |Q(\tau \mid z) - \hat{Q}^{(n)}(\tau \mid z)|$$

denote the uniform prediction error of a fitted quantile regression model, $\hat{Q}^{(n)}$, over the set of quantile levels $\mathcal{T} = [\tau_{\min}, \tau_{\max}]$. Below we write $X_n \in \mathcal{O}_P(a_n)$ when X_n is big-O in probability of a_n with respect to P. See Appendix B for the formal definition.

Assumption 1 *For each $P \in \mathcal{P}_0$ there exist*

(i) a deterministic rate function g_P tending to zero as $n \to \infty$ such that $\mathcal{D}_{\mathcal{T}}^{(n)} \in \mathcal{O}_P(g_P(n))$

(ii) and a finite constant C_P such that the conditional density $f_{X|Z}$ satisfies

$$\sup_{x \in [0,1]} f_{X|Z}(x \mid z) \leq C_P$$

for almost all $z \in [0,1]^d$.

Assumption 1 (i) is clearly necessary to achieve consistency of the estimator. Assumption 1 (ii) is a regularity condition that is used to ensure that $q_{k+1,z} - q_{k,z}$ does not tend to zero too fast as $\kappa_m \to 0$. We now have:

Proposition 4 *Let Assumption 1 be satisfied. Then*

$$\|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty} \in \mathcal{O}_P(g_P(n))$$

for each fixed $P \in \mathcal{P}_0$ and all equidistant grids $(\tau_k)_{k=1}^m$ in $\mathcal{T}$.

Consider letting the number of grid points m_n depend on the sample size n. By combining Corollary 3 and Proposition 4 we obtain the main pointwise consistency result.

Theorem 5 *Let Assumption 1 be satisfied. Then*

$$\|F - \hat{F}^{(m_n,n)}\|_{\mathcal{T},\infty} \in \mathcal{O}_P(g_P(n))$$

for each fixed $P \in \mathcal{P}_0$ given that the equidistant grids $(\tau_k)_{k=1}^{m_n}$ in $\mathcal{T}$ satisfy $\kappa_{m_n} \in o(g_P(n))$.

This shows that $\hat{F}^{(m_n,n)}$ is pointwise consistent over $\mathcal{P}_0$ given that the quantile regression procedure is pointwise consistent over $\mathcal{P}_0$. Moreover, we can transfer the rate of convergence g_P directly. In Section 4.4 we will use this type of pointwise consistency to show asymptotic pointwise level and power of our conditional independence test over $\mathcal{P}_0$.

Note that we can estimate conditional distribution functions in settings with high dimensional covariates to the extend that the quantile regression estimation procedure can deal with high dimensionality. An example of such a procedure is given in Section 3.5.

We chose to state Proposition 4 and Theorem 5 for equidistant grids only, but in the proof of Proposition 4 we only need that the ratio κ_m/γ_m between the coarseness κ_m and the smallest subinterval $\gamma_m = \min_{k=1,\dots,m-1}(\tau_{k+1} - \tau_k)$ must not diverge as $m \to \infty$. This is obviously ensured for an equidistant grid. Moreover, for an equidistant grid, $\kappa_m = (\tau_{\max} - \tau_{\min})/(m-1)$, and $\kappa_{m_n} \in o(g_P(n))$ if m_n grows with rate at least $g_P(n)^{-(1+\varepsilon)}$ for some $\varepsilon > 0$. Since the rate is unknown in practical applications we choose m to be the smallest integer larger than $\sqrt{n}$ as a rule of thumb, since this represents the optimal parametric rate.

3.4 Uniform consistency of $\hat{F}^{(m,n)}$

The pointwise consistency result of Theorem 5 can be extended to a uniform consistency over $\mathcal{P}_0$ by strengthening Assumption 1 to hold uniformly. Below we write $X_n \in \mathcal{O}_{\mathcal{M}}(a_n)$ when X_n is big-O in probability of a_n uniformly over a set of distributions $\mathcal{M}$. We refer to Appendix B for the formal definition.

Assumption 2 *For $\mathcal{P}_0 \subset \mathcal{P}$ there exist*

(i) a deterministic rate function g tending to zero as $n \to \infty$ such that $\mathcal{D}_{\mathcal{T}}^{(n)} \in \mathcal{O}_{\mathcal{P}_0}(g(n))$

(ii) and a finite constant C such that the conditional density $f_{X|Z}$ satisfies

$$\sup_{x \in [0,1]} f_{X|Z}(x \mid z) \leq C$$

for almost all $z \in [0,1]^d$.

With this stronger assumption we have a uniform extension of Proposition 4.

Proposition 6 *Let Assumption 2 be satisfied. Then*

$$\|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty} \in \mathcal{O}_{\mathcal{P}_0}(g(n)).$$

for all equidistant grids $(\tau_k)_{k=1}^m$ in $\mathcal{T}$.

We can now combine Corollary 3 with the stronger Proposition 6 to obtain the following uniform consistency of the estimator $\hat{F}^{(m,n)}$.

Theorem 7 *Suppose that Assumption 2 is satisfied. Then*

$$\|F - \hat{F}^{(m_n,n)}\|_{\mathcal{T},\infty} \in \mathcal{O}_{\mathcal{P}_0}(g(n))$$

given that the equidistant grids $(\tau_k)_{k=1}^{m_n}$ in $\mathcal{T}$ satisfy $\kappa_{m_n} \in o(g(n))$.

This shows that our estimator $\hat{F}^{(m,n)}$ can achieve uniform consistency over a set of distributions $\mathcal{P}_0 \subset \mathcal{P}$ given that the quantile regression procedure is uniformly consistent over $\mathcal{P}_0$. In Section 4.5 we show how this strenghtened result can be used to establish asymptotic uniform level and power of our conditional independence test over $\mathcal{P}_0$.

3.5 A quantile regression model

In this section we will provide an example of a flexible quantile regression model and estimation procedure where consistency results are available. Consider the model

$$Q(\tau \mid z) = h(z)^T \beta_\tau \tag{3}$$

where $h : [0,1]^d \to \mathbb{R}^p$ is a known and continuous transformation of Z, e.g., a polynomial or spline basis expansion to model non-linear effects. Inference in the model (3) was analyzed by Belloni and Chernozhukov (2011) and Belloni et al. (2019) in the high-dimensional

Testing Conditional Independence via Quantile Regression

setup $p \gg n$. In the following we describe a subset of their results that is relevant for our application. Given an i.i.d. sample $(X_i, Z_i)_{i=1}^n$ and a fixed quantile regression level $\tau \in (0,1)$, estimation of $\beta_\tau \in \mathbb{R}^p$ is carried out by penalized regression:

$$\hat{\beta}_\tau \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n L_\tau(X_i - h(Z_i)^T \beta) + \lambda_\tau \|\beta\|_1 \tag{4}$$

where $L_\tau(u) = u(\tau - 1(u < 0))$ is the check function, $\|\cdot\|_1$ is the 1-norm and $\lambda_\tau \geq 0$ is a tuning parameter that determines the degree of penalization. The tuning parameter λ_τ for a set $\mathcal{Q}$ of quantile regression levels can be chosen in a data driven way as follows (Belloni and Chernozhukov, 2011, Section 2.3). Let $W_i = h(Z_i)$ denote the transformed predictors for $i = 1, \ldots, n$. Then we set

$$\lambda_\tau = c\lambda\sqrt{\tau(1-\tau)} \tag{5}$$

where $c > 1$ is a constant with recommended value $c = 1.1$ and λ is the $(1 - n^{-1})$-quantile of the random variable

$$\sup_{\tau \in \mathcal{T}} \frac{\|\Gamma^{-1} \frac{1}{n} \sum_{i=1}^n (\tau - 1(U_i \leq \tau) W_i)\|_\infty}{\sqrt{\tau(1-\tau)}}$$

where $U_1, \ldots, U_n$ are i.i.d. $\mathcal{U}[0,1]$. Here $\Gamma \in \mathbb{R}^{p \times p}$ is a diagonal matrix with $\Gamma_{kk} = \frac{1}{n}\sum_{i=1}^n (W_i)_k^2$. The value of λ is determined by simulation.

Sufficient regularity conditions under which the above estimation procedure can be proven to be consistent are as follows.

Assumption 3 *Denote by $f_{X|Z}$ the conditional density of X given Z. Let $c > 0$ and $C > 0$ be constants.*

(i) There exists s such that $\|\beta_\tau\|_0 \leq s$ for all $\tau \in \mathcal{Q} := [c, 1-c]$.

(ii) $f_{X|Z}$ is continuously differentiable such that $f_{X|Z}(Q_{X|Z}(\tau \mid z) \mid z) \geq c$ for each $\tau \in \mathcal{Q}$ and $z \in [0,1]^d$. Moreover, $\sup_{x \in [0,1]} f_{X|Z}(x \mid z) \leq C$ and $\sup_{x \in [0,1]} \partial_x f_{X|Z}(x \mid z) \leq C$.

(iii) The transformed predictor $W = h(Z)$ satisfies $c \leq E((W^T\theta)^2) \leq C$ for all $\theta \in \mathbb{R}^p$ with $\|\theta\|_2 = 1$. Moreover, $(E(\|W\|_\infty^{2q}))^{1/(2q)} \leq M_n$ for some $q > 2$ where M_n satisfies

$$M_n^2 \leq \frac{\delta_n n^{1/2 - 1/q}}{s\sqrt{\log(p \vee n)}}$$

and δ_n is some sequence tending to zero.

Assumption 3 (i) is a sparsity assumption, (ii) is a regularity condition on the conditional distribution, while (iii) is an assumption on the predictors. Examples of distributions for which Assumption 3 is satisfied are given in Belloni and Chernozhukov (2011) Section 2.5. This includes location models with Gaussian noise and location-scale models with bounded covariates, which in our setup with $Z \in [0,1]^d$ means all location-scale models.

The following result (Belloni and Chernozhukov, 2011, Section 2.6) regarding the estimator $\hat{\beta}_\tau$ then holds.

Theorem 8 *Assume that the tuning parameters* $\{\lambda_\tau \mid \tau \in \mathcal{Q}\}$ *have been chosen according to* (5)*. Then*

$$\sup_{\tau \in \mathcal{Q}} \|\beta_\tau - \hat{\beta}_\tau\|_2 \in \mathcal{O}_P\left(\sqrt{\frac{s \log(p \vee n)}{n}}\right)$$

under Assumption 3.

As a corollary of this consistency result we have the following.

Corollary 9 *Let* $\hat{Q}(\tau \mid z) = h(z)^T \hat{\beta}_\tau$ *be the predicted conditional quantile using the estimator* $\hat{\beta}_\tau$*. Then*

$$\sup_{z \in [0,1]^d} \sup_{\tau \in \mathcal{Q}} |Q(\tau \mid z) - \hat{Q}^{(n)}(\tau \mid z)| \in \mathcal{O}_P\left(\sqrt{\frac{s \log(p \vee n)}{n}}\right)$$

under Assumption 3.

This shows that Assumption 1 is satisfied under the model (3) whenever Assumption 3 is satisfied with $\mathcal{T} \subset \mathcal{Q}$ and $\sqrt{s \log(p \vee n)/n} \to 0$, which is the key underlying assumption of Theorem 5. Note also that Assumption 1 (ii) is contained in Assumption 3 (ii). Theorem 8 and Corollary 9 can be extended to hold uniformly over $\mathcal{P}_0 \subset \mathcal{P}$ by assuming that the conditions of Assumption 3 hold uniformly over $\mathcal{P}_0$. This then gives the statement of Assumption 2 that is required for Theorem 7.

4. Testing Conditional Independence

In this section we describe the conditional independence testing framework in terms of the so-called partial copula. As above we let $(X, Y, Z) \sim P \in \mathcal{P}$ such that $X, Y \in [0,1]$ and $Z \in [0,1]^d$ where $\mathcal{P}$ are the distributions that are absolutely continuous with respect to Lebesgue measure on $[0,1]^{2+d}$. Also let f denote a generic density function. We then say that X is conditionally independent of Y given Z if

$$f(x, y \mid z) = f(x \mid z) f(y \mid z)$$

for almost all $x, y \in [0,1]$ and $z \in [0,1]^d$. See e.g. Dawid (1979). In this case we write that $X \perp\!\!\!\perp_P Y \mid Z$, where we usually omit the dependence on P when there is no ambiguity. By $\mathcal{H} \subset \mathcal{P}$ we denote the subset of distributions for which conditional independence is satisfied, and we let $\mathcal{Q} := \mathcal{P} \setminus \mathcal{H}$ be the alternative of conditional dependence.

4.1 The partial copula

We can regard the conditional distribution function as a mapping $(t, z) \mapsto F(t \mid z)$ for $t \in [0,1]$ and $z \in [0,1]^d$. Assuming that this mapping is measurable, we define a new pair of random variables U_1 and U_2 by the transformations

$$U_1 := F_{X|Z}(X \mid Z) \quad \text{and} \quad U_2 := F_{Y|Z}(Y \mid Z).$$

This transformation is usually called the probability integral transformation or Rosenblatt transformation due to Rosenblatt (1952), where the transformation was initially introduced and the following key result was shown.

Proposition 10 *It holds that* $U_\ell \sim \mathcal{U}[0,1]$ *and* $U_\ell \perp\!\!\!\perp Z$ *for* $\ell = 1, 2$.

Hence the transformation can be understood as a normalization, where marginal dependencies of X on Z and Y on Z are filtered away. The joint distribution of U_1 and U_2 has been termed the partial copula of X and Y given Z in the copula literature (Bergsma, 2011; Spanhel and Kurz, 2016). Independence in the partial copula relates to conditional independence in the following way.

Proposition 11 *If* $X \perp\!\!\!\perp Y \mid Z$ *then* $U_1 \perp\!\!\!\perp U_2$.

Therefore the question about conditional independence can be transformed into a question about independence. Note, however, that $U_1 \perp\!\!\!\perp U_2$ does not in general imply that $X \perp\!\!\!\perp Y \mid Z$. See Property 7 in Spanhel and Kurz (2016) for a counterexample

The variables U_ℓ were termed nonparametric residuals by Patra et al. (2016) due to the independence property $U_\ell \perp\!\!\!\perp Z$ which is analogues to the property of conventional residuals in additive Gaussian noise models. Note that the entire conditional distribution function is required in order to compute the nonparametric residual, while conventional residuals in additive noise models are computed using only the conditional expectation. In return, Proposition 10 provides the distribution of the nonparametric residuals without asumming any functional or distributional relationship between X (Y resp.) and Z, whereas the distribution of conventional residuals is not known without further assumptions. Moreover, the nonparametric residuals U_1 and U_2 are independent under conditional independence, while conventional residuals are only uncorrelated unless we make a Gaussian assumption, say.

4.2 Generic testing procedure

Suppose $(X_i, Y_i, Z_i)_{i=1}^n$ is a sample from $P \in \mathcal{P}_0$ where $\mathcal{P}_0$ is some subset of $\mathcal{P}$. Also let $\mathcal{H}_0 := \mathcal{P}_0 \cap \mathcal{H}$ and $\mathcal{Q}_0 := \mathcal{P}_0 \cap \mathcal{Q}$ be the distributions in $\mathcal{P}_0$ satisfying conditional independence and conditional dependence, respectively. Denote by

$$U_{1,i} := F_{X|Z}(X_i \mid Z_i) \quad \text{and} \quad U_{2,i} := F_{Y|Z}(Y_i \mid Z_i)$$

the nonparametric residuals for $i = 1, \ldots, n$. Let $\psi_n : [0,1]^{2n} \to \{0, 1\}$ denote a test for independence in a bivariate continuous distribution. The observed value of the test is

$$\Psi_n := \psi_n((U_{1,i}, U_{2,i})_{i=1}^n)$$

with $\Psi_n = 0$ indicating acceptance and $\Psi_n = 1$ rejection of the hypothesis. By Proposition 11 we then reject the hypothesis of conditional independence, $X \perp\!\!\!\perp Y \mid Z$, if $\Psi_n = 1$. However, in order to implement the test in practice, we will need to replace the conditional distribution functions $F_{X|Z}$ and $F_{Y|Z}$ by estimates.

Given some generic estimators of the conditional distribution functions we can formulate a generic version of the partial copula conditional independence test as follows.

Definition 12 *Let $(X_i, Y_i, Z_i)_{i=1}^n$ be an i.i.d. sample from $P \in \mathcal{P}_0$. Also let ψ_n be a test for independence in a bivariate continuous distribution.*

(i) Form the estimates $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$ based on $(X_i, Y_i, Z_i)_{i=1}^n$.

(ii) Compute the estimated nonparametric residuals

$$\hat{U}^{(n)}_{1,i} := \hat{F}^{(n)}_{X|Z}(X_i \mid Z_i) \quad \textit{and} \quad \hat{U}^{(n)}_{2,i} := \hat{F}^{(n)}_{Y|Z}(Y_i \mid Z_i)$$

for $i = 1, \ldots, n$.

(iii) Let $\hat{\Psi}_n := \psi_n\left((\hat{U}^{(n)}_{1,i}, \hat{U}^{(n)}_{2,i})_{i=1}^n\right)$ and reject the hypothesis $X \perp\!\!\!\perp Y \mid Z$ of conditional independence if $\hat{\Psi}_n = 1$.

This generic version of the conditional independence test is analogous to the approach of Bergsma (2011), but here we emphasize the modularity of the testing procedure. Firstly, one can use any method for estimating conditional distribution functions. Secondly, any test for independence in a bivariate continuous distribution can be utilized.

We note that under the assumptions of Theorem 5 it holds that

$$|(\hat{U}^{(n)}_{1,i}, \hat{U}^{(n)}_{2,i}) - (U_{1,i}, U_{2,i})|_{\mathcal{T},1} \xrightarrow{P} 0$$

where $|u - v|_{\mathcal{T},1} = |u_1 - v_1| 1(u_1, v_1 \in \mathcal{T}) + |u_2 - v_2| 1(u_2, v_2 \in \mathcal{T})$. That is, each estimated pair of nonparametric residuals has the partial copula as asymptotic distribution – except perhaps on the fringe part of the unit square outside of $\mathcal{T} \times \mathcal{T}$. This is a priori only a marginal result for each i, but it suggests that tests based on the estimated residuals behave as if they were i.i.d. observations from the partial copula.

Once we have chosen the test for independence, ψ_n, we can establish rigorous results on the properties of the test over the space of hypotheses $\mathcal{H}_0$ and alternatives $\mathcal{Q}_0$, but how exactly to transfer the consistency of the estimated residuals to results on level and power depends on the specific test statistic. We will in the following sections demonstrate this transfer for one particular class of test statistics.

4.3 Generalized measure of correlation

We will now introduce a generalized measure of correlation that will form the basis for an independence test between the nonparametric residuals U_1 and U_2.

Definition 13 *The generalized correlation, ρ, between U_1 and U_2 is defined in term of a multivariate function $\varphi = (\varphi_1, \ldots, \varphi_q) : [0,1] \to \mathbb{R}^q$ as*

$$\rho = E_P(\varphi(U_1)\varphi(U_2)^T) \tag{6}$$

such that ρ is a $q \times q$ matrix with entries $\rho_{k\ell} = E_P(\varphi_k(U_1)\varphi_\ell(U_2))$ for $k, \ell = 1, \ldots, q$.

We will assume that the function $\varphi = (\varphi_1, \ldots, \varphi_q)$ defining the generalized correlation satisfies the following assumptions for the remainder of the paper.

Assumption 4

(i) The support $\mathcal{T}_k$ of each coordinate function φ_k is a compact subset of $(0,1)$.

(ii) Each coordinate function $\varphi_k : [0,1] \to \mathbb{R}$ is Lipschitz continuous.

(iii) $\int_0^1 \varphi_k(u)\mathrm{d}u = 0$ and $\int_0^1 \varphi_k(u)^2\mathrm{d}u = 1$ for each $k = 1, \dots, q$.

(iv) The coordinate functions $\varphi_1, \dots, \varphi_q$ are linearly independent.

Let us provide some intuition about the interpretation of the generalized correlation ρ and explain the role of the assumptions on φ in Assumption 4.

Each entry $\rho_{k\ell}$ can be interpreted as an expected conditional correlation, and it can be understood in terms of the partial and conditional copula (Patton, 2006). Let $C(u_1, u_2 \mid z) = F(U_1 \leq u_1, U_2 \leq u_2 \mid Z = z)$ denote the conditional copula of X and Y given $Z = z$. Then the partial copula is the expected conditional copula, i.e., $C_p(u_1, u_2) = E_P(C(u_1, u_2 \mid Z))$ (Spanhel and Kurz, 2016). The conditional generalized correlation, $\rho_{k\ell}(z)$, between X and Y given $Z = z$ can be expressed in terms of the conditional copula by

$$\rho_{k\ell}(z) := E_P(\varphi_k(U_1)\varphi_\ell(U_2) \mid Z = z) = \int \varphi_k(u_1)\varphi_\ell(u_2)C(\mathrm{d}u_1, \mathrm{d}u_2 \mid z).$$

By the tower property of conditional expectations, $\rho_{k\ell}$ can be represented as an expected generalized correlation

$$\rho_{k\ell} = E_P(\rho_{k\ell}(Z)) = \int \varphi_k(u_1)\varphi_\ell(u_2)C_p(\mathrm{d}u_1, \mathrm{d}u_2).$$

Hence $\rho_{k\ell}$ measures the expected conditional generalized correlation of X and Y given Z w.r.t. the distribution of Z. Importantly, Assumption 4 (iii) implies that

$$\rho = E_P(\varphi(U_1)\varphi(U_2)^T) = E_P(\varphi(U_1))E(\varphi(U_2)^T) = 0$$

whenever $X \perp\!\!\!\perp Y \mid Z$ due to Proposition 11.

The purpose of Assumption 4 (i) is twofold. Firstly, letting the supports $\mathcal{T}_k$ and $\mathcal{T}_\ell$ of φ_k and φ_ℓ be subsets of $(0,1)$ implies that $\rho_{k\ell}$ focuses on dependence in the compact region $\mathcal{T}_k \times \mathcal{T}_\ell \subset (0,1)^2$ of the outcome space $[0,1]^2$ of (U_1, U_2). For $q \geq 2$ the generalized correlation ρ thus summarizes dependencies in different regions of the outcome space. See Figure 1 for an illustration of this idea. Secondly, the supports $(\mathcal{T}_k)_{k=1}^q$ will play the role as subsets of the possible quantile levels $\mathcal{T} = [\tau_{\min}, \tau_{\max}]$, when we choose the conditional distribution function estimators based on quantile regression from Section 3.2. This connection will be made clear in Section 4.4.

The functional form of φ_k and φ_ℓ determines the kind of dependence measured by $\rho_{k\ell}$. Ignoring Assumption 4 (i), consider letting $\varphi_k(u) = \varphi_\ell(u) = \sqrt{12}(u - 1/2)$ for $u \in [0,1]$. Then $\rho_{k\ell}$ measures the expected conditional Spearman correlation between X and Y given Z with respect to the distribution of Z. In Section 4.6 we describe a choice of functions φ_k that leads to a trimmed version of expected conditional Spearman correlation which satisfies Assumption 4 (i). As we shall see in Section 4.4, Assumption 4 (ii), i.e., that the coordinate functions φ_k are Lipschitz continuous, is crucial for deriving asymptotic properties for the empirical version of the generalized correlation ρ. Lastly, we assume that $\varphi_1, \dots, \varphi_q$ are linearly independent in Assumption 4 (iv) to avoid degeneracy of its empirical version, which we introduce in Section 4.4.

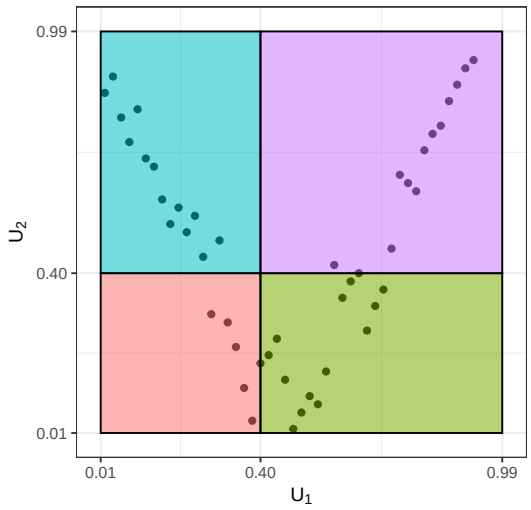

Figure 1: A sample from a copula (U_1, U_2) with clear dependence, but where the overall sample correlation is close to zero. The dependence is captured by considering sample correlation of observations in different regions of the outcome space.

4.4 Test based on generalized correlation

In this section we will analyze in depth the conditional independence test resulting from basing the test ψ_n in Definition 12 on the generalized correlation ρ. We will formulate the results in terms of a generic method for estimating conditional distribution functions in order to emphasize the generality of the method and illustrate the abstract assumptions needed for the test to be sound. Along the way we will explain when the assumptions are satisfied for the quantile regression based estimator $\hat{F}^{(m,n)}$ that we developed in Section 3.

With ρ the generalized correlation between U_1 and U_2 defined in terms of a function φ satisfying Assumption 4 we let $\rho_n : [0,1]^{2n} \to \mathbb{R}^{q \times q}$ be its corresponding empirical version:

$$\rho_n((u_i, v_i)_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n \varphi(u_i)\varphi(v_i)^T. \tag{7}$$

Soundness of a test based on ρ_n depends on consistency of the estimators $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$. Recall that we by $\mathcal{T}_1, \ldots, \mathcal{T}_q$ denote the supports of $\varphi_1, \ldots, \varphi_q$. Let $\tau_{\min} := \inf(\mathcal{T}_1 \cup \cdots \cup \mathcal{T}_q) > 0$ and $\tau_{\max} := \sup(\mathcal{T}_1 \cup \cdots \cup \mathcal{T}_q) < 1$, and then define $\mathcal{T} := [\tau_{\min}, \tau_{\max}]$. As in Section 3.2 we let the norm $\|\cdot\|_{\mathcal{T},\infty}$ be given by

$$\|f(t,z)\|_{\mathcal{T},\infty} = \sup_{z \in [0,1]^d} \sup_{t \in Q_{X|Z}(\mathcal{T}|z)} |f(t,z)|$$

when X given Z is the conditional distribution of interest. Similarly define $\|\cdot\|'_{\mathcal{T},\infty}$ by

$$\|f(t,z)\|'_{\mathcal{T},\infty} = \sup_{z \in [0,1]^d} \sup_{t \in Q_{Y|Z}(\mathcal{T}|z)} |f(t,z)|.$$

Testing Conditional Independence via Quantile Regression

Then we have the following assumption on our estimators.

Assumption 5 *For each distribution $P \in \mathcal{P}_0$ there exist deterministic rate functions g_P and h_P tending to zero as $n \to \infty$ and functions $\xi, \xi' : [0,1] \times [0,1]^d \to \mathbb{R}$ such that*

(i) $\|F_{X|Z} - \hat{F}^{(n)}_{X|Z}\|_{\mathcal{T},\infty} \in \mathcal{O}_P(g_P(n))$ *and* $\|F_{Y|Z} - \hat{F}^{(n)}_{Y|Z}\|'_{\mathcal{T},\infty} \in \mathcal{O}_P(h_P(n))$.

(ii) $||\xi - \hat{F}^{(n)}_{X|Z}||_{\mathcal{T}^c,\infty} \in \mathcal{O}_P(g_P(n))$ *and* $||\xi' - \hat{F}^{(n)}_{Y|Z}||'_{\mathcal{T}^c,\infty} \in \mathcal{O}_P(h_P(n))$.

Assumption 5 (i) states that our estimators $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$ are consistent with rates g_P and h_P over the conditional $\mathcal{T}$-quantiles in their respective conditional distributions. This is the result of Theorem 5 regarding our quantile regression based estimator $\hat{F}^{(m,n)}$ when $\mathcal{T}$ as above is taken as the set of potential quantile regression levels.

Assumption 5 (ii) is an assumption on the behavior of our estimator in the tails of the conditional distribution, i.e., over the conditional $\mathcal{T}^c$-quantiles. Here we do not assume consistency, but we do assume that the limit for $n \to \infty$ exists, and that our estimators are convergent to their limits with rates g_P and h_P respectively. This assumption is satisfied by our quantile regression based estimator $\hat{F}^{(m,n)}$ whenever it satisfies Assumption 5 (i).

With this assumption we first establish the asymptotic distribution of the test statistic

$$\hat{\rho}_n := \rho_n\left((\hat{U}_{1,i}, \hat{U}_{2,i})_{i=1}^n\right) = \frac{1}{n}\sum_{i=1}^n \varphi(\hat{U}_{1,i})\varphi(\hat{U}_{2,i})^T \tag{8}$$

under the hypothesis of conditional independence. Below we use $\Rightarrow_P$ to denote convergence in distribution with respect to P.

Theorem 14 *Suppose that Assumption 5 is satisfied with rate functions g_P and h_P such that $\sqrt{n}g_P(n)h_P(n) \to 0$ as $n \to \infty$ for each $P \in \mathcal{P}_0$. Then the statistic $\hat{\rho}_n$ given by* (8) *satisfies*

$$\sqrt{n}\hat{\rho}_n \Rightarrow_P \mathcal{N}(0, \Sigma \otimes \Sigma)$$

for each fixed $P \in \mathcal{H}_0$. The asymptotic covariance matrix is given by

$$\Sigma_{k,s} = \int_0^1 \varphi_k(u)\varphi_s(u)\mathrm{d}u$$

for $k, s = 1, \ldots, q$ and does not depend on P.

If the rate functions are $g_P(n) = n^{-a}$ and $h_P(n) = n^{-b}$, then we require that $a+b > 1/2$. Thus convergence slightly faster than rate $n^{-1/4}$ for both estimators is sufficient, but there can also be a tradeoff between the rates. Interestingly, Theorem 14 does not require sample splitting for the estimation of the conditional distribution function and computation of the test statistic. This is due to the fact that we are only interested in the asymptotic distribution under conditional independence. A similar phenomenon was found by Shah and Peters (2020), when they proved asymptotic normality of their Generalised Covariance Measure under conditional independence.

According to Corollary 9, the requirement $\sqrt{n} g_P(n) h_P(n) \to 0$ is satisfied for our quantile regression based estimator $\hat{F}^{(m,n)}$, if the quantile regression model (3) of Section 3.5 is valid for both X given Z and Y given Z for some continuous transformations $h_1 : [0,1]^d \to \mathbb{R}^{p_1}$ and $h_2 : [0,1]^d \to \mathbb{R}^{p_2}$ and Assumption 3 is satisfied with $s_1, s_2, p_1, p_2, n \to \infty$ such that

$$\sqrt{n} \cdot \sqrt{\frac{s_1 \log(p_1 \vee n)}{n}} \cdot \sqrt{\frac{s_2 \log(p_2 \vee n)}{n}} = \sqrt{\frac{s_1 s_2 \log(p_1 \vee n) \log(p_2 \vee n)}{n}} \to 0$$

where $s_1 = \sup_{\tau \in \mathcal{T}} \|\beta_{1,\tau}\|_0$ and $s_2 = \sup_{\tau \in \mathcal{T}} \|\beta_{2,\tau}\|_0$ are the sparsities of the model parameters.

With the test statistic

$$T_n := \|\Sigma^{-1/2} \hat{\rho}_n \Sigma^{-1/2}\|_F^2, \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, we have the following corollary of Theorem 14.

Corollary 15 *Let the condition of Theorem 14 be satisfied and let T_n be given by* (9)*. Then it holds that*

$$n T_n \Rightarrow_P \chi^2_{q^2}$$

for each fixed $P \in \mathcal{H}_0$.

In view of Theorem 14 and Corollary 15 we define the following conditional independence test based on the generalized correlation.

Definition 16 *Let $\alpha \in (0,1)$ be a desired significance level and T_n the test statistic* (9)*. Then we let $\hat{\Psi}_n$ be the test given by*

$$\hat{\Psi}_n = 1(T_n > n^{-1} z_{1-\alpha})$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of a $\chi^2_{q^2}$-distribution.

Control of the asymptotic pointwise level is then an easy corollary of Corollary 15.

Corollary 17 *Suppose that Assumption 5 is satisfied with rate functions g_P and h_P such that $\sqrt{n} g_P(n) h_P(n) \to 0$ as $n \to \infty$ for each $P \in \mathcal{P}_0$. Then the test $\hat{\Psi}_n$ given by Definition 16 has asymptotic pointwise level over $\mathcal{H}_0$, i.e.,*

$$\limsup_{n \to \infty} E_P(\hat{\Psi}_n) = \alpha$$

for each fixed $P \in \mathcal{H}_0$.

This shows that the test achieves correct level given consistency of the estimators $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$ with suitably fast rates. To obtain results on power of the test $\hat{\Psi}_n$ we only need to understand how $\hat{\rho}_n$ converges in probability and not its entire asymptotic distribution.

Theorem 18 *The test statistic $\hat{\rho}_n$ given by (8) satisfies*

$$\hat{\rho}_n \xrightarrow{P} \rho$$

for each fixed $P \in \mathcal{P}_0$ under Assumption 5.

One may note that the theorem does not require that the rate functions g_P and h_P converge to zero at a certain rate. Let $\mathcal{A}_0 \subseteq \mathcal{Q}_0$ be the subset of alternatives for which $\rho_{k\ell} \neq 0$ for at least one combination of $k, \ell = 1, \dots, q$. Then we have the following corollary of Theorem 18, which exploits that nT_n diverges in probability whenever $P \in \mathcal{A}_0$.

Corollary 19 *For each level $\alpha \in (0, 1)$ the test $\hat{\Psi}_n$ given by Definition 16 has asymptotic pointwise power against $\mathcal{A}_0$, i.e.,*

$$\liminf_{n\to\infty} E_P(\hat{\Psi}_n) = 1$$

for each fixed $P \in \mathcal{A}_0$ under Assumption 5.

Let us discuss the alternatives the test has power against. Firstly, note that we always have the implications

$$X \perp\!\!\!\perp Y \mid Z \quad \Rightarrow \quad U_1 \perp\!\!\!\perp U_2 \quad \Rightarrow \quad \rho = 0$$

However, none of the reverse implications are in general true. We do, however, have the following result stating a sufficient condition for the reverse implication of the first statement.

Proposition 20 *Assume that $(U_1, U_2) \perp\!\!\!\perp Z$. Then $X \perp\!\!\!\perp Y \mid Z$ if and only if $U_1 \perp\!\!\!\perp U_2$.*

This means that if Z only affects the marginal distributions of X and Y, then independence in the partial copula implies conditional independence. This is known as the simplifying assumption in the copula literature (Gijbels et al., 2015; Spanhel and Kurz, 2015). Naturally, $U_1 \not\perp\!\!\!\perp U_2$ always implies $X \not\perp\!\!\!\perp Y \mid Z$, so the simplifying assumption is not a necessary condition for our test to have power, but it does give some intuition about a subset of distributions for which the partial copula completely characterizes conditional independence. However, an unavoidable limitation of the method is that it can never have power against alternatives for which $U_1 \perp\!\!\!\perp U_2$ but $X \not\perp\!\!\!\perp Y \mid Z$.

Turning to the second implication, Corollary 19 tells us that we have power against alternatives for which $\rho_{k\ell} \neq 0$ for some $k, \ell = 1, \dots, q$. However, not all types of dependencies can be detected in this fashion, and it is possible that $\rho = 0$, while $U_1 \not\perp\!\!\!\perp U_2$. A test based on ρ will not have power against such an alternative. For an abstract interpretation of the generalized correlation ρ we refer to Section 4.3. In Section 4.6 we introduce a concrete generalized correlation and elaborate on its interpretation.

Finally, basing the test on values of T_n is natural since the asymptotic behaviour is readily available through Theorem 14, but other transformations of $\hat{\rho}_n$ could be considered such as taking the coordinatewise absolute maximum $\max_{k,l} |(\Sigma^{-1/2}\hat{\rho}_n\Sigma^{-1/2})_{k,l}| = \|\Sigma^{-1/2}\hat{\rho}_n\Sigma^{-1/2}\|_\infty$.

4.5 Uniform level and power results

The level and power results of Section 4.4 are pointwise over the space of hypotheses and alternatives, i.e., they state level and power of the test when fixing a distribution P. In this section we describe how these results can be extended to hold uniformly by strengthening the statements in Assumption 5 to hold uniformly.

Assumption 6 *For $\mathcal{P}_0 \subset \mathcal{P}$ there exist deterministic rate functions g and h tending to zero as $n \to \infty$ and functions $\xi, \xi' : [0,1] \times [0,1]^d \to \mathbb{R}$ such that*

(i) $\|F_{X|Z} - \hat{F}^{(n)}_{X|Z}\|_{\mathcal{T},\infty} \in \mathcal{O}_{\mathcal{P}_0}(g(n))$ *and* $\|F_{Y|Z} - \hat{F}^{(n)}_{Y|Z}\|'_{\mathcal{T},\infty} \in \mathcal{O}_{\mathcal{P}_0}(h(n))$.

(ii) $||\xi - \hat{F}^{(n)}_{X|Z}||_{\mathcal{T}^c,\infty} \in \mathcal{O}_{\mathcal{P}_0}(g(n))$ *and* $||\xi' - \hat{F}^{(n)}_{Y|Z}||'_{\mathcal{T}^c,\infty} \in \mathcal{O}_{\mathcal{P}_0}(h(n))$.

As before we note that Assumption 6 (i) is the result of Theorem 7 regarding our quantile regression based estimator $\hat{F}^{(m,n)}$. Moreover, Assumption 6 (ii) is valid for $\hat{F}^{(m,n)}$ whenever it satisfies Assumption 6 (i).

We will now describe the extensions of Theorem 14 and Theorem 18 that can be obtained under Assumption 6. Below we write $\Rightarrow_{\mathcal{M}}$ to denote uniform convergence in distribution over a set of distributions $\mathcal{M}$, and we use $\to_{\mathcal{M}}$ to denote uniform convergence in probability over $\mathcal{M}$. We refer to Appendix B for the formal definitions.

Theorem 21 *Let $\hat{\rho}_n$ be the statistic given by* (8)*. Then we have:*

(i) Under Assumption 6 with rate functions satisfying $\sqrt{n}g(n)h(n) \to 0$ it holds that

$$\sqrt{n}\hat{\rho}_n \Rightarrow_{\mathcal{H}_0} \mathcal{N}(0, \Sigma \otimes \Sigma)$$

where Σ is as in Theorem 14.

(ii) Under Assumption 6 it holds that $\hat{\rho}_n \to_{\mathcal{P}_0} \rho$.

As a straightforward corollary of Theorem 21 (i) we get the following uniform level result.

Corollary 22 *The test $\hat{\Psi}_n$ given by Definition 16 has asymptotic uniform level over $\mathcal{H}_0$, i.e.,*

$$\limsup_{n\to\infty} \sup_{P \in \mathcal{H}_0} E_P(\hat{\Psi}_n) = \alpha,$$

given that Assumption 6 is satisfied with $\sqrt{n}g(n)h(n) \to 0$ as $n \to \infty$.

The pointwise power result of Theorem 19 does not extend directly to a uniform version in the same way as the level result. For $\lambda > 0$ we let $\mathcal{A}_\lambda \subset \mathcal{Q}_0$ be the set of alternatives for which $|(\rho_P)_{k\ell}| > \lambda$ for at least one combination of $k, \ell = 1, \ldots, q$, where we emphasize that ρ_P depends on the distribution P. We then have the following uniform power result as a corollary of Theorem 21 (ii).

Corollary 23 *For all fixed levels $\alpha \in (0,1)$ the test $\hat{\Psi}_n$ given by Definition 16 has asymptotic uniform power against $\mathcal{A}_\lambda$ for each $\lambda > 0$, i.e.,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{A}_\lambda} E_P(\hat{\Psi}_n) = 1,$$

under Assumption 6.

The reason we need to restrict to the class of alternatives $\mathcal{A}_\lambda$ for a fixed $\lambda > 0$ is the following. If the infimum is taken over $\mathcal{A}_0$, then there could exist a sequence $(P_m)_{m=1}^\infty \subset \mathcal{A}_0$ of distributions such that $(\rho_{P_m})_{k\ell} \neq 0$ for each $m \geq 1$ but $(\rho_{P_m})_{k\ell} \to 0$ as $m \to \infty$. As a consequence nT_n will not necessarily diverge in probability, which is crucial to the proof of the corollary. However, when restricting to $\mathcal{A}_\lambda$ we are ensured that $\inf_{P\in\mathcal{A}_\lambda} |(\rho_P)_{k\ell}| \geq \lambda > 0$ for at least one combination of $k, \ell = 1, \ldots, q$.

We note that these uniform level and power results are not in contradiction with the impossibility result of Shah and Peters (2020) mentioned in Section 1 because our results apply to a restricted set of distributions, $\mathcal{P}_0$, where the conditional distribution functions are estimable with sufficiently fast rate.

4.6 Trimmed Spearman correlation

We will now define a specific family of functions φ defining the generalized correlation that can be shown to satisfy Assumption 4, which results in trimmed versions of the expected conditional Spearman correlation. As mentioned in Section 4.3, ignoring Assumption 4 (i), we could consider

$$\varphi_k(u) = \varphi_\ell(u) = \sqrt{12}\left(u - \frac{1}{2}\right) \tag{10}$$

for $u \in [0,1]$ which results in $\rho_{k\ell}$ being the expected conditional Spearman correlation of X and Y given Z with respect to the distribution of Z. Drawing inspiration from (10) we define a class of functions $\varphi : [0,1] \to \mathbb{R}^q$ by letting

$$\varphi_k(u) = c_k(u - m_k)\sigma_k(u) \tag{11}$$

such that each $\varphi_k : [0,1] \to \mathbb{R}$ is determined by a Lipschitz continuous function $\sigma_k : [0,1] \to \mathbb{R}$ with the support $\mathcal{T}_k$ of σ_k a compact interval in $(0,1)$, $\int_0^1 \sigma_k(u)\mathrm{d}u = 1$ and

$$m_k = \int u\sigma_k(u)\mathrm{d}u \quad \text{and} \quad c_k = \left(\int (u - m_k)^2\sigma_k(u)^2\mathrm{d}u\right)^{-1/2}.$$

The choice (11) satisfies Assumption 4 (i) – (iii) by construction, and if e.g. $\mathcal{T}_k \setminus \cup_{\ell\neq k}\mathcal{T}_\ell \neq \emptyset$ the functions are also linearly independent. We call the resulting generalized correlation ρ a trimmed Spearman correlation, and we refer to the functions σ_k as trimming functions. Note that if the supports $(\mathcal{T}_k)_{k=1}^q$ of $(\sigma_k)_{k=1}^q$ are choosen to be disjoint, then the covariance matrix Σ of Theorem 14 is the identity matrix.

A starting point for choosing a trimming function σ is the normalized indicator

$$u \mapsto (\lambda - \mu)^{-1} 1_{[\mu,\lambda]}(u) \tag{12}$$

for $u \in [0,1]$ where $0 < \mu < \lambda < 1$ are trimming parameters. However, (12) is not a valid trimming function, since it is not Lipschitz. Therefore, we consider a simple linear approximation $\sigma : [0,1] \to \mathbb{R}$ given by

$$\sigma(u) = Kf(u) \quad \text{and} \quad f(u) = \begin{cases} 1, & u \in [\mu+\delta, \lambda-\delta] \\ 0, & u \in [\mu,\lambda]^c \\ \delta^{-1}(u-\mu), & u \in [\mu, \mu+\delta) \\ \delta^{-1}(\lambda-u), & u \in (\lambda-\delta, \lambda] \end{cases} \tag{13}$$

and $K = (\lambda - \mu - \delta)^{-1}$. Here $0 < \delta < (\lambda-\mu)/2$ is a fixed parameter that determines the accuracy of the approximation. It is elementary to verify that σ given by (13) is a valid trimming function, i.e., σ is Lipschitz continuous with $\int \sigma(u)\mathrm{d}u = 1$ and support $[\mu,\lambda] \subset (0,1)$.

The interpretation of a generalised correlation ρ based on φ of the form (11) with trimming function σ of the form (13) is as follows. The entry $\rho_{k\ell}$ is an approximation of the expected conditional Spearman correlation between the observations of X and Y, respectively, that lie in the $\mathcal{T}_k$-quantile range of the distribution of X given Z and the $\mathcal{T}_\ell$-quantile range of the distribution of Y given Z, respectively, with respect to the distribution of Z. The matrix ρ then summarizes this type of dependence within different quantile ranges of X and Y given Z.

4.7 Practical considerations

Throughout Sections 4.4 and 4.5 we have analyzed our proposed test for conditional independence with an emphasis on modularity of the method regarding the choice of estimators $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$ of the conditional distribution functions and the choice of the function φ that defines the generalized correlation ρ of Section 4.3. This focus on the conceptual assumptions displays the generality of the method, but it also leaves the practitioner of conditional independence testing with a number of choices to be made. In this section we summarize a set of choices to make the method work out-of-the-box.

Throughout the paper we have assumed that all random variables take values in the unit interval, i.e., $(X,Y,Z) \in [0,1]^{d+2}$. This is not a restriction, since if e.g. $X \in \mathbb{R}$ we can always apply a strictly increasing, continuous transformation $t : \mathbb{R} \to [0,1]$ to obtain a new random variable $X' = t(X)$ with values in $[0,1]$. Moreover, the initial conditional independence structure of (X,Y,Z) is preserved since the transformation is marginal on X and bijective. The transformation t can be chosen to be e.g. the logistic function.

In principle, an arbitrary and fixed marginal transformation could be used for all variables, but we recommend to transform data to the unit interval via marginal empirical distribution functions. This results in transformed variables known in the copula literature as pseudo copula observations. The transformation creates dependence, similar to the dependence created by other common preprocessing techniques such as centering and scaling, which the theoretical analysis has not accounted for. We suggest, nevertheless, to use this preprocessing technique in practise, and in the simulation study in Section 5 we use pseudo copula observations since it reflects how a practitioner would transform the variables.

To estimate the conditional distribution functions $\hat{F}^{(m,n)}_{X|Z}$ and $\hat{F}^{(m,n)}_{Y|Z}$ using Definition 2, we suggest choosing $\tau_{\min} = 0.01$ and $\tau_{\max} = 0.99$ and form the equidistant grid $(\tau_k)_{k=1}^m$ in $\mathcal{T} = [\tau_{\min}, \tau_{\max}]$ with the number of gridpoints $m = \lceil \sqrt{n} \rceil$. We then suggest using a model of the form (3) for both the quantile regression model $Q_{X|Z}(\tau_k \mid \cdot)$ and $Q_{Y|Z}(\tau_k \mid \cdot)$ for each $k = 1, \dots, m$, where the bases h_1 and h_2 can be chosen to be e.g. an additive B-spline basis for each component of Z.

To test the hypothesis of conditional independence we suggest using the $\hat{\Psi}_n$ from Definition 16 based on the estimated nonparametric residuals $(\hat{U}_{1,i}, \hat{U}_{2,i})_{i=1}^n$. To this end we choose $q \geq 1$ and let $\tau_{\min} = \lambda_0 < \cdots < \lambda_q = \tau_{\max}$ be an equidistant grid in $\mathcal{T}$. We then define the trimming function σ_k to have the form (13) with trimming parameters λ_k and λ_{k+1} and approximation parameter $\delta = 0.01 \cdot (\lambda_{k+1} - \lambda_k)$ for each $k = 0, \dots, q-1$. We then define $(\sigma_k)_{k=1}^q$ according to (11), compute the test statistic $\hat{\rho}_n$ using (8) and compute $\hat{\Psi}_n$ as in Definition 16 using a desired significance level $\alpha \in (0, 1)$.

There are two non-trivial choices remaining. The first is the choice of bases h_1 and h_2 for the quantile regression models $Q_{X|Z}(\tau_k \mid z) = h_1(z)^T \beta_{\tau_k}$ and $Q_{Y|Z}(\tau_k \mid z) = h_2(z)^T \beta_{\tau_k}$. Here the practitioner needs to make a qualified model selection. We recommend using a flexible basis such as an additive B-spline basis, and perform penalized estimation using (4) to avoid overfitting. The second choice is the dimension of the generalized correlation $q \geq 1$, which corresponds to a choice of independence test in the partial copula. Note that the generalized correlation ρ as above is defined for any $q \geq 1$, and there is conditional dependence, $X \not\perp\!\!\!\perp Y \mid Z$, if there exists $q \geq 1$ for which $\rho \neq 0$. We suggest trying one or a few, small values, e.g. $q \in \{1, \dots, 5\}$, and reject the hypothesis of conditional independence if one of the tests rejects the hypothesis, but of course be aware of multiple testing issues.

5. Simulation Study

In this section we examine the performance of our conditional independence test $\hat{\Psi}_n$ of Definition 16, when combining it with the quantile regression based conditional distribution function estimator $\hat{F}^{(m,n)}$ from Definition 2. Firstly, we verify the level and power results obtained in Section 4.4 and Section 4.5 empirically. Secondly, we compare the test with other conditional independence tests. The test was implemented in the R language (R Core Team, 2021) using the `quantreg` package (Koenker, 2021) as the backend for performing quantile regression. The implementation and code for producing the simulations can be obtained from `https://github.com/lassepetersen/partial-copula-CI-test`.

5.1 Evaluation method

We will evaluate the tests by their ability to hold level when data is generated from a distribution where conditional independence holds, and by their power when data is generated from a distribution where conditional independence does not hold. In order to make the results independent of a chosen significance level we will base the evaluation on the p-values of the tests.

If a test has valid level, then we expect the p-value to be asymptotically $\mathcal{U}[0, 1]$-distributed. In Sections 5.3 and 5.4 we evaluate the level by a Kolmogorov-Smirnov (KS) statistics as a function of sample size n, which is independent of any specific significance level. A small KS

statistic is an indication of valid level. To examine power we consider in Sections 5.3 and 5.4 the p-values of the test as a function of the sample size, where we expect the p-values to tend to zero under the alternative of conditional dependence. Here a small p-value is an indication of large power. In Section 5.5 we evaluate the power against a local alternative, which shrinks with the sample size n toward the hypothesis of conditional independence with rate $n^{-\frac{1}{2}}$.

5.2 Data generating processes

This section gives an overview of the data generating processes that we use for benchmarking and comparison. The first category consists of data generating processes of the form

$$X = f_1(Z) + g_1(Z) \cdot \varepsilon_1 \quad \text{and} \quad Y = f_2(Z) + g_2(Z) \cdot \varepsilon_2 \tag{H}$$

where $f_1, f_2, g_1, g_2 : \mathbb{R}^d \to \mathbb{R}$ belong to some function class and $\varepsilon_1, \varepsilon_2$ are independent errors. For data generating processes of type (H), conditional independence is satisfied. The second category consists of data generating processes of the form

$$X = f_1(Z) + g_1(Z) \cdot \varepsilon_1 \quad \text{and} \quad Y = f_2(Z, X) + g_2(Z, X) \cdot \varepsilon_2 \tag{A}$$

where again $f_1, g_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2, g_2 : \mathbb{R}^{d+1} \to \mathbb{R}$ belong to some function class and $\varepsilon_1, \varepsilon_2$ are independent errors. Under data generating processes of type (A), conditional independence is not satisfied. We will consider four different data generating processes corresponding to different choices of functions f_1, g_1, f_2 and g_2 and error distributions.

(1) For data generating processes H_1 and A_1 we let

$$f_k(w_1, \ldots, w_d) = \sum_{j=1}^{d} \beta_{1,k,j} w_j + \beta_{2,k,j} w_j^2$$
$$g_k(w_1, \ldots, w_d) = \exp\left(- \left| \sum_{j=1}^{d} \alpha_{1,k,j} w_j + \alpha_{2,k,j} w_j^2 \right| \right)$$

for $k = 1, 2$ and real valued coefficients $(\alpha_{\ell,k,j}, \beta_{\ell,k,j})_{\ell=1,2,k=1,2,j=1,\ldots,d}$. Here each $Z_j \sim \mathcal{U}[-1, 1]$ independently, ε_1 follows an asymmetric Laplace distribution with location 0, scale 1 and skewness 0.8, and ε_2 follows a Gumpel distribution with location 0 and scale 1.

(2) For data generating processes H_2 and A_2 we let $g_1 = g_2 = 1$ and

$$f_k(w_1, \ldots, w_d) = \sum_{j=1}^{d} \beta_{k,j} w_j$$

for $k = 1, 2$ and real valued coefficients $(\beta_{k,j})_{k=1,2,j=1,\ldots,d}$. Here each $Z_j \sim \mathcal{U}[-1, 1]$ independently and both ε_1 and ε_2 follow a $\mathcal{N}(0, 1)$-distribution independently.

(3) For data generating processes H_3 and A_3 we let $g_1 = g_2 = 1$ and

$$f_k(w_1, \dots, w_d) = \sum_{j=1}^{d} \beta_{1,k,j} w_j + \beta_{2,k,j} w_j^2$$

for $k = 1, 2$ and real valued coefficients $(\beta_{\ell,k,j})_{\ell=1,2,k=1,2,j=1,\dots,d}$. Here each $Z_j \sim \mathcal{U}[-1, 1]$ independently and both ε_1 and ε_2 follow a $\mathcal{N}(0, 1)$-distribution independently.

(4) For data generating processes H_4 and A_4 we let $f_1 = f_2 = 0$ and

$$g_k(w_1, \dots, w_d) = \sum_{j=1}^{d} \beta_{1,k,j} w_j + \beta_{2,k,j} w_j^2$$

for $k = 1, 2$ for real valued coefficients $(\beta_{\ell,k,j})_{\ell=1,2,k=1,2,j=1,\dots,d}$. Here each $Z_j \sim \mathcal{U}[-1, 1]$ independently and both ε_1 and ε_2 follow a $\mathcal{N}(0, 1)$-distribution independently.

Each time we simulate from data generating processes $H_1, \dots, H_4$ we first draw the coefficients of the functions f_k, g_k from a $\mathcal{N}(0, 1)$-distribution in order to make the results independent of a certain combination of parameters. When we simulate from the data generating processes A_1, A_2 and A_3 we first draw the coefficients of f_k, g_k to be either -1 or 1 with equal probability in order to fix the signal to noise ratio between the predictors and responses. When simulating from A_4 we simulate the coefficients of g_k to be either -5 or 5, because the conditional dependence lies in the variance for A_4, and a stronger signal is needed to compare the power of the tests using the same samples sizes as for A_1, A_2 and A_3.

The data generating processes H_2, H_3, H_4, A_2, A_3 and A_4 can be shown to satisfy Assumption 3 that is needed for Corollary 9, since they are linear (in the parameters) location-scale models with bounded covariates (Belloni and Chernozhukov, 2011, Section 2.5). The processes H_1 and A_1 are not of this form, since g_1 and g_2 are nonlinear in the parameters. However, we include these in the simulation study to test the robustness of the test.

5.3 Level and power of partial copula test

In this section we examine the level and power properties of the test $\hat{\Psi}_n$. We examine the performance of the test on data generating processes H_1 and A_1 for dimensions $d \in \{1, 5, 10\}$ of Z. The test is performed as described in Section 4.7. As the quantile regression model we use an additive model with a B-spline basis of each variable with 5 degrees of freedom, and we try $q \in \{1, \dots, 5\}$. The result of the simulations can be seen in Figure 2. We observe that for $d = 1$ all five tests obtain level asymptotically under H_1, while for higher dimension $d \in \{5, 10\}$ the test with $q = 4$ has minor problems holding level. We also see that the p-values for all five tests tend to zero as the sample size increases under A_1. The convergence rate of the p-value depends on the dimension d such that a higher dimension gives a slower convergence rate. In conclusion we observe that our test holds level under a complicated data generating distribution (H_1), where there is a nonlinear conditional mean and variance dependence and skewed error distributions with super-Gaussian tails. Moreover, the test has power against the alternative of conditional dependence (A_1), however, for $d = 1$ we

Figure 2: Top: KS statistic for equality with a $\mathcal{U}[0,1]$-distribution of the p-values of the two tests computed from 500 simulations from H_1 for each combination of n and d. Bottom: Average p-values of the two tests separately computed over 200 simulations from A_1 for each combination of n and d. Dashed line indicates the common significance level 0.05. For visual purposes all p-values have been truncated at 10^{-10}.

see that $q = 5$ gives the best power, while $q = 1$ gives the best power for $d \in \{5, 10\}$. The testing procedure also displays robustness to the fact that the quantile regression models are misspecified.

5.4 Comparison with other tests

We now compare the partial copula based test $\hat{\Psi}_n$ with other nonparametric tests. We will compare with a residual based method, since this is another class of conditional independence test based on nonparametric regression. In order to describe this test we let

$$R_{1,i} = X_i - \hat{f}(Z_i) \quad \text{and} \quad R_{2,i} = Y_i - \hat{g}(Z_i)$$

for $i = 1, \ldots, n$ be the residuals obtained when performing conditional mean regression $\hat{f}$ of $f(z) = E(X \mid Z = z)$ and $\hat{g}$ of $g(z) = E(Y \mid Z = z)$ obtained from a sample $(X_i, Y_i, Z_i)_{i=1}^n$. We compare the following conditional independence tests:

- **GCM**: The Generalised Covariance Measure which tests for vanishing correlation between the residuals R_1 and R_2 given as above (Shah and Peters, 2020).
- **NPN correlation**: Testing for vanishing partial correlation in a nonparanormal distribution (Harris and Drton, 2013). This is a generalization of the partial correlation, which assumes a Gaussian dependence structure, but allows for arbitrary marginal distributions.
- **PC**: Our partial copula based test $\hat{\Psi}_n$ for $q \in \{1, 3, 5\}$ as described in Section 4.7.

We consider the behavior of the tests under H_2, A_2, H_3, A_3, H_4 and A_4. For fairness of comparison we choose our quantile and mean regression models to be the correct model class such that the tests perform at their oracle level, e.g., for H_3 we fit additive models with polynomial basis of degree 2. We fix the dimension d of Z to be 3 in all simulations for simplicity. The results of the simulations can be seen in Figure 3.

Under H_2 all five tests hold level, and we see that the NPN test has greatest power against A_2 followed by the GCM and $\hat{\Psi}_n$ with $q = 1$, while $\hat{\Psi}_n$ with $q \in \{3, 5\}$ does not have much power against A_2. In order to intuitively understand the effect of q see Figure 4. We see that in the estimated partial copula the dependence is captured by the overall correlation, while dividing $[\tau_{\min}, \tau_{\max}] \times [\tau_{\min}, \tau_{\max}]$ into subregions does not reveal finer dependence structure. Hence $q = 1$ is suitable to detect the dependence for A_1.

Under H_3 both the GCM test and $\hat{\Psi}_n$ with $q \in \{1, 3, 5\}$ hold level, but the NPN test does not hold level under H_3, which is due to the nonlinear response-predictor relationship. However, since both the GCM and $\hat{\Psi}_n$ test takes the nonlinearity into account, they can effectively filter away the Z-dependence. The NPN test has greatest power against the alternative A_3 following by $\hat{\Psi}_n$ with $q = 1$ and the GCM test. In Figure 4 we again see that the dependence in the estimated partial copula is described by the overall correlation, while dividing into subregions results a generalized correlation with elements that are close to zero, i.e., here $q = 1$ is suitable for capturing the dependence.

Under H_4, all test hold level. Note that the NPN test holds level even though there is a nonlinear conditional variance relation, since this is still a nonparanormal distribution. We also see that neither the GCM test nor the NPN test has power against A_4, while $\hat{\Psi}_n$ has some power against A_4 with the greatest power for $q = 3$. In Figure 4 we see that there is a clear dependence in the estimated partial copula, but that the overall correlation is close to zero. However, when dividing into subregions the generalized correlation is able to detect the dependencies in the tails of the distributions.

5.5 Power under local alternatives

Though GCM did not have power against the specific alternative A_4, it maintains level and it has power against a broad class of alternatives. To understand better when $\hat{\Psi}_n$ can be expected to have greater power than GCM, we consider a simulation, which is a small variation of the simulations presented in Section 5.2.

The dimension is fixed as $d = 1$, $Z \sim \mathcal{U}([0, 1])$ is uniformly distributed on $[0, 1]$, ε_1, ε_2 and W are independent and $\mathcal{N}(0, 1)$-distributed, and

$$X = (\beta Z^2 + 1)\varepsilon_1 + \gamma W \quad \text{and} \quad Y = (\beta Z^2 + 1)\varepsilon_2 + \gamma W \tag{A}$$

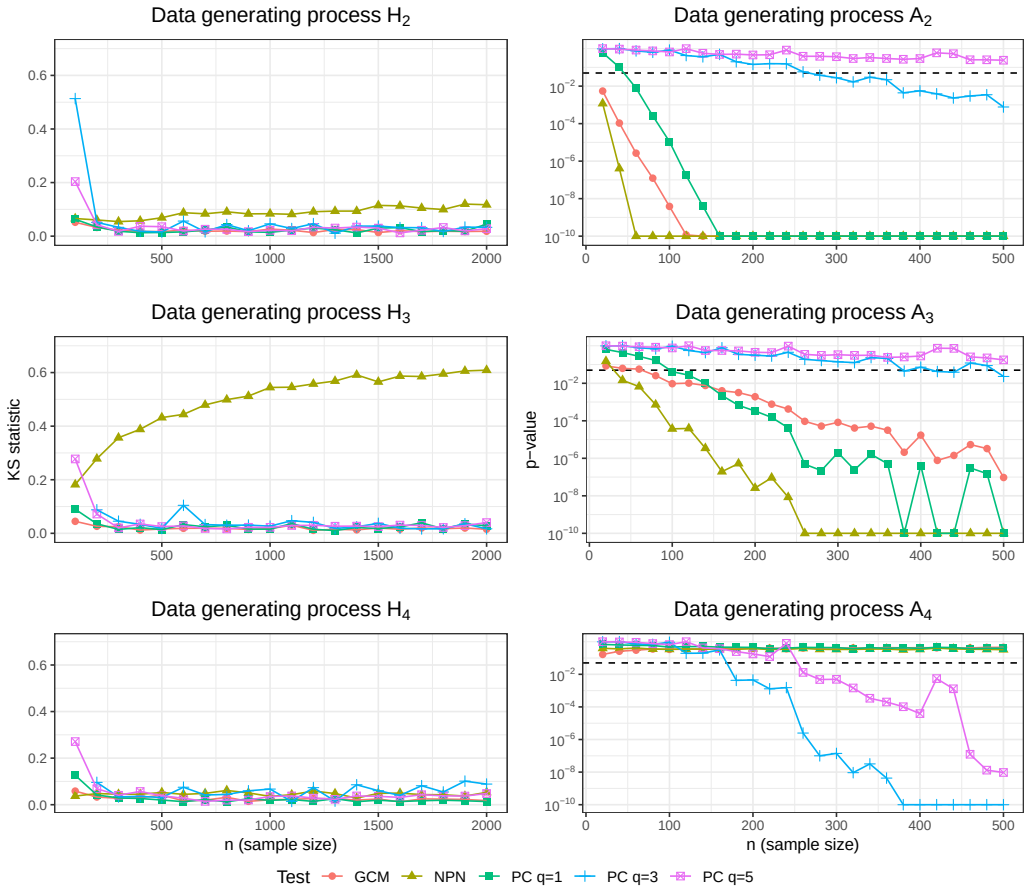

Figure 3: Left column: KS statistic for equality with a $\mathcal{U}[0,1]$-distribution of the p-values of the five tests computed from 500 simulations from $\mathrm{H}_2, \mathrm{H}_3$ and H_4, respectively, for each sample size n. Right column: Average p-values of the five tests separately computed over 200 simulations from $\mathrm{A}_2, \mathrm{A}_3$ and A_4, respectively, for each sample size n. For all simulations the dimension is fixed at $d = 3$. Dashed line indicates the common significance level 0.05. For visual purposes all p-values have been truncated at 10^{-10}.

for parameters $\beta, \gamma \in \mathbb{R}$. Conditionally on Z, the distribution of (X, Y) is a bivariate Gaussian distribution, and X and Y are conditionally independent if and only if $\gamma = 0$. We examine level and power by simulating 500 data sets for sample sizes $n \in \{100, 400, 1600\}$ and all combinations of parameters $\beta \in \{0, 1, 5, 10, 15, 20\}$, and local alternatives

$$\gamma^2 = \frac{\gamma_0^2}{\sqrt{n}}$$

for $\gamma_0^2 \in \{0, 50, 100, 150\}$. Note that $f(z) = E(X \mid Z = z) = 0$ and $g(z) = E(Y \mid Z = z) = 0$, which is exploited for GCM instead of estimating f and g. This should only increase the

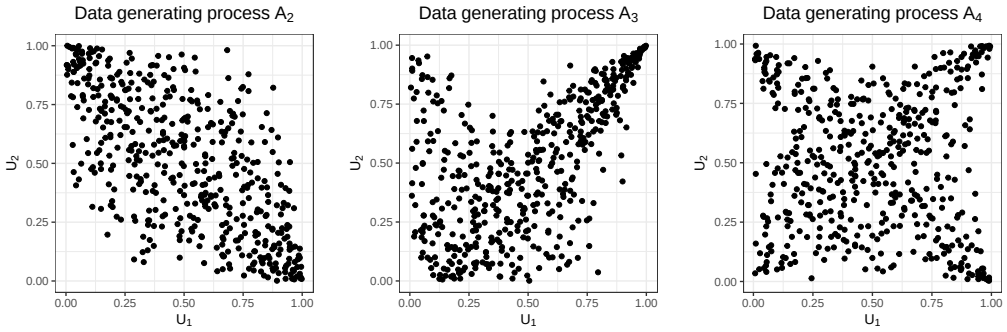

Figure 4: Estimated partial copula $(\hat{U}_{1,i}, \hat{U}_{2,i})_{i=1}^n$ from one realization from each of the data generating processes $\mathrm{A}_2, \mathrm{A}_3$ and A_4 for $d = 3$ and $n = 500$.

power of GCM relative to fitting any model of the conditional expectations. We perform the test $\hat{\Psi}_n$ as described in Section 4.7 using $q = 1$, and the quantile regression model is fitted using a polynomial basis of degree 2.

Figure 5 shows the results of the simulation. Both GCM and $\hat{\Psi}_n$ maintain level for $\gamma_0^2 = 0$. $\hat{\Psi}_n$ has comparable or superior power relative to GCM in all other cases. Both tests have decreasing power as a function of β, but $\hat{\Psi}_n$ maintains power even for large values of β, where GCM has almost no power. The power of $\hat{\Psi}_n$ against the local alternatives increases with the sample size, which shows how the increased precision for larger samples of the quantile regression based distribution functions improves power. We do not see the same for GCM, partly because no mean value model is fitted.

As β quantifies the conditional variance heterogeneity of X and Y given Z, we conclude that though GCM remains a valid test under conditional variance heterogeneity, its test statistic does not adequately account for the heterogeneity, and GCM has inferior power under local alternatives when compared to $\hat{\Psi}_n$.

6. Discussion

The first main contribution of this paper is an estimator of conditional distribution functions $\hat{F}^{(m,n)}$ based on quantile regression. We have shown that the estimator is pointwise (uniformly) consistent over a set of distributions $\mathcal{P}_0 \subset \mathcal{P}$ given that the quantile regression procedure is pointwise (uniformly) consistent over $\mathcal{P}_0$. Moreover, we showed that the convergence rate of the quantile regression procedure can be transferred directly to the estimator $\hat{F}^{(m,n)}$.

The second main contribution of this paper is an analysis of a nonparametric test for conditional independence based on the partial copula construction. We introduced a class of tests given in terms of a generalized correlation dependence measure ρ with the leading example being a trimmed version of the Spearman correlation. We showed that the test achieves asymptotic pointwise (uniform) level and power over $\mathcal{P}_0$ given that the conditional distribution function estimators are pointwise (uniformly) consistent over $\mathcal{P}_0$ with rate functions g_P and h_P satisfying $\sqrt{n} g_P(n) h_P(n) \to 0$. The partial copula has previously

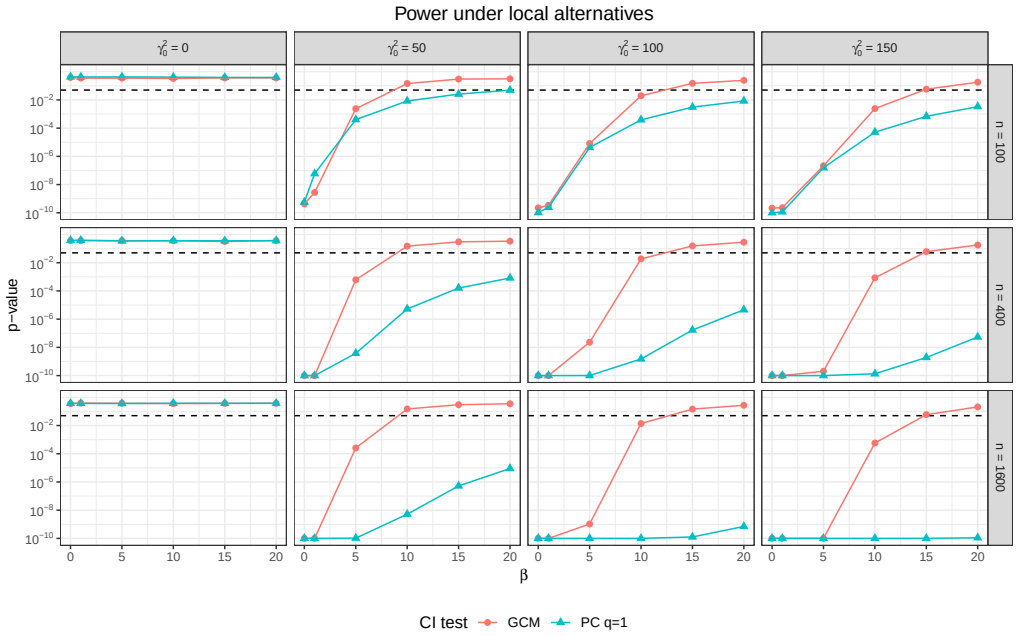

Figure 5: Average p-values of the three tests separately computed over 500 simulations for different values of the parameters β and γ_0 and different sample sizes n. The value $\gamma_0^2 = 0$ is equivalent to conditional independence. Other values correspond to the local alternatives $\gamma^2 = \gamma_0^2/\sqrt{n}$. The values of β determine the amount of heterogeneity of the mean regression residual variances with $\beta = 0$ meaning constant residual variance and $\beta = 20$ meaning substantial heterogeneity. Dashed line indicates the common significance level 0.05. For visual purposes all p-values have been truncated at 10^{-10}.

been considered for conditional independence testing in the literature, however, to the best of our knowledge, the results presented here are the first to explicitly connect the consistency requirements of the conditional distribution function estimators to level and power properties of the test.

Lastly, we established through a simulation study that the proposed test is sound under complicated data generating distributions, and that it has power comparable to or even better than other state-of-the-art nonparametric conditional independence tests. In particular, we demonstrated that our test has superior power against alternatives with variance heterogeneity between X and Y given Z when compared to conditional independence tests based on conventional residuals. We note that due to Daudin's lemma, tests based on conventional residuals can obtain power against any alternative if suitable transformations of X and Y are considered. In particular, if X^2 and Y^2 were used in our simulation study, GCM would have power against A_4. We tested the use of GCM in combination with X^2 and Y^2 in all our simulations (data not shown), and though it had some power against A_4, it was comparable to or inferior to just using GCM in all other simulations. Thus to

obtain good power properties, the specific choice of transformation appears important and to depend on the data generating distribution.

An important point about the test is the rate requirement $\sqrt{n}g_P(n)h_P(n) \to 0$ needed to achieve asymptotic level. The product structure means that the test is sound under quantile regression models with slower consistency rates than the usual parametric $n^{-1/2}$-rate. This opens up the methodology to nonparametric machine learning models. An interesting direction of research would be to empirically assess the performance of the test using machine learning inspired quantile regression models, such as deep neural networks, where explicit consistency rates are not available. We hypothesize that the method will still perform well in these scenarios due to the weak consistency requirement.

In this paper we have considered univariate X and Y. A possible extension of the test is to allow $X \in [0,1]^{r_1}$ and $Y \in [0,1]^{r_2}$ with $r_1, r_2 \geq 1$, and then consider the nonparametric residual $U_1 \in [0,1]^{r_1}$ of X given Z by performing coordinatewise probability integral transformations $U_{1,k} = F_{X_k|Z}(X_k \mid Z)$ for $k = 1, \ldots, r_1$, and similarly for constructing the nonparametric residual $U_2 \in [0,1]^{r_2}$ of Y given Z. Conditional independence $X \perp\!\!\!\perp Y \mid Z$ then implies pairwise independence of $U_{1,k}$ and $U_{2,l}$ for each $k = 1, \ldots, r_1$ and $l = 1, \ldots, r_2$. Combining our proposed test statistics for each such pair yields an $r_1 r_2 q^2$-dimensional test statistic, whose distribution under the hypothesis of conditional independence will be asymptotically Gaussian with mean 0. Its covariance matrix will only be partially known, though, due to the potential dependence between the pairs, but the unknown part could be estimated from the estimated nonparametric residuals. The multivariate statistic could be aggregated into a univariate test statistic in various ways, e.g. by a quadratic transformation as in (9), or by the maximum of the absolute values of its coordinates. In the low-dimensional case for fixed r_1 and r_2 our results would carry over immediately, and we expect that using the maximum could lead to high-dimensional results similar to Theorem 9 by Shah and Peters (2020).

A key property of the partial copula is that the nonparametric residuals U_1 and U_2 are independent under conditional independence and not only uncorrelated, which is the case for conventional residuals in additive noise models. Therefore, an important question is whether asymptotic level and power guaranties can be proven, when combining the partial copula with more general independence tests. In this paper we have focused on dependence measures of the form $\rho = E_P(\varphi(U_1)\varphi(U_2)^T)$ and tests based on

$$\hat{\rho}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(\hat{U}_{1,i})\varphi(\hat{U}_{2,i})^T$$

because it gives a flexible and general test for independence in the partial copula, it can be computed in linear time in the size of data, and most importantly its asymptotic theory is standard and easy to establish and apply. It also clearly illustrates the transfer of consistency of the conditional distribution function estimators to properties of the test. It is ongoing work to establish a parallel asymptotic theory for dependence measures of the form $\theta = E_P(h(U_1, U_2))$, where h is a kernel function, and whose estimators are U-statistics. This could potentially yield more powerfull tests against complicated alternatives of conditional dependence, but at the prize of increased computational complexity.

Acknowledgments

This work was supported by a research grant (13358) from VILLUM FONDEN.

Appendix A. Proofs

This appendix gives proofs of the main results of the paper. Throughout the proofs we will ignore the dependence of certain terms on the sample size to ease notation, e.g. we write $\hat{U}_{1,i}$ instead of $\hat{U}_{1,i}^{(n)}$ and $\hat{q}_{k,z}$ instead of $\hat{q}_{k,z}^{(n)}$.

A.1 Proof of Proposition 1

We need to bound the supremum

$$||F - \tilde{F}^{(m)}||_{\mathcal{T},\infty} = \sup_{z\in[0,1]^d} \sup_{t\in Q(\mathcal{T}|z)} |F(t \mid z) - \tilde{F}^{(m)}(t \mid z)|.$$

First we fix $z \in [0,1]^d$ and inspect the inner supremum. By construction we have

$$F(q_{k,z} \mid z) = \tilde{F}^{(m)}(q_{k,z} \mid z) = \tau_k$$

for $k = 1, \ldots, m$. Furthermore, since both F and $\tilde{F}^{(m)}$ are continuous and increasing in $t \in [0,1]$ we have that

$$\sup_{t\in[q_{k,z},q_{k+1,z}]} |F(t \mid z) - \tilde{F}^{(m)}(t \mid z)| \leq \tau_{k+1} - \tau_k$$

for each $k = 1, \ldots, m-1$. Since $Q(\mathcal{T} \mid z) = [q_{\min,z}, q_{\max,z}] = \bigcup_{k=1}^{m-1}[q_{k,z}, q_{k+1,z}]$ we now have

$$\begin{aligned}\sup_{t\in Q(\mathcal{T}|z)} |F(t \mid z) - \tilde{F}^{(m)}(t \mid z)| &= \max_{k=1,\ldots,m-1} \sup_{t\in[q_{k,z},q_{k+1,z}]} |F(t \mid z) - \tilde{F}^{(m)}(t \mid z)| \\ &\leq \max_{k=1,\ldots,m-1} (\tau_{k+1} - \tau_k) = \kappa_m.\end{aligned}$$

The result now follows from taking supremum over $z \in [0,1]^d$ as the right hand side of the inequality does not depend on z. □

A.2 Proof of Proposition 4

We need to bound the supremum

$$\|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty} = \sup_{z\in[0,1]^d} \sup_{t\in Q(\mathcal{T}|z)} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)|.$$

Our proof strategy is the following. First we evaluate the inner supremum over $t \in Q(\mathcal{T} \mid z)$ analytically to obtain a bound in terms of the quantile regression prediction error. Then we will evaluate the outer supremum over $z \in [0,1]^d$ and use the assumed consistency from Assumption 1. First define the two quantities

$$A(m,n,z) := \kappa_m \cdot \frac{\max_{k=1,\ldots,m} |q_{k,z} - \hat{q}_{k,z}^{(n)}|}{\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z})}$$

and

$$B(m,n,z) := \kappa_m \cdot \frac{\max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}^{(n)}|}{\min_{k=1,\dots,m-1}(\hat{q}_{k+1,z}^{(n)} - \hat{q}_{k,z}^{(n)})}.$$

We then have the following key result regarding the inner supremum over $t \in Q(\mathcal{T} \mid z)$.

Proposition 24 *Let Assumption 1 (i) be satisfied. Then for all $P \in \mathcal{P}_0$ and $\varepsilon > 0$ there exists $N \geq 1$ such that for all $n \geq N$,*

$$\sup_{t \in Q(\mathcal{T}|z)} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \leq \max\{A(m,n,z), B(m,n,z)\}$$

for all $z \in [0,1]^d$ and all grids $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ with probability at least $1-\varepsilon$.

We need a number of auxilliary results before proving Proposition 24. We start by proving the following key lemma that reduces the number of distinct cases of relative positions of the true conditional quantiles $q_{k,z}$ and the estimated conditional quantiles $\hat{q}_{k,z}$.

Lemma 25 *Let Assumption 1 (i) be satisfied. Then for each $P \in \mathcal{P}_0$ and $\varepsilon > 0$ there exists $N \geq 1$ such that for all $n \geq N$ we have that $\hat{q}_{k,z} \in (q_{k-1,z}, q_{k+1,z})$ for each $k = 1, \dots, m$ and $z \in [0,1]^d$ and for all grids $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ with probability at least $1-\varepsilon$.*

Proof Fix a distribution $P \in \mathcal{P}_0$. Let $\mathcal{G}$ be the set of all grids $(\tau_k)_{k=1}^m$ in $\mathcal{T}$. Then

$$\sup_{\mathcal{G}} \sup_{z \in [0,1]^d} \max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}^{(n)}| \leq \sup_{z \in [0,1]^d} \sup_{\tau \in \mathcal{T}} |Q(\tau \mid z) - \hat{Q}^{(n)}(\tau \mid z)| \xrightarrow{P} 0$$

under Assumption 1 (i). Since $q_{k,z} \in (q_{k-1,z}, q_{k+1,z})$ for each $k = 1, \dots, m$ and $z \in [0,1]^d$ for all grids $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ the result follows. ∎

Next we have some lemmas giving the supremum of certain functions over certain intervals that will be useful in the main proof.

Lemma 26 *Let $a \leq b < c \leq d$ and $f(t) = \frac{t-a}{c-a} - \frac{t-b}{d-b}$. Then $\sup_{t \in [b,c]} f(t) = \max\{\frac{b-a}{c-a}, \frac{d-c}{d-b}\}$.*

Proof Note that f is a linear function. Thus the supremum is obtained in one of the intervals endpoints, i.e., $\sup_{t \in [b,c]} f(t) = \max\{f(b), f(c)\}$. We see that

$$f(b) = \frac{b-a}{c-a} \quad \text{and} \quad f(c) = 1 - \frac{c-b}{d-b} = 1 - \frac{c-d+d-b}{d-b} = \frac{d-c}{d-b},$$

which shows the result. ∎

Lemma 27 *Let $a < b \leq c < d$ and $f(t) = \alpha + \beta \cdot \frac{t-b}{d-b} - \alpha \cdot \frac{t-a}{c-a}$ where $\alpha, \beta > 0$. Then we have $\sup_{t \in [b,c]} f(t) = \max\{\alpha \cdot \frac{c-b}{c-a}, \beta \cdot \frac{c-b}{d-b}\}$.*

Proof The function f is a linear function, and hence the supremum is obtained in one of the interval endpoints. We see that

$$f(b) = \alpha - \alpha \cdot \frac{b-a}{c-a} = \alpha \cdot \frac{c-b}{c-a} \quad \text{and} \quad f(c) = \beta \cdot \frac{c-b}{d-b},$$

which shows the claim. ■

Lemma 28 *Let $a \leq b < c < d$ and $f(t) = |g(t)|$ where $g(t) = \frac{t-b}{c-b} - \frac{t-a}{d-a}$. Then we have that $\sup_{t\in[b,c]} f(t) = \max\{\frac{b-a}{d-a}, \frac{d-c}{d-a}\}$.*

Proof Note that $f(t)$ is a convex function. Therefore the supremum of $f(t)$ is obtained in one of the interval endpoints. We see that

$$f(b) = \left|\frac{b-b}{c-b} - \frac{b-a}{d-a}\right| = \left|-\frac{b-a}{d-a}\right| = \frac{b-a}{d-a}$$

and

$$f(c) = \left|\frac{c-b}{c-b} - \frac{c-a}{d-a}\right| = \left|1 - \frac{c-a}{d-a}\right| = \left|1 - \frac{c-d+d-a}{d-a}\right| = \frac{d-c}{d-a}$$

which was what we wanted. ■

We are now ready to show Proposition 24.

Proof [Proof (of Proposition 24)]

We will compute the supremum over $t \in Q(\mathcal{T} \mid z) = [q_{\min,z}, q_{\max,z}]$ as the maximum of the suprema over the intervals $[q_{k,z}, q_{k+1,z}]$ for $k = 1, \ldots, m-1$, i.e.,

$$\sup_{t\in Q(\mathcal{T}|z)} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| = \max_{k=1,\ldots,m-1} \sup_{t\in[q_{k,z},q_{k+1,z}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)|.$$

This is useful since on each interval of the form $[q_{k,z}, q_{k+1,z}]$ we have that $\tilde{F}^{(m)}(\cdot \mid z)$ is a linear function, while $\hat{F}^{(m,n)}(\cdot \mid z)$ is a piecewise linear function.

First fix a distribution $P \in \mathcal{P}_0$ and $\varepsilon > 0$. Using Lemma 25 we choose $N \geq 1$ such that $\hat{q}_{k,z} \in (q_{k-1,z}, q_{k+1,z})$ for $k = 1, \ldots, m-1$ and $z \in [0,1]^d$ for each grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ with probability at least $1 - \varepsilon$. Now fix a $k = 1, \ldots, m-1$ such that we will examine the supremum on $[q_{k,z}, q_{k+1,z}]$. The relative position of the true and estimated conditional quantiles can be divided into four cases:

1) $q_{k,z} \geq \hat{q}_{k,z}$ and $q_{k+1,z} \geq \hat{q}_{k+1,z}$.

2) $q_{k,z} \geq \hat{q}_{k,z}$ and $q_{k+1,z} < \hat{q}_{k+1,z}$.

3) $q_{k,z} < \hat{q}_{k,z}$ and $q_{k+1,z} \geq \hat{q}_{k+1,z}$.

4) $q_{k,z} < \hat{q}_{k,z}$ and $q_{k+1,z} < \hat{q}_{k+1,z}$.

We start with case 1). First we compute the supremum over $t \in [q_{k,z}, \hat{q}_{k+1,z}]$ and then over $t \in [\hat{q}_{k+1,z}, q_{k+1,z}]$. We have that

$$|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| = (\tau_{k+1} - \tau_k) \left(\frac{t - \hat{q}_{k,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}} - \frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} \right)$$

for $t \in [q_{k,z}, \hat{q}_{k+1,z}]$. Hence we can compute the supremum as

$$\begin{aligned}
&\sup_{t \in [q_{k,z}, \hat{q}_{k+1,z}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= (\tau_{k+1} - \tau_k) \max \left\{ \frac{q_{k,z} - \hat{q}_{k,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}}, \frac{q_{k+1,z} - \hat{q}_{k+1,z}}{q_{k+1,z} - q_{k,z}} \right\} \\
&\leq \kappa_m \max \left\{ \frac{\max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\dots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z})}, \frac{\max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\dots,m-1}(q_{k+1,z} - q_{k,z})} \right\}
\end{aligned}$$

where we have used Lemma 26. Now we see that

$$\begin{aligned}
&|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= (\tau_{k+1} - \tau_k) + (\tau_{k+2} - \tau_{k+1}) \frac{t - \hat{q}_{k+1,z}}{\hat{q}_{k+2,z} - \hat{q}_{k+1,z}} - (\tau_{k+1} - \tau_k) \frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}}
\end{aligned}$$

for $t \in [\hat{q}_{k+1,z}, q_{k+1,z}]$. We compute the supremum to be

$$\begin{aligned}
&\sup_{t \in [\hat{q}_{k+1,z}, q_{k+1,z}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= \max \left\{ (\tau_{k+1} - \tau_k) \frac{\hat{q}_{k+1,z} - q_{k+1,z}}{q_{k+1,z} - q_{k,z}}, (\tau_{k+2} - \tau_{k+1}) \frac{\hat{q}_{k+1,z} - q_{k+1,z}}{\hat{q}_{k+2,z} - \hat{q}_{k+1,z}} \right\} \\
&\leq \kappa_m \max \left\{ \frac{\max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\dots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z})}, \frac{\max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\dots,m-1}(q_{k+1,z} - q_{k,z})} \right\}
\end{aligned}$$

where we have used Lemma 27. This covers case 1).

Now let us proceed to case 2). Here we can evaluate the supremum over $t \in [q_{k,z}, q_{k+1,z}]$ directly. We have that

$$|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| = (\tau_{k+1} - \tau_k) \left| \frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} - \frac{t - \hat{q}_{k,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}} \right|.$$

The supremum can now be evaluated using Lemma 28 to be

$$\begin{aligned}
&\sup_{t \in [q_{k,z}, q_{k+1,z}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= (\tau_{k+1} - \tau_k) \max \left\{ \frac{q_{k,z} - \hat{q}_{k,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}}, \frac{\hat{q}_{k+1,z} - q_{k+1,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}} \right\} \\
&\leq \kappa_m \frac{\max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\dots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z})}.
\end{aligned}$$

In case 3) we need to divide into three cases, namely when $t \in [q_{k,z}, \hat{q}_{z,k}]$, $t \in [\hat{q}_{k,z}, \hat{q}_{k+1,z}]$ and $t \in [\hat{q}_{k+1,z}, q_{k+1,z}]$. In the first case we have

$$\begin{aligned}
&|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= (\tau_k - \tau_{k-1}) + (\tau_{k+1} - \tau_k)\frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} - (\tau_k - \tau_{k-1})\frac{t - \hat{q}_{k-1,z}}{\hat{q}_{k,z} - \hat{q}_{k-1,z}}
\end{aligned}$$

for $t \in [q_{k,z}, \hat{q}_{k,z}]$. Therefore we have

$$\begin{aligned}
&\sup_{t \in [q_{k,z}, \hat{q}_{z,k}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&\leq \max\left\{(\tau_k - \tau_{k-1})\frac{\hat{q}_{k,z} - q_{k,z}}{\hat{q}_{k,z} - \hat{q}_{k-1,z}}, (\tau_{k+1} - \tau_k)\frac{\hat{q}_{k,z} - q_{k,z}}{q_{k+1,z} - q_{k,z}}\right\} \\
&\leq \kappa_m \max\left\{\frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z})}, \frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z})}\right\}
\end{aligned}$$

where we have used Lemma 27. In the second case we have

$$|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| = (\tau_{k+1} - \tau_k)\left|\frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} - \frac{t - \hat{q}_{k,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}}\right|,$$

for $t \in [\hat{q}_{k,z}, \hat{q}_{k+1,z}]$ and therefore we obtain

$$\begin{aligned}
&\sup_{t \in [\hat{q}_{k,z}, \hat{q}_{k+1,z}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&\leq (\tau_{k+1} - \tau_k)\max\left\{\frac{\hat{q}_{k,z} - q_{k,z}}{q_{k+1,z} - q_{k,z}}, \frac{q_{k+1,z} - \hat{q}_{k+1,z}}{q_{k+1,z} - q_{k,z}}\right\} \\
&\leq \kappa_m \cdot \frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z})}
\end{aligned}$$

where we have used Lemma 28. In the third case we have

$$\begin{aligned}
&|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= (\tau_{k+1} - \tau_k) + (\tau_{k+2} - \tau_{k+1})\frac{t - \hat{q}_{k+1,z}}{\hat{q}_{k+2,z} - \hat{q}_{k+1,z}} - (\tau_{k+1} - \tau_k)\frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}}
\end{aligned}$$

for $t \in [\hat{q}_{k+1,z}, q_{k+1,z}]$. So we obtain

$$\begin{aligned}
&\sup_{t \in [\hat{q}_{k+1,z}, q_{k+1,z}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= \max\left\{(\tau_{k+1} - \tau_k)\frac{\hat{q}_{k+1,z} - q_{k+1,z}}{q_{k+1,z} - q_{k,z}}, (\tau_{k+2} - \tau_{k+1})\frac{\hat{q}_{k+1,z} - q_{k+1,z}}{\hat{q}_{k+2,z} - \hat{q}_{k+1,z}}\right\} \\
&\leq \kappa_m \max\left\{\frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z})}, \frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z})}\right\}
\end{aligned}$$

where we have used Lemma 27.

Let us now examine case 4). Here we have the two sub cases $t \in [q_{k,z}, \hat{q}_{k,z}]$ and $t \in [\hat{q}_{k,z}, q_{k+1,z}]$. First we see that

$$\begin{aligned}
&|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= (\tau_k - \tau_{k-1}) + (\tau_{k+1} - \tau_k)\frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} - (\tau_k - \tau_{k-1})\frac{t - \hat{q}_{k-1,z}}{\hat{q}_{k,z} - \hat{q}_{k-1,z}}
\end{aligned}$$

for $t \in [q_{k,z}, \hat{q}_{k,z}]$. Thus we have

$$\begin{aligned}
&\sup_{t \in [q_{k,z}, \hat{q}_{z,k}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&\leq \max\left\{(\tau_k - \tau_{k-1})\frac{\hat{q}_{k,z} - q_{k,z}}{\hat{q}_{k,z} - \hat{q}_{k-1,z}}, (\tau_{k+1} - \tau_k)\frac{\hat{q}_{k,z} - q_{k,z}}{q_{k+1,z} - q_{k,z}}\right\} \\
&\leq \kappa_m \max\left\{\frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z})}, \frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z})}\right\}
\end{aligned}$$

where we have used 27. Now in the second case we have

$$|\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| = (\tau_{k+1} - \tau_k)\left(\frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} - \frac{t - \hat{q}_{k,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}}\right)$$

for $t \in [\hat{q}_{k,z}, q_{k+1,z}]$. From this we get the supremum to be

$$\begin{aligned}
&\sup_{t \in [\hat{q}_{k,z}, q_{k+1,z}]} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \\
&= (\tau_{k+1} - \tau_k)\max\left\{\frac{\hat{q}_{k,z} - q_{k,z}}{q_{k+1,z} - q_{k,z}}, \frac{q_{k+1,z} - \hat{q}_{k+1,z}}{\hat{q}_{k+1,z} - \hat{q}_{k,z}}\right\} \\
&\leq \kappa_m \max\left\{\frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z})}, \frac{\max_{k=1,\ldots,m}|q_{k,z} - \hat{q}_{k,z}|}{\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z})}\right\}
\end{aligned}$$

where we have used Lemma 26. Taking maximum of all cases and sub cases yields the desired result. ■

We will now move on to tackling the problem of controlling the outer supremum over $z \in [0,1]^d$. First we prove the following technical lemma that gives control over the denominators in $A(m,n,z)$ and $B(m,n,z)$.

Lemma 29 *Let Assumption 1 be satisfied. Let $\gamma_m := \min_{k=1,\ldots,m-1}(\tau_{k+1} - \tau_k)$ denote the finest subinterval of the grid. Then for each $P \in \mathcal{P}_0$ we have*

$$\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z}) \geq \frac{\gamma_m}{C_P}$$

for almost all $z \in [0,1]^d$ for each grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$. Also for all $\varepsilon > 0$ there is $N \geq 1$ such that for all $n \geq N$ we have

$$\min_{k=1,\ldots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z}) \geq \frac{\gamma_m}{3C_P}$$

for almost all $z \in [0,1]^d$ for each grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ with probability at least $1 - \varepsilon$.

Proof Fix a distribution $P \in \mathcal{P}_0$. We see that

$$\begin{aligned}\tau_{k+1} - \tau_k &= F(q_{k+1,z} \mid z) - F(q_{k,z} \mid z) \\ &= \int_{q_{k,z}}^{q_{k+1,z}} f(x \mid z)\mathrm{d}x \leq C_P \cdot (q_{k+1,z} - q_{k,z})\end{aligned}$$

for each $k = 1, \ldots, m-1$ and almost all $z \in [0,1]^d$ for each grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$. Here we have used Assumption 1 (ii). Rearranging and taking minimum, we have that

$$\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z}) \geq \min_{k=1,\ldots,m-1} \frac{\tau_{k+1} - \tau_k}{C_P} = \frac{\gamma_m}{C_P}$$

for almost all $z \in [0,1]^d$ and each grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$. Now let $\varepsilon > 0$ be given. Choose $N \geq 1$ such that for all $n \geq N$ we have

$$P\left(\hat{q}_{k,z} \in \left(q_{k,z} - \frac{\gamma_m}{3C_P}, q_{k,z} + \frac{\gamma_m}{3C}\right)\right) \geq 1 - \varepsilon$$

for all $k = 1, \ldots, m$ and all $z \in [0,1]^d$ for each $(\tau_k)_{k=1}^m$ in $\mathcal{T}$, which is possible due to Assumption 1 (i). In this case

$$\hat{q}_{k,z} \leq q_{k,z} + \frac{\gamma_m}{3C_P} \quad \text{and} \quad \hat{q}_{k+1,z} \geq q_{k+1,z} - \frac{\gamma_m}{3C_P}$$

for all $k = 1, \ldots, m-1$ and $z \in [0,1]^d$ with probability at least $1 - \varepsilon$. Thus for $n \geq N$,

$$\begin{aligned}\min_{k=1,\ldots,m-1}(\hat{q}_{k+1,z} - \hat{q}_{k,z}) &\geq \min_{k=1,\ldots,m-1}\left(q_{k+1,z} - \frac{\gamma_m}{3C_P} - \left(q_{k,z} + \frac{\gamma_m}{3C_P}\right)\right) \\ &= \min_{k=1,\ldots,m-1}\left(q_{k+1,z} - q_{k,z} - \frac{2\gamma_m}{3C_P}\right) \\ &\geq \frac{\gamma_m}{C_P} - \frac{2\gamma_m}{3C_P} = \frac{\gamma_m}{3C_P}\end{aligned}$$

for all $z \in [0,1]^d$ and each grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ with probability at least $1 - \varepsilon$. ■

We are now ready to prove the main result.

Proof [Proof (of Proposition 4)]

Fix a distribution $P \in \mathcal{P}_0$. Let $\varepsilon \in (0,1)$ be given. Firstly, we use Proposition 24 to choose $N_1 \geq 1$ such that the event

$$E_1 = \left(\sup_{t \in Q(\mathcal{T}|z)} |\tilde{F}^{(m)}(t \mid z) - \hat{F}^{(m,n)}(t \mid z)| \leq \max\{A(m,n,z), B(m,n,z)\}\right)$$

has probability at least $1 - \varepsilon/3$ for all $n \geq N_1$ and every grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$. Secondly, according to Lemma 29 we have that

$$\min_{k=1,\ldots,m-1}(q_{k+1,z} - q_{k,z}) \geq \frac{\gamma_m}{C_P}$$

and we can choose $N_2 \geq 1$ such that the event

$$E_2 = \left(\min_{k=1,\dots,m-1} (\hat{q}_{k+1,z} - \hat{q}_{k,z}) \geq \frac{\gamma_m}{3C_P} \right)$$

has probability at least $1 - \varepsilon/3$ for all $n \geq N_2$ and every grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$. Thirdly, we can choose $N_3 \geq 1$ and $M'_P > 0$ such that the event

$$E_3 = \left(\frac{\sup_{z\in[0,1]^d} \max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}|}{g_P(n)} \leq M'_P \right)$$

has probability at least $1 - \varepsilon/3$ for all $n \geq N_3$ and every $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ using Assumption 1 (i). Now we note that on the event $E := E_1 \cap E_2 \cap E_3$ we have

$$\begin{aligned} \|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty} &\leq \sup_{z\in[0,1]^d} \max\{A(m,n,z), B(m,n,z)\} \\ &= 3C_P \cdot \frac{\kappa_m}{\gamma_m} \sup_{z\in[0,1]^d} \max_{k=1,\dots,m} |q_{k,z} - \hat{q}_{k,z}| \\ &\leq 3C_P \cdot M'_P \cdot g_P(n) \end{aligned}$$

with probability $P(E) \geq 1 - \varepsilon$ for all $n \geq N$ and every grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ where $N := \max\{N_1, N_2, N_3\}$. Here we have used that $\kappa_m/\gamma_m = 1$ due to the grids being equidistant. We can now set $M_P := 3C_P \cdot M'_P$ such that

$$P\left(\frac{\|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty}}{g_P(n)} > M_P \right) < \varepsilon$$

whenever $n \geq N$. This shows that $\|\tilde{F}^{(m)} - \hat{F}^{(m,n)}\|_{\mathcal{T},\infty} \in \mathcal{O}_P(g_P(n))$ for every equidistant grid $(\tau_k)_{k=1}^m$ in $\mathcal{T}$ as wanted. ■

A.3 Proof of Theorem 5

According to Corollary 3 we have

$$\|F - \hat{F}^{(m_n,n)}\|_\infty \leq \kappa_{m_n} + \|\tilde{F}^{(m_n)} - \hat{F}^{(m_n,n)}\|_\infty.$$

Here $\|\tilde{F}^{(m_n)} - \hat{F}^{(m_n,n)}\|_\infty \in \mathcal{O}_P(g_P(n))$ for each equidistant grid $(\tau_k)_{k=1}^{m_n}$ in $\mathcal{T}$ due to Proposition 4. Since we have assumed that $\kappa_{m_n} \in o(g_P(n))$ we have the result. □

A.4 Proof of Proposition 6

The proof follows immediately from the proof of Proposition 4 and the stronger Assumption 2 in the following way. Note that the statement of Lemma 25 holds uniformly over $P \in \mathcal{P}_0$ under Assumption 2 (i). Therefore Proposition 24 also holds uniformly over $P \in \mathcal{P}_0$. Furthermore, the result of Lemma 29 also holds uniformly in $P \in \mathcal{P}_0$ under Assumption 2. Therefore the probability of the events E_1, E_2 and E_3 can be controlled uniformly over $P \in \mathcal{P}_0$ from which the result follows. □

A.5 Proof of Theorem 7

The corollary follows from Proposition 6 using the same argument as in the proof of Theorem 5. □

A.6 Proof of Corollary 9

Using Theorem 8 we have that

$$\begin{aligned}\sup_{z\in[0,1]^d}\sup_{\tau\in\mathcal{Q}}|Q(\tau\mid z)-\hat{Q}(\tau\mid z)| &= \sup_{z\in[0,1]^d}\sup_{\tau\in\mathcal{Q}}|h(z)^T(\beta_\tau-\hat{\beta}_\tau)| \\ &\leq \sup_{z\in[0,1]^d}\|h(z)\|_2\sup_{\tau\in\mathcal{Q}}\|\beta_\tau-\hat{\beta}_\tau\|_2 \\ &\in \mathcal{O}_P\left(\sqrt{\frac{s_n\log(p\vee n)}{n}}\right)\end{aligned}$$

since $\sup_{z\in[0,1]^d}\|h(z)\|_2<\infty$ because $[0,1]^d$ is compact and h is continuous. □

A.7 Proof of Proposition 11

Assume that $X \perp\!\!\!\perp Y \mid Z$. Then it also holds that $(X,Z) \perp\!\!\!\perp (Y,Z) \perp\!\!\!\perp Z$ and thus $U_1 \perp\!\!\!\perp U_2 \mid Z$. Letting f denote a generic density function, we now have that

$$\begin{aligned}f(u_1,u_2) &= \int f(u_1,u_2\mid z)f(z)\mathrm{d}z = \int f(u_1\mid z)f(u_2\mid z)f(z)\mathrm{d}z \\ &= \int f(u_1)f(u_2)f(z)\mathrm{d}z = f(u_1)f(u_2)\end{aligned}$$

for all $u_1,u_2\in[0,1]$, where we have used Proposition 10. □

A.8 Proof of Theorem 14

Before proving the theorem, we will supply a lemma that will aid us during the proof.

Lemma 30 *Let $\hat{F}^{(n)}_{X|Z}$ and $\hat{F}^{(n)}_{Y|Z}$ satisfy Assumption 5. Then*

$$\|\varphi_k\circ F_{X|Z}-\varphi_k\circ\hat{F}^{(n)}_{X|Z}\|_\infty\in\mathcal{O}_P(g_P(n)) \quad \textit{and} \quad \|\varphi_k\circ F_{Y|Z}-\varphi_k\circ\hat{F}^{(n)}_{Y|Z}\|_\infty\in\mathcal{O}_P(h_P(n))$$

for each $k=1,\ldots,q$ given that φ satisfies Assumption 4.

Proof We only show the first statement. Fix $k=1,\ldots,q$. We need to control the supremum

$$\sup_{z\in[0,1]^d}\sup_{t\in[0,1]}|\varphi_k(F_{X|Z}(t\mid z))-\varphi_k(\hat{F}^{(n)}_{X|Z}(t\mid z))|.$$

We will divide the supremum over $t\in[0,1]$ into two cases. Namely, when $t\in Q(\mathcal{T}\mid z)=[q_{\min,z},q_{\max,z}]$ and when $t\in Q(\mathcal{T}^c\mid z)=[q_{\min,z},q_{\max,z}]^c$. First we see that

$$\begin{aligned}&\sup_{z\in[0,1]^d}\sup_{t\in Q(\mathcal{T}|z)}|\varphi_k(F_{X|Z}(t\mid z))-\varphi_k(\hat{F}^{(n)}_{X|Z}(t\mid z))| \\ &\qquad\leq L_k\cdot\|F_{X|Z}-\hat{F}^{(n)}_{X|Z}\|_{\mathcal{T},\infty}\in\mathcal{O}_P(g_P(n))\end{aligned}$$

where L_k is the Lipschitz constant of φ_k under Assumption 4 (ii). Here we have used the consistency in Assumption 5 (i). Next we examine the supremum over $t \in Q(\mathcal{T}^c \mid z)$. First note that $F_{X|Z}(t \mid z) \in [\tau_{\min}, \tau_{\max}]^c$ whenever $t \in Q(\mathcal{T}^c \mid z)$. Also recall that the support of φ_k is $\mathcal{T}_k \subset \mathcal{T} = [\tau_{\min}, \tau_{\max}]$. Therefore $\varphi_k(F_{X|Z}(t \mid z)) = 0$ for $t \in Q(\mathcal{T}^c \mid z)$. Hence we have

$$\sup_{z\in[0,1]^d} \sup_{t\in Q(\mathcal{T}^c|z)} |\varphi_k(F_{X|Z}(t \mid z)) - \varphi_k(\hat{F}^{(n)}_{X|Z}(t \mid z))| = \sup_{z\in[0,1]^d} \sup_{t\in Q(\mathcal{T}^c|z)} |\varphi_k(\hat{F}^{(n)}_{X|Z}(t \mid z))|.$$

By Assumption 5 (i) we know that

$$\hat{F}^{(n)}_{X|Z}(q_{\min,z} \mid z) \xrightarrow{P} \tau_{\min} \quad \text{and} \quad \hat{F}^{(n)}_{X|Z}(q_{\max,z} \mid z) \xrightarrow{P} \tau_{\max}$$

for all $z \in [0,1]^d$. Since $\hat{F}^{(n)}_{X|Z}(\cdot \mid z)$ is increasing we thus know that the limit $\xi(t,z)$ from Assumption 5 (ii) must satisfy $\xi(t,z) \in [\tau_{\min}, \tau_{\max}]^c$ for $t \in Q(\mathcal{T}^c \mid z)$ and $z \in [0,1]^d$. Again, since the support of φ_k is $\mathcal{T}_k \subset \mathcal{T} = [\tau_{\min}, \tau_{\max}]$ we have that $\varphi_k(\xi(t,z)) = 0$ when $t \in Q(\mathcal{T}^c \mid z)$ and $z \in [0,1]^d$. Therefore we have that

$$\begin{aligned}\sup_{z\in[0,1]^d} \sup_{t\in Q(\mathcal{T}^c|z)} |\varphi_k(\hat{F}^{(n)}_{X|Z}(t \mid z))| &= \sup_{z\in[0,1]^d} \sup_{t\in Q(\mathcal{T}^c|z)} |\varphi_k(\xi(t,z)) - \varphi_k(\hat{F}^{(n)}_{X|Z}(t \mid z))| \\ &\leq L_k \cdot \|\xi - \hat{F}^{(n)}_{X|Z}\|_{\mathcal{T}^c,\infty} \in \mathcal{O}_P(g_P(n)),\end{aligned}$$

where we have used Assumption 5 (ii). Putting the two cases together we have that

$$\|\varphi_k \circ F_{X|Z} - \varphi_k \circ \hat{F}^{(n)}_{X|Z}\|_\infty \in \mathcal{O}_P(g_P(n))$$

which was what we wanted. ■

We can now prove the main theorem.

Proof [Proof (of Theorem 14)] Fix a distribution $P \in \mathcal{H}_0$. The key to proving the theorem is the decomposition

$$\hat{\rho}_n = \alpha_n + \beta_n + \gamma_n + \delta_n$$

where $\alpha_n, \beta_n, \gamma_n$ and δ_n are given by

$$\begin{aligned}\alpha_n &= \frac{1}{n}\sum_{i=1}^n \varphi(U_{1,i})\varphi(U_{2,i})^T, \\ \beta_n &= \frac{1}{n}\sum_{i=1}^n \Big(\varphi(\hat{U}_{1,i}) - \varphi(U_{1,i})\Big)\Big(\varphi(\hat{U}_{2,i}) - \varphi(U_{2,i})\Big)^T \\ \gamma_n &= \frac{1}{n}\sum_{i=1}^n \varphi(U_{1,i})\Big(\varphi(\hat{U}_{2,i}) - \varphi(U_{2,i})\Big)^T, \\ \delta_n &= \frac{1}{n}\sum_{i=1}^n \Big(\varphi(\hat{U}_{1,i}) - \varphi(U_{1,i})\Big)\varphi(U_{2,i})^T\end{aligned}$$

The term α_n will be driving the asymptotics of the test statistics, while β_n, γ_n and δ_n are error terms that we wish to show converge to zero sufficiently fast.

Let us start by examining α_n. Under Assumption 4 (iii) we see that

$$E_P(\varphi(U_{1,i})\varphi(U_{2,i})^T) = E_P(\varphi(U_{1,i}))E_P(\varphi(U_{2,i}))^T = 0$$

because $P \in \mathcal{H}_0$ and furthermore we see that

$$\begin{aligned}\mathrm{Cov}_P(\varphi_k(U_{1,i})\varphi_\ell(U_{2,i}), \varphi_s(U_{1,i})\varphi_t(U_{2,i})) &= E_P(\varphi_k(U_{1,i})\varphi_\ell(U_{2,i})\varphi_s(U_{1,i})\varphi_t(U_{2,i})) \\ &= E_P(\varphi_k(U_{1,i})\varphi_s(U_{1,i}))E_P(\varphi_\ell(U_{2,i})\varphi_t(U_{2,i})) \\ &= \int_0^1 \varphi_k(u)\varphi_s(u)\mathrm{d}u \int_0^1 \varphi_\ell(u)\varphi_t(u)\mathrm{d}u \\ &= \Sigma_{ks}\Sigma_{\ell t} = (\Sigma \otimes \Sigma)_{k\ell,st}\end{aligned}$$

for $k, \ell, s, t = 1, \ldots, q$. Observe that $\Sigma_{k,k} = 1$. Since α_n is the average of i.i.d. terms with zero mean and covariance $\Sigma \otimes \Sigma$, the central limit theorem states that

$$\sqrt{n}\alpha_n \Rightarrow_P \mathcal{N}(0, \Sigma \otimes \Sigma)$$

for each $P \in \mathcal{H}_0$.

Now let us examine the term $\sqrt{n}\beta_n$. Fix $k, \ell = 1, \ldots, q$. Then we have

$$\begin{aligned}|\sqrt{n}\beta_{k\ell,n}| &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^n \left|\varphi_k(\hat{U}_{1,i}) - \varphi_k(U_{1,i})\right| \cdot \left|\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right| \\ &\leq \frac{n}{\sqrt{n}} \|\varphi_k \circ \hat{F}^{(n)}_{X|Z} - \varphi_k \circ F_{X|Z}\|_\infty \cdot \|\varphi_\ell \circ \hat{F}^{(n)}_{Y|Z} - \varphi_\ell \circ F_{Y|Z}\|_\infty \\ &\in \mathcal{O}_P(\sqrt{n} g_P(n) h_P(n))\end{aligned}$$

where we have used Lemma 30, which is valid due to Assumption 5. Since we have assumed that the rate functions satisfy $\sqrt{n}g_P(n)h_P(n) \to 0$ we can conclude that $|\sqrt{n}\beta_{k\ell,n}| \to_P 0$ for each $k, \ell = 1, \ldots, q$. Hence $\sqrt{n}\beta_n \to_P 0$.

Now we turn to the cross terms γ_n and δ_n. The two terms are dealt with analogously, so we only examine γ_n. Fix $k, \ell = 1, \ldots, q$ and consider writing

$$\gamma_{k\ell,n} = \frac{1}{n}\sum_{i=1}^n C_i \quad \text{where} \quad C_i = \varphi_k(U_{1,i})\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right).$$

We will compute the mean and variance of $\sqrt{n}\gamma_{k\ell,n}$ conditionally on $(Y_j, Z_j)_{j=1}^n$ in order to use Chebyshev's inequality to show that it converges to zero in probability. Observe that

$$\begin{aligned}E_P(C_i \mid (Y_j, Z_j)_{j=1}^n) &= E_P\left(\varphi_k(U_{1,i})\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right) \mid (Y_j, Z_j)_{j=1}^n\right) \\ &= \left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right) E_P\left(\varphi_k(U_{1,i}) \mid (Y_j, Z_j)_{j=1}^n\right) \quad \text{a.s.}\end{aligned}$$

Here we have exploited that $\varphi_\ell(U_{2,i})$ and $\varphi_\ell(\hat{U}_{2,i}) = \varphi_\ell(\hat{F}_{Y|Z}(Y_i \mid Z_i))$ are measurable functions of $(Y_j, Z_j)_{j=1}^n$. Now since $P \in \mathcal{H}_0$ we have $\varphi_k(U_{1,i}) \perp\!\!\!\perp Y_i \mid Z_i$ and $\varphi_k(U_{1,i}) \perp\!\!\!\perp Z_i$ due to Proposition 10. Therefore

$$E_P\left(\varphi_k(U_{1,i}) \mid (Y_j, Z_j)_{j=1}^n\right) = E_P\left(\varphi_k(U_{1,i})\right) = 0 \quad \text{a.s.}$$

Testing Conditional Independence via Quantile Regression

where we have used Assumption 4 (iii). Hence $E_P(C_i \mid (Y_j, Z_j)_{j=1}^n) = 0$ a.s. From the tower property we also obtain that $E_P(C_i) = 0$ and therefore $\sqrt{n}\gamma_{k\ell,n}$ has mean zero. Let us turn to the conditional variance. Conditionally on $(Y_j, Z_j)_{j=1}^n$ the terms $(C_i)_{i=1}^n$ are i.i.d. because $\varphi_\ell \circ \hat{F}_{Y|Z}$ is $(Y_j, Z_j)_{j=1}^n$-measurable as exploited before. So we have

$$V_P(\sqrt{n}\gamma_{k\ell,n} \mid (Y_j, Z_j)_{j=1}^n) = \frac{1}{n}\sum_{i=1}^n V_P(C_i \mid (Y_j, Z_j)_{j=1}^n) = V_P(C_i \mid (Y_j, Z_j)_{j=1}^n).$$

We compute the conditional variance to be

$$\begin{aligned} V_P(C_i \mid (Y_j, Z_j)_{j=1}^n) &= E_P\left(\varphi_k(U_{1,i})^2 \left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 \mid (Y_j, Z_j)_{j=1}^n\right) \\ &= \left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 E_P\left(\varphi_k(U_{1,i})^2 \mid (Y_j, Z_j)_{j=1}^n\right) \\ &= \left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 E_P\left(\varphi_k(U_{1,i})^2\right) \\ &= \left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 \quad \text{a.s.} \end{aligned}$$

where we have used Assumption 4 (iii). We can use the the law of total variance to see that

$$\begin{aligned} V_P(\sqrt{n}\gamma_{k\ell,n}) &= E_P(V_P(\sqrt{n}\gamma_{k\ell,n} \mid (Y_j, Z_j)_{j=1}^n)) + V_P(E_P(\sqrt{n}\gamma_{k\ell,n} \mid (Y_j, Z_j)_{j=1}^n)) \\ &= E_P\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 + 0 = E_P\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2. \end{aligned}$$

By Lemma 30 we have that $\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 \to_P 0$ with similar arguments as before. Note that $\varphi_\ell : [0,1] \to \mathbb{R}$ is bounded due to continuity of φ_ℓ and compactness of $[0,1]$. Hence each term in the sequence

$$\left(\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2\right)_{i=1,\dots,n}$$

is bounded. Therefore we also have $E_P\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 \to 0$. For given $\varepsilon > 0$ we have by Chebyshev's inequality that

$$P(|\sqrt{n}\gamma_{k\ell,n}| > \varepsilon) \leq \frac{V_P(\sqrt{n}\gamma_{k\ell,n})}{\varepsilon^2} = \frac{1}{\varepsilon^2} \cdot E_P\left(\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right)^2 \to 0$$

for each $P \in \mathcal{H}_0$. This shows $\sqrt{n}\gamma_n \to_P 0$. By the same argument it can be shown that $\sqrt{n}\delta_n \to_P 0$. By Slutsky's lemma we now have that

$$\sqrt{n}\hat{\rho}_n = \sqrt{n}\alpha_n + \sqrt{n}\beta_n + \sqrt{n}\gamma_n + \sqrt{n}\delta_n \;\Rightarrow_P\; \mathcal{N}(0, \Sigma \otimes \Sigma)$$

for each $P \in \mathcal{H}_0$. This shows the theorem. ■

A.9 Proof of Corollary 15

First note that Σ is a positive definite matrix as $\varphi_1, \ldots, \varphi_q$ are assumed linearly independent. It thus has a positive definite matrix square root $\Sigma^{-1/2}$ satisfying $\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I$, and we have that

$$\sqrt{n}\Sigma^{-1/2}\hat{\rho}_n\Sigma^{-1/2} \Rightarrow_P \mathcal{N}(0, I \otimes I)$$

for $P \in \mathcal{H}_0$ where we have used Theorem 18. The test statistics T_n is therefore well defined and

$$nT_n = \|\sqrt{n}\Sigma^{-1/2}\hat{\rho}_n\Sigma^{-1/2}\|_F^2 \Rightarrow_P \chi^2_{q^2}$$

for $P \in \mathcal{H}_0$ by the continuous mapping theorem. □

A.10 Proof of Corollary 17

Under Assumption 5 we have by Corollary 15 that $nT_n \Rightarrow_P \chi^2_{q^2}$. Therefore

$$\limsup_{n\to\infty} E_P(\hat{\Psi}_n) = \limsup_{n\to\infty} P(nT_n > z_{1-\alpha}) = \limsup_{n\to\infty}(1 - (F_{nT_n}(z_{1-\alpha})) = 1 - (1-\alpha) = \alpha.$$

because $F_{nT_n}(t) \to \Phi(t)$ as $n \to \infty$ for all $t \in \mathbb{R}$ where Φ is the distribution function of a $\chi^2_{q^2}$-distribution and $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of a $\chi^2_{q^2}$-distribution. □

A.11 Proof of Theorem 18

The proof uses the same decomposition as in the proof of Theorem 14, i.e., $\hat{\rho}_n = \alpha_n + \beta_n + \gamma_n + \delta_n$. Let us first comment on the large sample properties of α_n. Since α_n is the i.i.d. average of terms with expectation ρ for all $P \in \mathcal{P}_0$ we have that $\alpha_n \to_P \rho$ for all $P \in \mathcal{P}_0$. The term $\hat{\beta}_n$ is dealt with similarly as in the proof of Theorem 14. For fixed $k, \ell = 1, \ldots, q$ we have that

$$\begin{aligned}|\beta_{k\ell,n}| &\leq \frac{1}{n}\sum_{i=1}^n \left|\varphi_k(\hat{U}_{1,i}) - \varphi_k(U_{1,i})\right| \left|\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})\right| \\ &\leq \|\varphi_k \circ \hat{F}^{(n)}_{X|Z} - \varphi_k \circ F_{X|Z}\|_\infty \cdot \|\varphi_\ell \circ \hat{F}^{(n)}_{Y|Z} - \varphi_\ell \circ F_{Y|Z}\|_\infty \in \mathcal{O}_P(g_P(n)h_P(n))\end{aligned}$$

where we have used Lemma 30. From Assumption 5 we get that $\beta_{k\ell,n} \to_P 0$ for each $k, \ell = 1, \ldots, q$, and so $\beta_n \to_P 0$ for all $P \in \mathcal{P}_0$. The terms γ_n and δ_n are analyzed similarly, so we only look at γ_n. We see that for $k, \ell = 1, \ldots, q$,

$$\begin{aligned}|\gamma_{k\ell,n}| &\leq \frac{1}{n}\sum_{i=1}^n |\varphi_k(U_{1,i})||\varphi_\ell(\hat{U}_{2,i}) - \varphi_\ell(U_{2,i})| \\ &\leq \|\varphi_k\|_\infty \cdot \|\varphi_\ell \circ \hat{F}^{(n)}_{Y|Z} - \varphi_\ell \circ F_{Y|Z}\|_\infty \in \mathcal{O}_P(h_P(n))\end{aligned}$$

where we have used that $\|\varphi_k\|_\infty < \infty$ since $\varphi_k : [0,1] \to \mathbb{R}$ is continuous and $[0,1]$ is compact. Here we have used Lemma 30, and we conclude that $\gamma_{k\ell,n} \to_P 0$ due to Assumption 5, which shows that $\gamma_n \to_P 0$ for all $P \in \mathcal{P}_0$. Conclusively, we have $\hat{\rho}_n \to_P \rho$ for all $P \in \mathcal{P}_0$. □

A.12 Proof of Corollary 19

Assume that $P \in \mathcal{A}_0$ such that $\rho_{k\ell} \neq 0$ for some $k, \ell = 1, \ldots, q$. Then we have

$$T_n = \|\Sigma^{-1/2}\hat{\rho}_n\Sigma^{-1/2}\|_F^2 \xrightarrow{P} \|\Sigma^{-1/2}\rho\Sigma^{-1/2}\|_F^2 > 0$$

for all $P \in \mathcal{A}_0$ because $\rho \neq 0$. Here we have used Theorem 18. Therefore we obtain that

$$nT_n = n\|\Sigma^{-1/2}\hat{\rho}_n\Sigma^{-1/2}\|_F^2 \xrightarrow{P} \infty$$

for all $P \in \mathcal{A}_0$. This means that

$$P(nT_n > c) \to 1$$

as $n \to \infty$ for all $c \in \mathbb{R}$. From this we obtain that

$$\liminf_{n\to\infty} E_P(\hat{\Psi}_n) = \liminf_{n\to\infty} P(nT_n > z_{1-\alpha}) = 1$$

for all $\alpha \in (0,1)$ whenever $P \in \mathcal{A}_0$. □

A.13 Proof of Proposition 20

Assume $(U_1, U_2) \perp\!\!\!\perp Z$ and $U_1 \perp\!\!\!\perp U_2$. Then it also holds that $U_1 \perp\!\!\!\perp U_2 \mid Z$, which gives $(U_1, Z) \perp\!\!\!\perp (U_2, Z) \mid Z$. More explicitly we have

$$(F_{X|Z}(X \mid Z), Z) \perp\!\!\!\perp (F_{Y|Z}(Y \mid Z), Z) \mid Z.$$

Transforming with the conditional quantile functions gives

$$Q_{X|Z}(F_{X|Z}(X \mid Z) \mid Z) \perp\!\!\!\perp Q_{Y|Z}(F_{Y|Z}(X \mid Z) \mid Z) \mid Z.$$

Since we assume throughout the paper that the conditional distributions $X \mid Z = z$ and $Y \mid Z = z$ are continuous for each $z \in [0,1]^d$ we get that $(X, Z) \perp\!\!\!\perp (Y, Z) \mid Z$ which reduces to $X \perp\!\!\!\perp Y \mid Z$. □

A.14 Proof of Theorem 21

We start by showing (i). Again we consider the decomposition $\hat{\rho}_n = \alpha_n + \beta_n + \gamma_n + \delta_n$ introduced in the proof of Theorem 14. By the stronger condition of Assumption 6 we immediately have that $\sqrt{n}\beta_n \to_{\mathcal{P}_0} 0, \sqrt{n}\gamma_n \to_{\mathcal{P}_0} 0$ and $\sqrt{n}\delta_n \to_{\mathcal{P}_0} 0$ by following the same arguments as in the proof of Theorem 14. The fact that $\sqrt{n}\alpha_n$ converges uniformly in distribution to a $\mathcal{N}(0, \Sigma\otimes\Sigma)$-distribution over $\mathcal{H}_0$ follows from the fact that the distribution of $(U_{1,i}, U_{2,i})_{i=1}^n$ is unchanged whenever $P \in \mathcal{H}_0$. By Lemma 37 we have that

$$\sqrt{n}\hat{\rho}_n = \sqrt{n}\alpha_n + \sqrt{n}\beta_n + \sqrt{n}\gamma_n + \sqrt{n}\delta_n \Rightarrow_{\mathcal{H}_0} \mathcal{N}(0, \Sigma \otimes \Sigma)$$

which shows part (i) of the theorem. Next we turn to part (ii) of the theorem. Analogously to the proof of Theorem 18 we have that $\beta_n \to_{\mathcal{P}_0} 0, \gamma_n \to_{\mathcal{P}_0} 0$ and $\delta_n \to_{\mathcal{P}_0} 0$ under Assumption 6. Now consider writing

$$\alpha_{k\ell,n} = \frac{1}{n}\sum_{i=1}^n A_i \quad \text{where} \quad A_i = \varphi_k(U_{1,i})\varphi_\ell(U_{2,i})$$

for $k, \ell = 1, \ldots, q$. Then $(A_i)_{i=1}^n$ are i.i.d. with $E_P(A_i) = \rho_{k\ell}$ and

$$V_P(A_i) = E_P(\varphi_k(U_{1,i})^2 \varphi_\ell(U_{2,i})^2) - \rho^2 \leq \|\varphi_k\|_\infty \|\varphi_\ell\|_\infty < \infty$$

for all $P \in \mathcal{P}_0$. Therefore, for given $\varepsilon > 0$, we have by Chebyshev's inequality that

$$\sup_{P \in \mathcal{P}_0} P(|\alpha_{k\ell,n} - \rho_{k\ell}| > \varepsilon) \leq \sup_{P \in \mathcal{P}_0} \frac{V_P(\frac{1}{n}\sum_{i=1}^n A_i)}{\varepsilon^2} = \sup_{P \in \mathcal{P}_0} \frac{V_P(A_i)}{n\varepsilon^2} \leq \frac{||\varphi_1||_\infty^2 ||\varphi_2||_\infty^2}{n\varepsilon^2} \to 0$$

for $n \to \infty$ which shows that $\alpha_n \to_{\mathcal{P}_0} \rho$. From this we get $\hat{\rho}_n \to_{\mathcal{P}_0} \rho$ as wanted.

A.15 Proof of Corollary 22

Note that due to Theorem 21 (i) we have that $nT_n \Rightarrow_{\mathcal{H}_0} \chi^2_{q^2}$ under Assumption 6 using the same argument as in the proof of Corollary 15. Then the result is obtained by the same argument as in the proof of Corollary 17 by noting that $\sup_{P \in \mathcal{H}_0} |F_{nT_n}(t) - \Phi(t)| \to 0$ as $n \to \infty$ for all $t \in \mathbb{R}$. □

A.16 Proof of Corollary 23

Let $\lambda > 0$ be fixed. By Theorem 21 (ii) we have $\hat{\rho}_n \to_{\mathcal{A}_\lambda} \rho$ where $|\rho_{k\ell}| > \lambda > 0$ for some $k, \ell = 1, \ldots, q$. Therefore $\inf_{P \in \mathcal{A}_\lambda} |\rho_{k\ell}| \geq \lambda > 0$ and so

$$T_n = \|\Sigma^{-1/2} \hat{\rho}_n \Sigma^{-1/2}\|_F^2 \to_{\mathcal{A}_\lambda} \|\Sigma^{-1/2} \rho \Sigma^{-1/2}\|_F^2 > 0$$

since $\inf_{P \in \mathcal{A}_\lambda} |\rho_{k\ell}^P| > 0$ and $\Sigma^{-1/2}$ is positive definite. Therefore $nT_n \to_{\mathcal{A}_\lambda} \infty$, and so we have

$$\begin{aligned} \inf_{P \in \mathcal{A}_\lambda} P(nT_n > c) &= \inf_{P \in \mathcal{A}_\lambda} (1 - P(nT_n \leq c)) \\ &= - \sup_{P \in \mathcal{A}_\lambda} (P(nT_n \leq c)) - 1) \to 1 \end{aligned}$$

as $n \to \infty$ for all $c \in \mathbb{R}$. From this we have

$$\liminf_{n \to \infty} \inf_{P \in \mathcal{A}_\lambda} E_P(\hat{\Psi}_n) = \liminf_{n \to \infty} \inf_{P \in \mathcal{A}_\lambda} P(nT_n > z_{1-\alpha}) = 1$$

for all $\alpha \in (0, 1)$. □

Appendix B. Modes of Stochastic Convergence

Let $\mathcal{M}$ denote some class of distributions. We start by defining the notions of small and big O in probability.

B.1 Small and big-O in probability

All sequences (a_n) and (b_n) below are assumed to be non-zero.

Definition 31 *Let (X_n) and (a_n) be sequences of random variables in $\mathbb{R}$. If for every $\varepsilon > 0$*

$$\sup_{P \in \mathcal{M}} P(|X_n/a_n| > \varepsilon) \to 0$$

for $n \to \infty$ then we say that X_n is small O of a_n in probability uniformly over $\mathcal{M}$ and write $X_n \in o_{\mathcal{M}}(a_n)$. If for every $\varepsilon > 0$ there is $M > 0$ such that

$$\sup_{n \in \mathbb{N}} \sup_{P \in \mathcal{M}} P(|X_n/a_n| > M) < \varepsilon$$

then we say that X_n is big O of a_n in probability uniformly over $\mathcal{M}$ and write $X_n \in \mathcal{O}_{\mathcal{M}}(a_n)$.

When $X_n \in \mathcal{O}_{\mathcal{M}}(a_n)$ we also say that X_n is stochastically bounded by a_n uniformly over $\mathcal{M}$. When $X_n \in o_{\mathcal{M}}(1)$ we will typically write $X_n \to_{\mathcal{M}} 0$.

Lemma 32 *Let $(X_n), (a_n)$ and (b_n) be sequences of random variables in $\mathbb{R}$ such that $X_n \in \mathcal{O}_{\mathcal{M}}(a_n)$. Then it holds that $b_n X_n \in \mathcal{O}_{\mathcal{M}}(a_n b_n)$.*

Lemma 33 *Assume that $X_n \in \mathcal{O}_{\mathcal{M}}(a_n)$ and $Y_n \in \mathcal{O}_{\mathcal{M}}(b_n)$. Then $X_n Y_n \in \mathcal{O}_{\mathcal{M}}(a_n b_n)$.*

Lemma 34 *Assume $X_n \in \mathcal{O}_{\mathcal{M}}(a_n)$ and that $a_n \in o(1)$. Then $X_n \in o_{\mathcal{M}}(1)$.*

Lemma 35 *Assume that $X_n \in o_{\mathcal{M}}(1)$ and that $|X_n| \leq C$ for all $n \geq 1$ for a constant C that does not depend on P. Then $\sup_{P \in \mathcal{M}} E_P |X_n| \to 0$ for $n \to \infty$.*

We now turn to uniform convergence in distribution.

B.2 Uniform convergence in distribution

We follow Kasy (2019) and Bengs and Holzmann (2019).

Definition 36 *Let $X, X_1, X_2, \ldots$ be real valued random variables with distribution determined by $P \in \mathcal{M}$. If it holds that*

$$\sup_{P \in \mathcal{M}} |E_P(f(X_n)) - E_P(f(X))| \to 0$$

for $n \to \infty$ for all functions $f : \mathbb{R} \to \mathbb{R}$ that are bounded and continuous, then we say that (X_n) converges uniformly in distribution to X over $\mathcal{M}$. In this case we write $X_n \Rightarrow_{\mathcal{M}} X$.

Lemma 37 (Uniform Slutsky's Lemma) *Assume that $X_n \Rightarrow_{\mathcal{M}} X$ and that $Y_n \to_{\mathcal{M}} 0$. Then $X_n + Y_n \Rightarrow_{\mathcal{M}} X$.*

Proof See Bengs and Holzmann (2019) Theorem 6.3. ■

References

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114(526):749–758, 2019.

Viktor Bengs and Hajo Holzmann. Uniform approximation in classical weak convergence theory. *arXiv preprint arXiv:1903.09864*, 2019.

Wicher Bergsma. *Testing conditional independence for continuous random variables*. Eurandom, 2004.

Wicher Bergsma. Nonparametric testing of conditional independence by means of the partial copula. *SSRN Electronic Journal*, 01 2011.

Taoufik Bouezmarni, Jeroen VK Rombouts, and Abderrahim Taamouti. Nonparametric copula-based test for conditional independence with applications to Granger causality. *Journal of Business & Economic Statistics*, 30(2):275–287, 2012.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

Uwe Einmahl and David M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380 – 1403, 2005.

Jianqing Fan, Yang Feng, and Lucy Xia. A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*, 218(1):119–139, 2020.

Irène Gijbels, Marek Omelka, and Noël Veraverbeke. Estimation of a copula when a covariate affects only marginal distributions. *Scandinavian Journal of Statistics*, 42(4): 1109–1126, 2015.

Ingrid Hobæk Haff and Johan Segers. Nonparametric estimation of pair-copula constructions with the empirical pair-copula. *Computational Statistics & Data Analysis*, 84:1–13, 2015.

Naftali Harris and Mathias Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(1):3365–3383, 2013.

Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.

Maximilian Kasy. Uniformity and the delta method. *Journal of Econometric Methods*, 8 (1), 2019.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

Roger Koenker. *quantreg: Quantile Regression*, 2021. URL `https://CRAN.R-project.org/package=quantreg`. R package version 5.83.

Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of Quantile Regression*. CRC press, 2017.

Steffen L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

Qi Liu, Chun Li, Valentine Wanga, and Bryan E Shepherd. Covariate-adjusted Spearman's rank correlation with probability-scale residuals. *Biometrics*, 74(2):595–605, 2018.

Rohit K Patra, Bodhisattva Sen, and Gábor J Székely. On a nonparametric notion of residual and its applications. *Statistics & Probability Letters*, 109:208–213, 2016.

Andrew J Patton. Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2):527–556, 2006.

Judea Pearl. *Causality*. Cambridge University Press, second edition, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

Joseph D Ramsey. A scalable conditional independence test for nonlinear, non-Gaussian data. *arXiv preprint arXiv:1401.5031*, 2014.

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020.

Kyungchul Song. Testing conditional independence via Rosenblatt transforms. *The Annals of Statistics*, 37(6B):4011–4045, 2009.

Fabian Spanhel and Malte S Kurz. The partial vine copula: A dependence measure and approximation based on the simplifying assumption. *arXiv preprint arXiv:1510.06971*, 2015.

Fabian Spanhel and Malte S. Kurz. The partial copula: Properties and associated dependence measures. *Statistics & Probability Letters*, 119:76–83, 2016.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

Winfried Stute et al. On almost sure convergence of conditional empirical distribution functions. *The Annals of Probability*, 14(3):891–901, 1986.

4.1 The role of uniform level and power

Let us discuss the role of the asymptotic uniform level and power results presented in Section 4.5 of the paper. In the context of constraint-based causal structure learning, this is related to the notion of uniform consistency in causal inference discussed in Robins et al. [2003]. Here the authors argue that asymptotic uniform level and power is necessary in order to link large sample properties of the test to finite sample properties.

More specifically, let $\mathcal{P}$ be a set of distributions, $\mathcal{H} \subset \mathcal{P}$ a hypothesis and $\mathcal{A} \subset \mathcal{P} \setminus \mathcal{H}$ a set of alternatives. Also let Ψ_n denote the observed value of a test based on a sample of size $n \geq 1$. Assuming that Ψ has asymptotic uniform level over $\mathcal{H}$ and power against $\mathcal{A}$, then for every $\varepsilon > 0$ there exist a sample size $n_0(\varepsilon)$ such that $P(\Psi_n = 1) \leq \varepsilon$ for all $P \in \mathcal{H}$ and $P(\Psi_n = 0) \leq \varepsilon$ for all $P \in \mathcal{A}$ for all $n \geq n_0(\varepsilon)$. In other words, we are guaranteed the existence of a sample size for which we can bound the probability of wrongly rejecting a true hypothesis or wrongly accepting a false hypothesis. On the contrary, if we only have asymptotic pointwise level over $\mathcal{H}$ and power against $\mathcal{A}$, then the sample size $n_P(\varepsilon)$ for which we can bound the probability of making a wrong decision depends on the distribution P. Since the true distribution P_n could potentially depend on the sample size, we could be in a situation where $P_n \in \mathcal{H}$ for each $n \geq 1$ but $P_n \to P_\infty \in \mathcal{A}$ as $n \to \infty$. Thus, even for large sample size, we are not guaranteed to have control over the probability of accepting a false hypothesis.

Since constraint-based causal structure learning algorithms use conditional independence tests to construct the skeleton of the Markov equivalence class of interest, asymptotic uniform level and power of the conditional independence test is necessary for uniform consistent estimation of the equivalence class.

The way we ensure asymptotic uniform level and power of our test can be compared to the λ-strong faithfulness condition given in Zhang and Spirtes [2002]. Let X be a random variable which follows a multivariate Gaussian distribution P. Let $\rho_{ab|C}$ denote the partial correlation of X_a and X_b given $X_C = (X_c)_{c \in C}$. Then P is called λ-strong faithful to a directed acyclic graph $\mathcal{G}$ if for all $a, b \notin C$ with $a \neq b$ it holds that X_a and X_b are d-connected given X_C in $\mathcal{G}$ if and only if $|\rho_{ab|C}| > \lambda$. Note that 0-strong faithfulness just corresponds to usual faithfulness of a multivariate Gaussian distribution to a directed acyclic graph, so it is a strengthening of the notion of faithfulness. Our restriction to the set of alternatives $\mathcal{A}_\lambda$, defined in Section 4.5 of the paper, where the generalized correlation ρ has an element which is bounded away from zero, i.e., $|\rho_{k\ell}| > \lambda$ for some $k, \ell = 1, \ldots, q$, is analogous to the λ-strong faithfulness condition.

Zhang and Spirtes [2002] prove that a uniformly consistent estimator for the Markov equivalence class exists given that the true distribution is Markov and λ-strong faithful to a directed acyclic graph and that causal sufficiency is satisfied. Kalisch and Bühlmann [2007] extends this result to a high dimensional setting, where they assume sparsity of the true graph, but additionally allows for the λ in the λ-strong faithfulness condition to depend on the sample size, such that $\lambda_n \to 0$ at a slow rate.

We hypothesize that it will be possible to show uniform consistency of a PC-algorithm that uses our partial copula based conditional independence test by assuming that the true distribution P belongs to $\mathcal{P}_0$ where the quantile regressions are uniformly consistent, as given in Assumption 2 of the paper, and P is Markov, faithful and causally sufficient with respect to a directed acyclic graph $\mathcal{G}$ such that whenever X_a and X_b are d-connected given X_C in $\mathcal{G}$ then the joint distribution (X_a, X_b, X_C) belong to $\mathcal{A}_\lambda$ for some $\lambda \in (0, 1)$. However, this is the topic of future research.

4.2 Implementation of the test

An implementation of the test as described in Section 4.7 of the paper in the R programming language [R Core Team, 2021] is available at `https://github.com/lassepetersen/partial-copula-CI-test`. Here we will briefly showcase the functionality of the implementation and discuss further possible improvements. The core of the implementation is the function **test_CI**, which is implemented as described in Section 4.7 of the paper. At the time of writing the available quantile regression models include linear models, polynomial basis models and B-spline basis models. Example usage of the function can be seen below:

```
set.seed(1)

N    <- 100
Z    <- rnorm(N)
eps1 <- rnorm(N)
eps2 <- rnorm(N)
X    <- 2 * Z + eps1
Y    <- 5 * Z + eps2

test_results <- test_CI(X = X, Y = Y, Z = Z, alpha = 0.05,
                        q = c(1, 2, 3), quantile_reg = 'B-Spline', bspline_df = 3)

print(test_results)
-------------------------------------------------------------------------------------
$statistic
[1] 0.6104795 4.3223138 4.1058109

$p_value
[1] 0.4346073 0.3641393 0.9043112

$test_decision
[1] 0 0 0

$q_vals
[1] 1 2 3
```

The test returns the test statistics, p-values and test decisions based on the chosen significance level α for each provided value of q. One can also tell the function to return all information related to performing the test. Setting `return_all = TRUE`, **test_CI** also returns the predicted values of the quantile regression models for doing model diagnostics, and the estimated nonparametric residuals $(\hat{U}_{1,i}, \hat{U}_{2,i})_{i=1}^{n}$. Having direct access to the estimated nonparametric residuals, one can visually inspect the dependence between them. Also, one can experiment with using alternative tests for independence in the partial copula.

```
test_results <- test_CI(X = X, Y = Y, Z = Z, alpha = 0.05,
                        q = c(1, 2, 3), quantile_reg = 'B-Spline', bspline_df = 3,
```

```
                        return_all = TRUE)

head(cbind(test_results$U1, test_results$U2), n=10)
-------------------------------------------------------------------------------
           [,1]       [,2]
 [1,] 0.5001913 0.52318487
 [2,] 0.5658928 0.99000000
 [3,] 0.2584495 0.88111111
 [4,] 0.7289587 0.22777778
 [5,] 0.1921994 0.08003584
 [6,] 0.9900000 0.98649876
 [7,] 0.8047336 0.89497542
 [8,] 0.8201127 0.82772475
 [9,] 0.6325964 0.55444444
[10,] 0.8811111 0.69203920
```

Further work on the implementation includes the opportunity to add penalization to the quantile regression as described in Section 3.5 of the paper, and also performing quantile regression using non-parametric machine learning models such as neural networks. Moreover, since the intended use of the test is for constraint-based causal structure learning, we plan on writing wrappers for the **test_CI** functions such that they can be used together with popular structure learning packages such as **pcalg** [Kalisch et al., 2012] and **bnlearn** [Scutari, 2010].

4.3 More general independence tests

Throughout the paper we restrict our attention to measures of dependence in the partial copula of the form $\rho = E_P(\varphi(U_1)\varphi(U_2)^T)$ with test statistic given by

$$\hat{\rho}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(\hat{U}_{1,i}^{(n)})\varphi(\hat{U}_{2,i}^{(n)})^T, \tag{4.1}$$

where $\varphi : [0,1] \to \mathbb{R}^q$ satisfies Assumption 4 of the paper. We always have that $U_1 \perp\!\!\!\perp U_2$ implies $\rho = 0$, but as we emphasize throughout the paper, the reverse implication does not always hold true. In this section we empirically examine the performance of the partial copula conditional independence test if we combine it with more general independence tests, while still using our quantile regression based conditional distribution function estimator.

Hoeffding's independence test [Hoeffding, 1948] and the Hilbert-Schmidt Independence Criterion (HSIC) test [Gretton et al., 2008, Pfister et al., 2018] are independence tests that both have power against general alternatives of independence. Let us briefly review the ideas behind these two tests.

Let X and Y be continuous real valued random variable, and let F_X, F_Y and $F_{X,Y}$ denote the conditional distribution functions of X, Y and (X, Y) respectively. Define

$$\Delta = \int D(x,y)^2 \mathrm{d}F(x,y) \quad \text{where} \quad D(x,y) = F_{X,Y}(x,y) - F_X(x)F_Y(y).$$

Here the integral is a Lebesgue-Stieltjes integral. The test exploits that $F_{X,Y} = F_X F_Y$ under independence, and one can show that $\Delta = 0$ if and only if $X \perp\!\!\!\perp Y$. Note that this is not necessarily true if (X, Y)

has discontinuities. See e.g. Blum et al. [1961] for an extension to random variables with discontinuities. However, for our application the standard Hoeffding's independence test suffices. The dependence measure Δ is estimated using a U-statistic, which has a known asymptotic distribution under independence. We will use the implementation given in the **DescTools** R package [Andri et mult. al., 2021].

The HSIC test is more involved to describe. The HSIC dependence measure is defined as the Hilbert-Schmidt norm of the so-called cross-covariance operator $C_{xy} : \mathcal{G} \to \mathcal{F}$ between reproducing kernel Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ on the outcome spaces of X and Y respectively. The full description is technical, so we refer to Gretton et al. [2008] for details. Most importantly, the HSIC dependence measure is zero if and only if X and Y are independent. The HSIC dependence measure is estimated using a V-statistic, and its asymptotic distribution under independence can be approximated by a Γ-distribution. Alternatively, the test can be carried out as a permutation test at a greater computational cost. An extension to test for joint independence of multiple random variables $X_1, \ldots, X_d$, known as the dHSIC test, is given in Pfister et al. [2018]. We will use the implementation given in the **dHSIC** R package [Pfister and Peters, 2019] even though we are only dealing with the case $d = 2$.

We will evaluate the soundness of using Hoeffding's independence test and the HSIC test with the partial copula by repeating the simulation studies in Sections 5.4 and 5.5 of the paper. In addition to the two partial copula based test, we also examine the effect of applying the Generalised Covariance Measure (GCM) with squared responses $X' = X^2$ and $Y' = Y^2$, since this could potentially increase the power of the test against alternatives with variance heterogeneity. Here we make sure to fit the correct mean regression models $f(z) = E(X' \mid Z = z)$ and $g(z) = E(Y' \mid Z = z)$ of the transformed responses given Z for a fair comparison. We also include the results of the partial copula test with generalized correlation $\hat{\Psi}_n$ studied throughout the paper with $q \in \{1, 3\}$ for reference.

First consider the simulation results seen in Figure 4.1, where we assess the asymptotic level and power of the tests using the same methodology as in Section 5.4 of the paper. First we observe that the partial copula test with Hoeffding's independence test has power against all alternatives $\mathrm{A}_2, \mathrm{A}_3$ and A_4, but that it fails to maintain level over $\mathrm{H}_2, \mathrm{H}_3$ and H_4. Second, we see that the partial copula test with the HSIC independence test has power against all $\mathrm{A}_2, \mathrm{A}_3$ and A_4, and has valid level over H_2 and H_3, while having some issues maintaining level over H_4. Finally, the GCM with squared responses maintains level in all cases, and has power against A_2 and A_3, however, it does not increase the power compared to the standard GCM. It does not have power against A_4.

Next we consider the simulation results seen in Figure 4.2, where we investigate the power of the tests against the local alternative described in Section 5.5 of the paper. We observe that the partial copula test with Hoeffding's independence test and the HSIC test provide better low sample power than $\hat{\Psi}_n$ with $q = 1$, while holding level under conditional independence $\gamma = 0$. Since we are now fitting mean regression models in order to apply the GCM with squared responses, the power of the test improves with increasing sample size. However, we observe that transforming the responses still does not yield power against the local alternative, when there is a large degree of variance heterogeneity between X and Y given Z. We cannot rule out another transformation of the responses, which could yield a greater power, but choosing such a transformation adds a layer of ad-hoc decisions to be made when performing the GCM.

In conclusion, our simulation study suggests that the partial copula conditional independence test can be combined with more general independence tests to yield greater power, but that it can come at the expense of loss of level. Here we saw that the partial copula test did not maintain level when combining it with Hoeffding's independence test, while it had minor problems holding level when combining it with the HSIC test under H_4. The explanation lies in the more complicated U- and V-statistic used

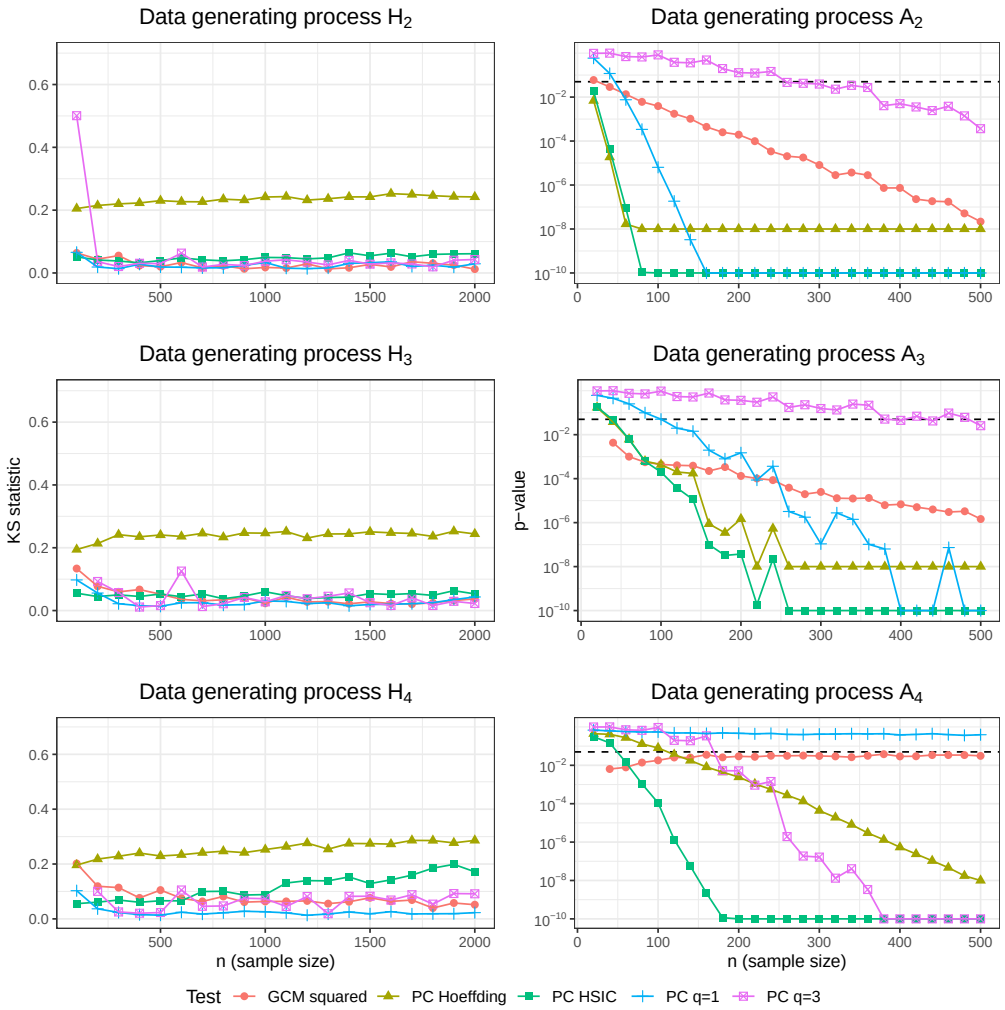

Figure 4.1: The result of repeating the simulation study in Section 5.4 with the addition of applying GCM to squared responses (GCM squared) and combining the partial copula test with Hoeffding's independence test (PC Hoeffding) and the HSIC test (PC HSIC). Note that the p-values of Hoeffding's independence test is truncated at 10^{-8}, which is due to its implementation in the **DescTools** R package.

for estimating the dependence measures, which do not have a representation of the form (4.1). The challenge is to control the nested estimation uncertainty involved with estimating the nonparametric residuals $(U_{1,i})_{i=1}^n$ and $(U_{2,i})_{i=1}^n$, and then plugging these estimates $(\hat{U}_{1,i})_{i=1}^n$ and $(\hat{U}_{2,i})_{i=1}^n$ into the test statistics for the dependence measure. The representation (4.1) has several advantages in this regard. Consider the decomposition of the test statistic,

$$\hat{\rho}_n = \alpha_n + \beta_n + \gamma_n + \delta_n,$$

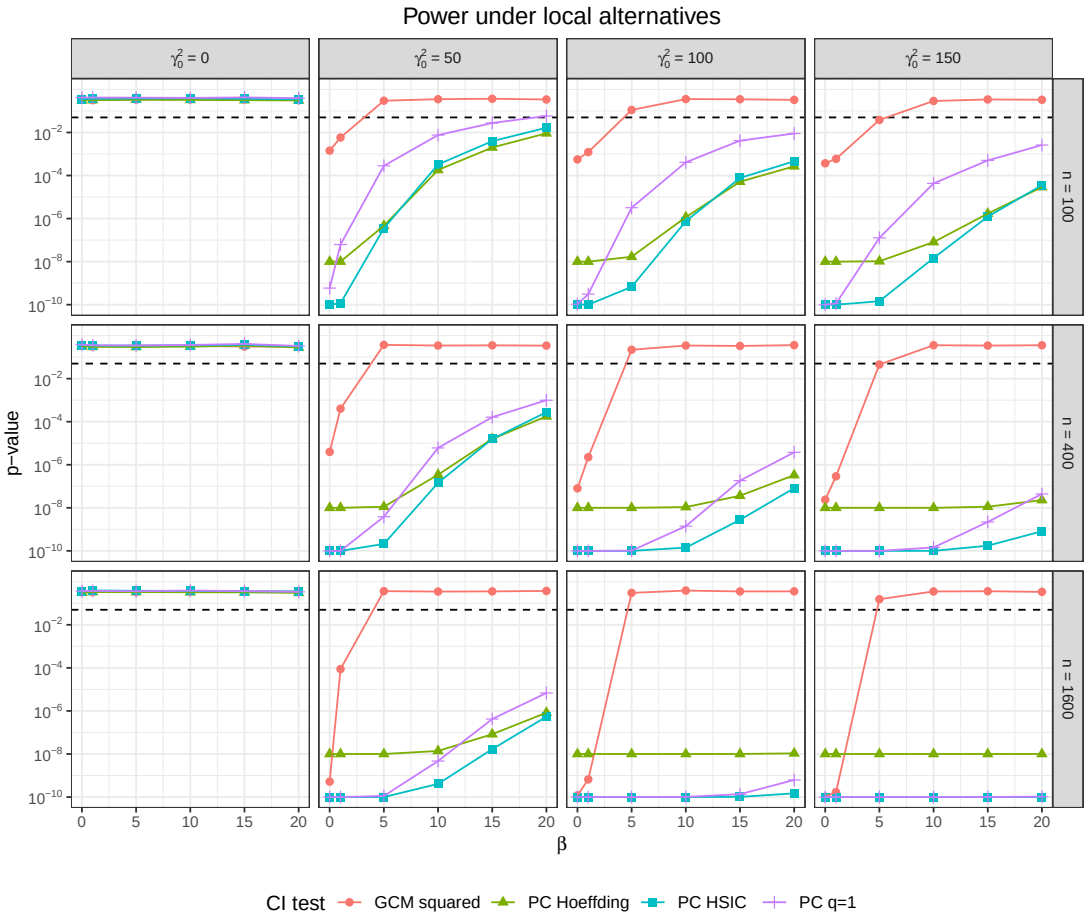

Figure 4.2: A repetition of the simulation study described in Section 5.5. In addition, we have applied GCM to squared responses (GCM squared) and combining the partial copula test with Hoeffding's independence test (PC Hoeffding) and the HSIC test (PC HSIC). Note that the p-values of Hoeffding's independence test is truncated at 10^{-8}, which is due to its implementation in the **DescTools** R package.

where

$$\alpha_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(U_{1,i})\varphi(U_{2,i})^T,$$

$$\beta_n = \frac{1}{n}\sum_{i=1}^{n} \left(\varphi(\hat{U}_{1,i}) - \varphi(U_{1,i})\right)\left(\varphi(\hat{U}_{2,i}) - \varphi(U_{2,i})\right)^T,$$

$$\gamma_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(U_{1,i})\left(\varphi(\hat{U}_{2,i}) - \varphi(U_{2,i})\right)^T,$$

$$\delta_n = \frac{1}{n}\sum_{i=1}^{n} \left(\varphi(\hat{U}_{1,i}) - \varphi(U_{1,i})\right)\varphi(U_{2,i})^T,$$

which we used in the proof of Theorem 14 of the paper. The first advantage of this decomposition is the separation into a term, α_n, which does not involve estimation uncertainty of the nonparametric residuals and drives the asymptotic distribution of the test statistic, and remainder terms, β_n, γ_n and δ_n, which do involve estimation uncertainty, but only need to converge to zero in probability (at rate faster than

$1/\sqrt{n}$). Secondly, the product structure of the test statistic gives a product structure of the remainder term β_n, which is responsible for the rate condition $\sqrt{n} g_P(n) h_P(n) \to 0$ of the theorem. The product structure of the test statistic and the decomposition above lie at the heart of the technique used for analyzing double machine learning estimators.

Since U- and V-statistics do not have this product structure, we cannot control the nested estimation uncertainty in the same way. As we mention in the discussion, it is ongoing work to establish an analogous asymptotic theory for more complicated test statistics, but our simulations suggest that it cannot in general be proved about all U-statistics under the same conditions as in Theorem 14 of the paper, due to our negative simulation results regarding Hoeffding's independence test. However, our simulations also suggest that it could be safe to combine the partial copula with the HSIC independence test, but we cannot provide a theoretical guarantee of this statement.

Chapter 5

Local Independence Testing

Lasse Petersen and Niels Richard Hansen. Nonparametric conditional local independence testing. 2021a.

NONPARAMETRIC CONDITIONAL LOCAL INDEPENDENCE TESTING

LASSE PETERSEN AND NIELS RICHARD HANSEN

Abstract. Conditional local independence is an independence relation among continuous time stochastic processes. It describes how the infinitesimal evolution of one process is unaffected by another process given the histories of additional processes, and it is important for the description and learning of causal relations among processes. This paper proposes a nonparametric test for conditional local independence based on double machine learning. We construct a stochastic process as a stochastic integral, which is a zero-mean, local martingale under the hypothesis of conditional local independence. We derive the weak limit of its test statistic process under the hypothesis, which we show is a Gaussian martingale with a variance function that can be estimated from data. Based on the limiting Gaussian martingale, we propose test statistics for the hypothesis as finite dimensional functionals of the test statistic process.

1. Introduction

Conditional local independence was introduced by Schweder (1970) for continuous time composable Markov processes. It is a formalization of how the evolution of one stochastic process depends on the past of other processes in a dynamical system. As such, it is closely related to Granger causality (Granger 1969), which has been popular in econometrics and for the analysis of time series data.

In words, a process N_t is *conditionally locally independent* of a process Z_t given a history $\mathcal{F}_t$ (a filtration) if Z_t does not add any predictable information to $\mathcal{F}_t$ about the infinitesimal evolution of N_t. Moreover, with a structural assumption about the stochastic processes, a causal interpretation of conditional local independence is possible (Aalen 1987, Aalen et al. 2012, Commenges & Gégout-Petit 2009). Being able to test conditional local independence is therefore an important tool for investigating causal relations. In the context of time series analysis, such a test is known as a test of Granger non-causality.

A systematic investigation of algebraic properties of conditional local independence was initiated by Didelez (2006, 2008, 2015). She introduced graphical representations of conditional local independence using directed graphs and studied the semantics of such *local independence graphs*. This work was extended further by Mogensen & Hansen (2020) to graphical representations of partially observed systems, which is of importance for causal representations of systems with e.g. unobserved confounders.

A constraint based learning algorithm of local independence graphs was given by Mogensen et al. (2018) in terms of a conditional local independence oracle, and the learning of models for multivariate dynamic event systems with a causal interpretation has also caught some attention in the machine learning community, (Xu et al. 2016, Achab et al. 2017, Xiao et al. 2019). In

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

E-mail addresses: lp@math.ku.dk, Niels.R.Hansen@math.ku.dk.

practice, conditional local independence is then discovered from data by either sparsity inducing learning algorithms or statistical tests.

One challenge is that any particular model class may be misspecified. For the popular class of linear Hawkes processes, conditional local independence is equivalent to a kernel being zero, and a statistical test can readily be based on parametric or nonparametric estimation of kernels within this model class. However, model misspecification will generally invalidate tests of conditional local independence based on the model. This is compounded by the model class of Hawkes processes not being closed under marginalization (also known as non-collapsability), which means that even within a subsystem of a linear Hawkes process, conditional local independence cannot be tested correctly using a Hawkes process model.

A simple example of how model misspecification affects conditional local independence testing is presented in Section 1.1 based on Cox's survival model. The non-collapsability of that model illustrates the need of a completely nonparametric test of conditional local independence.

We propose a new test based on the ideas of double machine learning (Chernozhukov et al. 2018). The main novelty is that within our time-dynamic framework the test is based on an infinite dimensional parameter – a function of time. We model the target process N_t as well as the covariate process Z_t conditionally on $\mathcal{F}_t$. As we show, we need to learn the predictable projection of Z_t onto $\mathcal{F}_t$ to achieve the orthogonalization at the core of double machine learning. If we can learn this model at rate $g(n)$ and the model of N_t at rate $h(n)$, the main result states that our test statistic converges under conditional local independence to a Gaussian martingale at rate $n^{-\frac{1}{2}}$ if $\sqrt{n}g(n)h(n) \to 0$ for $n \to \infty$.

1.1. **A Cox model with a partially observed covariate process.** To motivate the importance of nonparametric local independence testing and the benefits of our proposed solution, we consider an example based on Cox's survival model with time dependent covariates. In this example, we consider a model of death time τ, which depends on three time-dynamic processes X, Y and Z.

An interpretation of the processes is as follows:

$$X = \text{BMI}$$
$$Y = \text{Blood pressure}$$
$$Z = \text{Pension savings}$$

Periods of overweight or obesity may have long-term effects on blood pressure, and due to e.g. job market discrimination, high BMI could also affect pension savings negatively. Death risk is affected directly by BMI and blood pressure but not the size of your pension savings. Figure 1 illustrates the dependence structure among the processes and the death time as a local independence graph.

We assume that $\tau \in [0, T]$ for a fixed $T > 0$ and that X, Y and Z have continuous sample paths. The Cox model of death is given via the intensity

$$\lambda_t^{\text{full}} = 1_{[0,\tau)}(t)\lambda_t^0 e^{\beta_1 X_t + Y_t} \tag{1}$$

with λ_t^0 a baseline intensity. Let also

$$\mathcal{F}_t^X = \sigma(\tau > t, X_s; s \leq t)$$

denote the filtration generated by τ and the X-process, and similarly for other processes and combinations of processes. That is, $\mathcal{F}_t^{X,Y,Z}$ is the filtration generated by τ and all three X-, Y-,

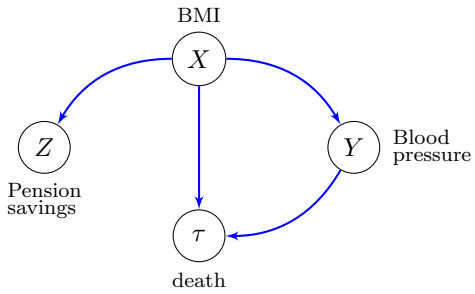

FIGURE 1. Local independence graph illustrating how the three processes X, Y, and Z affect each other and time of death in the toy example. Death is conditionally locally independent of Z given X in this example.

and Z-processes. By definition, λ_t^{full} is the $\mathcal{F}_t^{X,Y,Z}$-intensity based on the full history of all three processes. It is not important that λ_t^{full} is a Cox model for our general procedure, but it allows for certain theoretical computations in this example.

The fact that λ_t^{full} does not depend upon Z means that *death is conditionally locally independent* of Z given $\mathcal{F}_t^{X,Y}$. We ask if death is conditionally locally independent of Z given only $\mathcal{F}_t^X$? That is, with Y unobserved we want to know if the intensity of death given the history of the X-process depends on Z. The intensity given the observed history $\mathcal{F}_t^{X,Z}$ is

$$\lambda_t^{\text{marg}} = E(\lambda_t^{\text{full}} \mid \mathcal{F}_t^{X,Z}),$$

and the hypothesis we are interesting in can be formulated as

$$H_0 : \lambda_t^{\text{marg}} \text{ does not depend on } Z. \tag{2}$$

We could investigate the hypothesis H_0 via a marginal Cox model

$$\lambda_t^{\text{marg-cox}} = 1_{[0,\tau)}(t)\lambda_t^0 e^{\beta_1 X_t + \beta_2 Z_t} \tag{3}$$

and test if $\beta_2 = 0$, but the Cox model is non-collapsable, even if Z is not a confounder (Martinussen & Vansteelandt 2013), and this semi-parametric model is quite likely misspecified. Consequently, the test of $\beta_2 = 0$ is not equivalent to a test of conditional local independence.

Knowledge about the joint distribution of the processes beyond (1) is needed to decide if death is conditionally locally independent of Z given $\mathcal{F}_t^X$. The local independence graph in Figure 1 encodes such knowledge in terms of the node X blocking the paths between the node Z and the nodes Y and τ.

With the Cox model given by (1), a sufficient additional condition for (2) to hold is that

$$\sigma(Y_s; s \leq t) \perp\!\!\!\perp \sigma(Z_s; s \leq t) \mid \mathcal{F}_t^X, \tag{4}$$

that is, knowing the entire history of the X-process (and whether $\tau > t$), the Z- and Y-process histories are independent. To see this, note that (4) implies

$$E(e^{Y_t} \mid \mathcal{F}_t^{X,Z}) = E(e^{Y_t} \mid \mathcal{F}_t^X),$$

thus the $\mathcal{F}_t^{X,Z}$-intensity is

$$\lambda_t^{\text{marg}} = E(\lambda_t^{\text{full}} \mid \mathcal{F}_t^{X,Z}) = 1_{[0,\tau)}(t)\lambda_t^0 e^{\beta_1 X_t} E(e^{Y_t} \mid \mathcal{F}_t^X),$$

and since it does not depend on Z, we conclude that death is conditionally locally independent of Z given $\mathcal{F}_t^X$.

The factor $E(e^{Y_t} \mid \mathcal{F}_t^X)$ in λ_t^{marg} indicates that the dependence on the X-history is not only via X_t but via its entire history. If $Z_t \not\perp\!\!\!\perp Y_t \mid X_t$, which can be true even when (4) holds, Z_t is predictive of Y_t conditionally on X_t, and Z_t will act as a proxy of Y_t in the marginal Cox model (3). This illustrates the non-collapsability of the Cox model.

With our toy interpretation of the three processes, the Cox model (1) implies that BMI at time t has a direct effect on the risk of death at time t, but the entire history of BMI has an indirect effect mediated via blood pressure at time t. In the absence of measuring blood pressure, and conditioning on the value of BMI at time t only, pension savings act as a proxy of blood pressure – via its dependence on blood pressure through the history of BMI. When (4) holds, meaning that blood pressure and pension savings are independent given the full BMI-history, conditioning on the BMI-history will render death conditionally locally independent of pension savings.

Instead of testing H_0 within a parametric model of λ_t^{marg}, we will consider a nonparametric test. We introduce

$$\lambda_t = E(\lambda_t^{\text{full}} \mid \mathcal{F}_t^X)$$

as the $\mathcal{F}_t^X$-intensity, and H_0 is then equivalent to $\lambda_t = \lambda_t^{\text{marg}}$. A test of conditional local independence can then be based on the parameter

$$\gamma = E\left(Z_\tau - \int_0^T Z_s \lambda_s \mathrm{d}s\right),$$

which under H_0 is zero, see Proposition 2.2. Whence conditional local independence implies $\gamma = 0$.

With i.i.d. observations $(\tau_1, X_1, Z_1), \ldots, (\tau_n, X_n, Z_n)$ and (nonparametric) estimates, $\hat{\lambda}_{j,t}$, based on $(\tau_1, X_1), \ldots, (\tau_n, X_n)$, we can compute the plug-in estimate

$$\hat{\gamma}_{\text{plug-in}} = \frac{1}{n}\sum_{j=1}^n \left(Z_{j,\tau_j} - \int_0^T Z_{j,s}\hat{\lambda}_{j,s}\mathrm{d}s\right).$$

However, it is well known that to achieve $\sqrt{n}$-rate convergence of this estimator, we generally need Donsker class conditions on the model of the intensity, and low variance but biased estimators of λ can lead to severe bias of $\hat{\gamma}_{\text{plug-in}}$, see e.g. Chernozhukov et al. (2018).

Our proposal is based on the double machine learning idea by Chernozhukov et al. (2018). Defining $\Pi_s = E(Z_s \mid \mathcal{F}_{s-}^X)$, we show that

$$\gamma = E\left(Z_\tau - \Pi_\tau - \int_0^T (Z_s - \Pi_s)\lambda_s \mathrm{d}s\right),$$

which follows from Π being $\mathcal{F}_t^X$-predictable by construction, see Section 2.1. Plugging in two nonparametric estimators gives the estimator

$$\hat{\gamma}_{\text{double}} = \frac{1}{n}\sum_{j=1}^n \left(Z_{j,\tau_j} - \hat{\Pi}_{j,\tau_j} - \int_0^T (Z_{j,s} - \hat{\Pi}_{j,s})\hat{\lambda}_{j,s}\mathrm{d}s\right).$$

To achieve a small bias and a $\sqrt{n}$-rate of convergence, we will use data splitting. With e.g. a total of $2n$ observations, the nonparametric estimates $\hat{\Pi}_j$ and $\hat{\lambda}_j$ will be based on the second half only, and be independent of the first half of the sample used for testing. The details are given in Section 2.2.

1.2. **Organization.** The organization of the paper is as follows. In Section 2 we introduce the mathematical framework for formulating the basic hypothesis of conditional local independence. We introduce a test statistic as a stochastic process and describe how sample splitting is to be used for its computation via the estimation of two unknown components. In Section 3 we state the main result of the paper, which is a functional CLT of the test statistic process under weak rate conditions on the estimators. Section 4 elaborates on the motivating Cox model example above. Proofs and auxiliary results are in Appendix A.

2. Setup

We consider a counting process $N = (N_t)$ and another caglad real value process $Z = (Z_t)$, both defined on the probability space $(\Omega, \mathcal{F}, P)$. All processes considered are assumed to have their time index in the compact interval $[0, T]$ for a fixed $T > 0$ unless otherwise specified. We will assume that N is adapted w.r.t. a right continuous and complete filtration $\mathcal{F}_t$, and we denote by $\mathcal{G}_t$ the right continuous and complete filtration generated by $\mathcal{F}_t$ and Z_t.

In the survival example of the introduction, $N_t = 1(\tau \leq t)$ is the indicator of whether death has happened by time t, and there can only be one event per individual observed. Furthermore, $\mathcal{F}_t = \mathcal{F}_t^X$ and $\mathcal{G}_t = \mathcal{F}_t^{X,Z}$. Our general setup works for any counting process, thus it allows for recurrent events, and the filtration $\mathcal{F}_t$ can contain the histories of any number of processes in addition to the history of N itself.

2.1. **The hypothesis of conditional local independence.** The counting process N is assumed to have an $\mathcal{F}_t$-intensity λ_t, that is, λ_t is $\mathcal{F}_t$-predictable and with

$$\Lambda_t = \int_0^t \lambda_s \mathrm{d}s$$

being the compensator of N then

$$M_t := N_t - \Lambda_t \tag{5}$$

is a local $\mathcal{F}_t$-martingale. With this framework we can define the hypothesis of conditional local independence precisely.

Definition 2.1 (Conditional local independence)**.** We say that N is conditionally locally independent of Z given $\mathcal{F}_t$ if the local $\mathcal{F}_t$-martingale M defined by (5) is also a local $\mathcal{G}_t$-martingale.

For simplicity, we will refer to this hypothesis as simply *local independence* and write

$$H_0 : M \text{ is a local } \mathcal{G}_t\text{-martingale.} \tag{6}$$

As argued in the introduction, the hypothesis of local independence is the hypothesis that observing Z on $[0, t]$ does not add any information to $\mathcal{F}_{t-}$ about whether an N-event will happen in an infinitesimal time interval $[t, t + \delta)$.

A test of the hypothesis of local independence could be based on the stochastic integral

$$\int_0^t Z_s \mathrm{d}M_s,$$

which under the hypothesis is a local martingale. We can also introduce

$$\gamma_t = E\left(\int_0^t Z_s \mathrm{d}M_s\right), \tag{7}$$

provided that the expectation is well defined. If the stochastic integral is a martingale under H_0, $\gamma_t = 0$ for all $t \in [0, T]$.

However, for the reasons presented in Section 1.1, we will base the test on the process

$$I_t = \int_0^t (Z_s - \Pi_s)\,\mathrm{d}M_s \tag{8}$$

where

$$\Pi_s = E(Z_s \mid \mathcal{F}_{s-}) \tag{9}$$

is the predictable projection of the caglad process Z_s, see Theorem VI.19.2 in (Rogers & Williams 2000).

Proposition 2.2. *Under H_0, that is, when N is conditionally locally independent of Z given $\mathcal{F}_t$, the process $I = (I_t)$ with $I_0 = 0$ is a local $\mathcal{G}_t$-martingale. If it is a martingale, $E(I_t) = 0$ for $t \in [0, T]$.*

Note that if Z is $\mathcal{F}_t$-adapted, then $\mathcal{G}_t = \mathcal{F}_t$ and N is trivially locally independent of Z. In this case, we also have that $\Pi = Z$ and $I_t = 0$. The hypothesis is only of interest when $\mathcal{G}_t$ is a strictly larger filtration than $\mathcal{F}_t$, that is, when Z provides information not already in $\mathcal{F}_t$.

Since the predictable projection Π by definition is $\mathcal{F}_t$-predictable and M is a local $\mathcal{F}_t$-martingale,

$$\int_0^t \Pi_s \mathrm{d}M_s$$

is a local $\mathcal{F}_t$-martingale. If it is a martingale and the expectation in (7) is well defined,

$$\gamma_t = E\left(\int_0^t (Z_s - \Pi_s)\mathrm{d}M_s\right) + E\left(\int_0^t \Pi_s \mathrm{d}M_s\right) = E(I_t).$$

Thus the function $t \mapsto \gamma_t$ quantifies deviations from the hypothesis H_0, and our proposal is effectively to test if γ_t is constantly equal to 0 for $t \in [0, T]$.

2.2. **The test statistic.** We will consider the setup where we have observed n i.i.d. replications of the processes, $(N_1, \mathcal{F}_1, Z_1), \ldots, (N_n, \mathcal{F}_n, Z_n)$, where observing $\mathcal{F}_j = (\mathcal{F}_{j,t})$ signifies that anything adapted to the j-th filtration is computable from observations. We randomly split the data into $J_n \subseteq \{1, \ldots, n\}$ and J_n^c, and we let $\hat{\lambda}^{(n)}$ and $\hat{\Pi}^{(n)}$ denote nonparametric estimates of the intensity and the predictable projection, respectively, based on J_n^c only. See Section 4 for an example of nonparametric estimators. Then we define the estimator

$$\hat{I}_t^{(n)} = \frac{1}{|J_n|} \sum_{j \in J_n} \int_0^t \left(Z_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right) \mathrm{d}\hat{M}_{j,s}^{(n)} \tag{10}$$

where $\hat{M}_{j,t}^{(n)} = N_{j,t} - \int_0^t \hat{\lambda}_{j,s}^{(n)} \mathrm{d}s$.

We can regard $\hat{I}_t^{(n)}$ as a double machine learning estimator of γ_t, with the observations indexed by J_n^c used to learn models of λ and Π, and with observations in J_n used to estimate γ_t based on these two models. The test statistic that we will use to test H_0 is the process $\sqrt{|J_n|}\hat{I}^{(n)}$, whose asymptotic distribution under H_0 is derived below.

We note that by an estimate, $\hat{\lambda}^{(n)}$, of λ we mean a function that can be computed on the basis of $\mathcal{F}_{j,t}$ for any j, and similarly for $\hat{\Pi}^{(n)}$. That is, based on data in J_n^c the functional forms of these processes is determined, and these are then applied to compute $\hat{\Pi}_{j,s}^{(n)}$ and $\hat{\lambda}_{j,s}^{(n)}$, that enter into the computation of $\hat{I}^{(n)}$, for $j \in J_n$.

3. The asymptotic distribution of the test statistic

In this section we will derive the asymptotic weak limit of $\sqrt{|J_n|}\hat{I}^{(n)}$ defined by (10) as a stochastic process. In order to do so, we will need the decomposition

$$\sqrt{|J_n|}\hat{I}^{(n)} = U^{(n)} + R_1^{(n)} + R_2^{(n)} + R_3^{(n)} \tag{11}$$

where the processes $U^{(n)}$ and $R_k^{(n)}$, $k = 1, 2, 3$, are given by

$$U_t^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t (Z_{j,s} - \Pi_{j,s}) \, \mathrm{d}M_{j,s} \tag{12}$$

$$R_{1,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t (Z_{j,s} - \Pi_{j,s}) \left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)} \right) \mathrm{d}s \tag{13}$$

$$R_{2,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t \left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)} \right) \mathrm{d}M_{j,s}, \tag{14}$$

$$R_{3,t}^{(n)} = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} \int_0^t \left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)} \right) \left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)} \right) \mathrm{d}s. \tag{15}$$

We now proceed to show that $U^{(n)}$ drives the asymptotic limit of $\sqrt{|J_n|}\hat{I}^{(n)}$, and that the processes $R_k^{(n)}$ can be considered remainder terms, which we will show converge weakly to the zero process. Before doing so, we introduce regularity conditions on the data generating distribution, and consistency assumptions on the estimators $\hat{\lambda}^{(n)}$ and $\hat{\Pi}^{(n)}$. Define also

$$\sigma^2(t) := E\left(\int_0^t (Z_s - \Pi_s)^2 \mathrm{d}N_s \right) = E\left(\int_0^t (Z_s - \Pi_s)^2 \lambda_s \mathrm{d}s \right). \tag{16}$$

From this end we will assume the following on the data generating distribution and its interplay with our choice of auxiliary process Z.

Assumption 3.1.

i) The $\mathcal{F}_t$-intensity λ of N is caglad with $\sup_{0 \le t \le T} \lambda_t \le C$ almost surely.
ii) The process Z is caglad with $\sup_{0 \le t \le T} |Z_t| \le C'$ almost surely.
iii) $\sigma^2(t) < \infty$ for each $t \in [0, T]$.
iv) $E((Z_t - \Pi_t)^2 \lambda_t) < \infty$ for each $t \in [0, T]$.

For a caglad process $X \in L_2(\Omega \times [0, T])$ we let $|||\cdot|||_{\infty,T}^2$ denote the norm

$$|||X|||_{\infty,T}^2 = E\left(\left(\sup_{0 \le t \le T} |X_t| \right)^2 \right).$$

We then make the following consistency assumptions on $\hat{\lambda}^{(n)}$ and $\hat{\Pi}^{(n)}$.

Assumption 3.2. Let

$$g(n) = \left| \left| \left| \Pi - \hat{\Pi}^{(n)} \right| \right| \right|_{\infty,T},$$
$$h(n) = \left| \left| \left| \lambda - \hat{\lambda}^{(n)} \right| \right| \right|_{\infty,T},$$
$$k(n) = \left| \left| \left| \left(\lambda - \hat{\lambda}^{(n)} \right)^2 \right| \right| \right|_{\infty,T}.$$

Then it holds that $\sqrt{|J_n|}g(n)h(n) \to 0$ and $g(n), h(n), k(n) \to 0$ as $n \to \infty$.

Let $\tilde{\mathcal{G}}_t$ denote the smallest filtration that has the filtrations $\{\mathcal{G}_{j,t} \mid j \in J_n, n \in \mathbb{N}\}$ as subfiltrations. First we have the following proposition about the weak limit of the process $U^{(n)}$, where we write $\Longrightarrow$ to denote weak convergence in $D[0,T]$, the space of cadlag function $[0,T] \to \mathbb{R}$, endowed with the Skorokhod topology.

Proposition 3.3. *Assume H_0 and that Assumption 3.1 holds. Then*

$$U^{(n)} \Longrightarrow U$$

where $U = (U_t)_{t\in[0,T]}$ is a zero-mean Gaussian martingale with respect to $\tilde{\mathcal{G}}_t$ with variance function σ^2 given by (16)*.*

Next we turn our attention to the remainder terms $R_k^{(n)}, k = 1, 2, 3$.

Proposition 3.4. *Under H_0 and Assumptions 3.1 and 3.2 it holds that*

$$R_k^{(n)} \Longrightarrow 0$$

for $k = 1, 2, 3$ where 0 is the zero-process.

By combining the two proposition, we have the following theorem, which is a simple consequence of the continuous mapping theorem.

Theorem 3.5. *Assume H_0 and Assumptions 3.1 and 3.2. Then*

$$\sqrt{|J_n|}I^{(n)} \Longrightarrow U$$

where $U = (U_t)_{t\in[0,T]}$ is a zero-mean Gaussian martingale with respect to $\tilde{\mathcal{G}}_t$ with variance function σ^2 given by (16)*.*

Hence we have established the weak asymptotic limit of our test process $\sqrt{|J_n|}\hat{I}^{(n)}$. However, the variance function σ^2 of the limiting Gaussian martingale is unknown and must be estimated from data, e.g. using the estimator

$$\begin{aligned}\hat{\sigma}_n^2(t) &= \frac{1}{|J_n|}\sum_{j\in J_n}\int_0^t (Z_{j,s} - \hat{\Pi}_{j,s}^{(n)})^2 \mathrm{d}N_{j,s} \\ &= \frac{1}{|J_n|}\sum_{j\in J_n}\sum_{\tau\le t:\Delta N_{j,\tau}=1} (Z_{j,\tau} - \hat{\Pi}_{j,\tau}^{(n)})^2.\end{aligned}$$

We can now construct test statistics for conditional local independence as finite dimensional functionals of $\hat{I}^{(n)}$ that quantifies the deviance of the parameter $t \mapsto \gamma_t$ from the zero-function. As a simple example consider the following.

Corollary 3.6. *Under H_0 and Assumptions 3.1 and 3.2 it holds that*

$$\sqrt{\frac{|J_n|}{\hat{\sigma}_n^2(T)}}\hat{I}_T^{(n)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \tag{17}$$

as $n \to \infty$ if $\hat{\sigma}_n^2(T) \xrightarrow{P} \sigma^2(T)$.

The intuition behind the test statistic is as follows. Since I is a zero-mean, local martingale under H_0, we expect the fluctuations of its estimator process $\hat{I}^{(n)}$ to be well-behaved. On the contrary, under the alternative the process I is not necessarily a zero-mean, local martingale, and could

have a drift. Therefore, we expect its estimator $\hat{I}^{(n)}$ to have a drift, which makes the fluctuations more extreme than what we expect under the null hypothesis. However, the test statistic in (17) does not fully make use of this idea, since it only considers the end point of the process, and not the fluctuations of the entire sample path of $\hat{I}^{(n)}$ over $[0,T]$. In order to fully leverage this idea, consider the test statistic

$$\hat{T}_n = \sup_{0\leq t\leq T} \left|\hat{I}_t^{(n)}\right|. \tag{18}$$

We then have the following result, which exploits the Dubins-Schwarz theorem, which states that a continuous, local martingale can be represented by a time-changed Brownian motion.

Theorem 3.7. *Let $\hat{T}_n$ be given by* (18)*. Under H_0 and Assumptions 3.1 and 3.2 we have that*

$$\sqrt{|J_n|}\hat{T}_n \xrightarrow{\mathcal{D}} S$$

where S has the distribution of the supremum of the absolute value of a Brownian motion on $[0,\sigma^2(T)]$.

In practice, the p-value for the test can be determined by bootstrapping by plugging in the estimate $\hat{\sigma}_n^2(T)$ for $\sigma^2(T)$, and then simulate a large number of sample paths from a Brownian motion on $[0,\hat{\sigma}_n^2(T)]$.

4. Cox Example continued

With the setup as in Section 1.1, we let $X = (X_t)_{0\leq t\leq T}$ be a stochastic process with continuous sample path and values in $[0,C_0]$. In terms of X we let

$$Y_t = \int_0^t X_s\beta(s,t)\mathrm{d}s + V_t$$
$$Z_t = \int_0^t X_s\rho(s,t)\mathrm{d}s + W_t$$

where β and ρ are two functions and $V = (V_t)_{0\leq t\leq T}$ and $W = (W_t)_{0\leq t\leq T}$ are two bounded stochastic processes with continuous sample paths and mean 0. The processes X, V and W are assumed independent, which implies (4). The functions β and ρ are defined on the triangle $\mathcal{T} = \{(s,t)\in[0,T]^2 \mid s\leq t\}$ and are assumed suitably smooth. By the boundedness of X and hence λ and Z it follows that Assumption 3.1 is fulfilled.

Within this framework,

$$\Pi_s = E(Z_s \mid \mathcal{F}_{s-}^X) = \int_0^t X_s\rho(s,t)\mathrm{d}s, \tag{19}$$

and on $(\tau > t)$

$$\begin{aligned} E(e^{Y_t} \mid \mathcal{F}_t^X) &= e^{\int_0^t X_s\beta(s,t)\mathrm{d}s}E(e^{V_t} \mid \tau > t) \\ &= e^{\tilde{\beta}_0(t)+\int_0^t X_s\beta(s,t)\mathrm{d}s}, \end{aligned}$$

where $\tilde{\beta}_0(t) = \log(E(e^{V_t} \mid \tau > t))$. Since

$$\lambda_t = 1_{[0,\tau)}(t)\lambda_t^0 e^{\beta_1 X_t}E(e^{Y_t} \mid \mathcal{F}_t^X)$$

it follows that on $(\tau > t)$,

$$\begin{aligned}\log(\lambda_t) &= \log(\lambda_t^0) + \tilde{\beta}_0(t) + \beta_1 X_t + \int_0^t X_s \beta(s,t) \mathrm{d}s \\ &= \beta_0(t) + \beta_1 X_t + \int_0^t X_s \beta(s,t) \mathrm{d}s, \end{aligned} \tag{20}$$

where the two baseline terms depending only on time have been merged into β_0. The functional forms of how Π and $\log(\lambda)$ depend on the history of the X-process are thus similar and known as the *historical functional linear model* in functional data analysis (Malfait & Ramsay 2003). We can therefore use standard methods and implementations for nonparametric estimation of ρ and β, and thus Π and λ, such as boosting (Brockhaus et al. 2017, 2020).

5. Discussion

Our proposed test statistic, its decomposition and the strategies used to bound the three remainder terms, e.g. data splitting, follow the general pattern used for deriving properties of double machine learning procedures. The functional limit of the leading term, $U^{(n)}$, is likewise not surprising, see e.g. Section V.4 in (Andersen et al. 1993) for similar nonparametric test statistics in the context of survival analysis. However, targeting the infinite dimensional parameter $t \mapsto \gamma_t$ requires several novel ideas.

First, the orthogonalization is based on a model of Z_t for each t given the history up to time t, and here it is important that we model the *predictable* projection, $E(Z_t \mid \mathcal{F}_{t-})$. This will ensure the necessary martingale properties used for deriving the asymptotic limit.

Second, we need to control the remainder terms uniformly over t. The third term, $R_3^{(n)}$ is simple to bound, and $R_2^{(n)}$ can also be bounded fairly easily by exploiting Doob's submartingal inequality. However, for the first term, $R_1^{(n)}$, we need to establish stochastic equicontinuity via an exponential tail bound and the use of the chaining lemma. It is this argument that requires most of the strong assumptions made, e.g. uniform bounds on λ and Z.

A major practical question is whether we can estimate λ and Π with sufficient rates, e.g. $n^{-\frac{1+\varepsilon}{4}}$. This is, of course, possible with parametric models, but of much greater interest if it can be achieved with nonparametric estimators. The example in Section 4 makes it plausible that this is, indeed, possible with somewhat strong assumptions about the structure of λ and Π. The rate results by Yao et al. (2005) for function-on-function regression suggest that good rates can be achieved if e.g. Z, X and ρ are sufficiently smooth and sampled on a sufficiently fine grid. Though the results by Yao et al. (2005) are *not* for the historical model, but a slightly different model, we conjecture that similar results can be obtained for the historical model in Section 4. Whether it is possible to achieve good rates with weak assumptions about the structure of the λ and Π processes remains an open question.

Appendix A. Proofs

This appendix contains proofs of the results of the paper.

A.1. **Proof of Proposition 2.2.** The process Z is preditable since it is caglad, and Π is predictable by construction. Thus $Z - \Pi$ is predictable, and the process $I = (I_t)$ being a stochastic integral of $Z - \Pi$ w.r.t. a local $\mathcal{G}_t$-martingale is likewise a local $\mathcal{G}_t$-martingale under the hypothesis. By definition, $I_0 = 0$, and if I is a martingale, $E(I_t) = E(I_0) = 0$. □

A.2. **Proof of Proposition 3.3.** For simplicity of notation we write

$$U_t^{(n)} = \sum_{j \in J_n} \int_0^t H_{j,s}^{(n)} \mathrm{d}M_{j,s} \quad \text{where} \quad H_{j,s}^{(n)} = \frac{Z_{j,s} - \Pi_{j,s}}{\sqrt{|J_n|}}.$$

We will use the functional martingale central limit theorem, Theorem B.1, described in Appendix B on the sequence $(U^{(n)})_{n \geq 1}$ to show the result. In the following we let $\tilde{\mathcal{G}}_t$ be the smallest right continuous and complete filtration having the filtrations $\{\mathcal{G}_{j,t} \mid j \in J_n, n \geq 1\}$ as subfiltrations.

First note that each compensator Λ_j is continuous and bounded, since its intensity λ_j is caglad and bounded by Assumption 3.1 (i). Therefore, M_j is a square integrable, zero-mean, local $\tilde{\mathcal{G}}_t$-martingale under the hypothesis of conditional local independence. Secondly, every $H_j^{(n)}$ is bounded and $\tilde{\mathcal{G}}_t$-predictable, since Z_j is bounded and caglad by Assumption 3.1 (ii). Therefore, under the hypothesis of conditional local independence, $(U^{(n)})_{n \geq 1}$ is also a sequence of square integrable, zero-mean, local $\tilde{\mathcal{G}}_t$-martingales. Now we see that

$$\begin{aligned}
\left\langle U^{(n)}, U^{(n)} \right\rangle (t) &= \sum_{j \in J_n} \int_0^t \left(H_{j,s}^{(n)} \right)^2 \mathrm{d}\Lambda_{j,s} \\
&= \sum_{j \in J_n} \int_0^t \left(H_{j,s}^{(n)} \right)^2 \mathrm{d}N_{j,s} - \sum_{j \in J_n} \int_0^t \left(H_{j,s}^{(n)} \right)^2 \mathrm{d}M_{j,s}.
\end{aligned}$$

By the law of large numbers we have that

$$\sum_{j \in J_n} \int_0^t \left(H_{j,s}^{(n)} \right)^2 \mathrm{d}M_{j,s} = \frac{1}{|J_n|} \sum_{j \in J_n} \int_0^t (Z_{j,s} - \Pi_{j,s})^2 \mathrm{d}M_{j,s} \xrightarrow{P} 0$$

as $n \to \infty$ for each fixed $t \in [0, T]$, since the integrals are i.i.d. zero-mean local martingales. Similarly, we have that

$$\begin{aligned}
\sum_{j \in J_n} \int_0^t \left(H_{j,s}^{(n)} \right)^2 \mathrm{d}N_{j,s} &= \frac{1}{|J_n|} \sum_{j \in J_n} \int_0^t (Z_{j,s} - \Pi_{j,s})^2 \mathrm{d}N_{j,s} \\
&\xrightarrow{P} E\left(\int_0^t (Z_s - \Pi_s)^2 \mathrm{d}N_s \right) = \sigma^2(t)
\end{aligned}$$

as $n \to \infty$ for each fixed $t \in [0, T]$. Here we note that $\sigma^2(t) < \infty$ for each $t \in [0, T]$ by Assumption 3.1 (iii). All in all we have

$$\left\langle U^{(n)}, U^{(n)} \right\rangle (t) \xrightarrow{P} \sigma^2(t) < \infty$$

as $n \to \infty$ for $t \in [0, T]$. Using the notation of Theorem B.1 we have that

$$\begin{aligned}
\left\langle U_\varepsilon^{(n)}, U_\varepsilon^{(n)} \right\rangle (t) &= \sum_{j \in J_n} \int_0^t \left(H_{j,s}^{(n)} \right)^2 1\left(|H_{j,s}^{(n)}| \geq \varepsilon \right) \mathrm{d}\Lambda_{j,s} \\
&= \int_0^t \frac{1}{|J_n|} \sum_{j \in J_n} (Z_{j,s} - \Pi_{j,s})^2 \, 1\left(\left| \frac{Z_{j,s} - \Pi_{j,s}}{\sqrt{|J_n|}} \right| \geq \varepsilon \right) \lambda_{j,s} \mathrm{d}s
\end{aligned}$$

for each $t \in [0, T]$ and $\varepsilon > 0$. We will show that this integral converges to zero in probability by showing that it converges to zero in expectation. Let $W_n(s) = E(F_n(s))$ where

$$F_n(s) = \frac{1}{|J_n|} \sum_{j \in J_n} (Z_{j,s} - \Pi_{j,s})^2 \, 1\left(\left| \frac{Z_{j,s} - \Pi_{j,s}}{\sqrt{|J_n|}} \right| \geq \varepsilon \right) \lambda_{j,s}$$

for $s \in [0,t]$ and $\varepsilon > 0$. Firstly, we note that

$$E\left(\int_0^t F_n(s)\mathrm{d}s\right) \leq E\left(\int_0^t \left(Z_{j,s} - \Pi_{j,s}\right)^2 \lambda_{j,s}\mathrm{d}s\right) = \sigma^2(t) < \infty$$

by Assumption 3.1 (iii), where we have used that the observations $j \in J_n$ are i.i.d. Tonelli's theorem therefore gives

$$E\left(\int_0^t F_n(s)\mathrm{d}s\right) = \int_0^t W_n(s)\mathrm{d}s.$$

Now secondly, we have by Assumption 3.1 (iv) that

$$W_n(s) \leq E\left(\left(Z_{j,s} - \Pi_{j,s}\right)^2 \lambda_{j,s}\right) < \infty,$$

and because $F_n(s) \to 0$ a.s. for $n \to \infty$ we have by dominated convergence that $W_n(s) \to$ for $n \to \infty$. All in all this gives that

$$E\left(\left\langle U_\varepsilon^{(n)}, U_\varepsilon^{(n)} \right\rangle(t)\right) = \int_0^t W_n(s)\mathrm{d}s \to 0 \tag{21}$$

as $n \to \infty$. By Theorem B.1 we conclude that

$$U^{(n)} \Longrightarrow U$$

where U is a zero-mean, Gaussian martingale with respect to $\tilde{\mathcal{G}}_t$ with variance function σ^2, which was what we wanted. □

A.3. **Proof of Proposition 3.4.** We will divide the proof into lemmas concerning each of the remainder terms $R_1^{(n)}, R_2^{(n)}$ and $R_3^{(n)}$. Our proof strategy is as follows. Instead of directly showing weak convergence to the zero-process, we will show uniform convergence in probability to the zero-process, since this implies the former. For a general discussion of the relation between weak convergence and uniform convergence in probability see Newey (1991). Throughout the following we let $\tilde{\mathcal{G}}_t$ be the smallest right continuous and complete filtration having the filtrations $\{\mathcal{G}_{j,t} \mid j \in J_n, n \in \mathbb{N}\}$ as subfiltrations. Analogously, we let $\tilde{\mathcal{G}}_t^c$ be the smallest right continuous and complete filtration having the filtrations $\{\mathcal{G}_{j,t} \mid j \in J_n^c, n \in \mathbb{N}\}$ as subfiltrations. We start by considering $R_3^{(n)}$, since this is the easiest case.

Lemma A.1. *Under Assumption 3.2 it holds that* $R_3^{(n)} \Longrightarrow 0$.

Proof. We will show the result by showing that

$$E\left(\sup_{0\leq t\leq T} |R_{3,t}^{(n)}|\right) \to 0$$

as $n \to \infty$. Using that the collection of random variables

$$\left(\sup_{0\leq t\leq T} \left|\Pi_{j,t} - \hat{\Pi}_{j,t}^{(n)}\right| \cdot \sup_{0\leq t\leq T} \left|\lambda_{j,t} - \hat{\lambda}_{j,t}^{(n)}\right|\right)_{j\in J_n}$$

is identically distributed for each fixed $n \geq 2$ we have that

$$\begin{aligned}
&E\left(\sup_{0\leq t\leq T}|R_{3,t}^{(n)}|\right)\\
&= E\left(\sup_{0\leq t\leq T}\left|\frac{1}{\sqrt{|J_n|}}\sum_{j\in J_n}\int_0^t\left(\Pi_{j,s}-\hat{\Pi}_{j,s}^{(n)}\right)\left(\lambda_{j,s}-\hat{\lambda}_{j,s}^{(n)}\right)\mathrm{d}s\right|\right)\\
&\leq E\left(\frac{T}{\sqrt{|J_n|}}\sum_{j\in J_n}\sup_{0\leq t\leq T}\left|\Pi_{j,t}-\hat{\Pi}_{j,t}^{(n)}\right|\cdot\sup_{0\leq t\leq T}\left|\lambda_{j,t}-\hat{\lambda}_{j,t}^{(n)}\right|\right)\\
&= T\cdot\sqrt{|J_n|}E\left(\sup_{0\leq t\leq T}\left|\Pi_{j,t}-\hat{\Pi}_{j,t}^{(n)}\right|\cdot\sup_{0\leq t\leq T}\left|\lambda_{j,t}-\hat{\lambda}_{j,t}^{(n)}\right|\right)\\
&\leq T\cdot\sqrt{|J_n|}\sqrt{E\left(\left(\sup_{0\leq t\leq T}\left|\Pi_{j,t}-\hat{\Pi}_{j,t}^{(n)}\right|\right)^2\right)}\sqrt{E\left(\left(\sup_{0\leq t\leq T}\left|\lambda_{j,t}-\hat{\lambda}_{j,t}^{(n)}\right|\right)^2\right)}\\
&= T\cdot\sqrt{|J_n|}g(n)h(n)\to 0
\end{aligned}$$

as $n\to\infty$ by Assumption 3.2. Therefore, we also have uniform convergence in probability to the zero process. □

Next we proceed to the remainder process $R_2^{(n)}$.

Lemma A.2. *Under H_0 and Assumptions 3.1 and 3.2 it holds that $R_2^{(n)} \Longrightarrow 0$.*

Proof. Note that

$$R_{2,t}^{(n)} = \frac{1}{\sqrt{|J_n|}}\sum_{j\in J_n}\int_0^t\left(\Pi_{j,s}-\hat{\Pi}_{j,s}^{(n)}\right)\mathrm{d}M_{j,s}$$

is a mean-zero, local $\tilde{\mathcal{G}}_t$-martingale conditionally on $\tilde{\mathcal{G}}_T^c$ under the hypothesis of conditional local independence since then each M_j is a mean-zero, local $\tilde{\mathcal{G}}_t$-martingale and each $\Pi_j - \hat{\Pi}_j^{(n)}$ is $\tilde{\mathcal{G}}_t$-predictable. Therefore, the squared process $(R_2^{(n)})^2$ is a local $\tilde{\mathcal{G}}_t$-submartingale. By Doob's submartingale inequality we have that

$$\begin{aligned}
P\left(\sup_{0\leq t\leq T}|R_{2,t}^{(n)}|\geq\varepsilon\right) &= P\left(\sup_{0\leq t\leq T}\left(R_{2,t}^{(n)}\right)^2\geq\varepsilon^2\right)\\
&= E\left(P\left(\sup_{0\leq t\leq T}\left(R_{2,t}^{(n)}\right)^2\geq\varepsilon^2\mid\tilde{\mathcal{G}}_T^c\right)\right)\\
&\leq \frac{E\left(V\left(R_{2,T}^{(n)}\mid\tilde{\mathcal{G}}_T^c\right)\right)}{\varepsilon^2}
\end{aligned}$$

for $\varepsilon>0$. The collection of random variables

$$\left(\int_0^T\left(\Pi_{j,s}-\hat{\Pi}_{j,s}^{(n)}\right)\mathrm{d}M_{j,s}\right)_{j\in J_n}$$

are i.i.d. conditionally on $\tilde{\mathcal{G}}_T^c$. Therefore,

$$\begin{aligned} V\left(R_{2,T}^{(n)} \mid \tilde{\mathcal{G}}_T^c\right) &= \frac{1}{|J_n|}\sum_{j\in J_n} V\left(\int_0^T \left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right) \mathrm{d}M_{j,s} \mid \tilde{\mathcal{G}}_T^c\right) \\ &= E\left(\int_0^T \left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right)^2 \mathrm{d}\langle M_j\rangle_s \mid \tilde{\mathcal{G}}_T^c\right) \\ &= E\left(\int_0^T \left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right)^2 \lambda_{j,s}\mathrm{d}s \mid \tilde{\mathcal{G}}_T^c\right) \\ &\leq T\cdot E\left(\sup_{0\leq t\leq T}\left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right)^2 \lambda_{j,t} \mid \tilde{\mathcal{G}}_T^c\right) \\ &\leq T\cdot C\cdot E\left(\sup_{0\leq t\leq T}\left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right)^2 \mid \tilde{\mathcal{G}}_T^c\right) \end{aligned}$$

where we have used that λ_j is bounded by Assumption 3.1 (i). Thus

$$E\left(V\left(R_{2,T}^{(n)} \mid \tilde{\mathcal{G}}_T^c\right)\right) \leq T\cdot C\cdot E\left(\sup_{0\leq t\leq T}\left(\Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right)^2\right) = T\cdot C\cdot g(n)^2,$$

and we conclude that

$$P\left(\sup_{0\leq t\leq T}|\sqrt{|J_n|}R_{2,t}^{(n)}| \geq \varepsilon\right) \leq \frac{T\cdot C}{\varepsilon^2} g(n)^2 \to 0$$

as $n\to\infty$ by Assumption 3.2. □

Before proving that $R_1^{(n)}$ converges weakly to the zero-process, we will need an auxiliary lemma. For $s,t\in[0,T]$ with $s<t$ define

$$X^{s,t} = \frac{1}{t-s}\int_s^t (Z_u - \Pi_u)(\lambda_u - \hat{\lambda}_u^{(n)})\mathrm{d}u.$$

Recall that a random variable S is called sub-Gaussian with variance factor ν if

$$\log E(e^{\lambda S}) \leq \frac{\lambda^2\nu}{2}$$

for all $\lambda\in\mathbb{R}$.

Lemma A.3. *Let Assumption 3.1 hold true. There exists $\nu>0$ such that for all $s<t$ it holds that $X^{s,t}$ is sub-Gaussian conditionally on $\tilde{\mathcal{G}}_T^c$ with variance factor ν, that is,*

$$\log E(e^{\lambda X^{s,t}} \mid \tilde{\mathcal{G}}_T^c) \leq \frac{\lambda^2\nu}{2}$$

for all $s<t$ and $\lambda\in\mathbb{R}$.

Proof. This follows from Boucheron et al. (2013) Lemma 2.2, which states that bounded random variables are sub-Gaussian. Indeed, we have that

$$\begin{aligned} |X^{s,t}| &\leq \frac{1}{t-s}\int_s^t |(Z_u-\Pi_u)||(\lambda_u - \hat{\lambda}_u^{(n)})|\mathrm{d}u \\ &\leq \sup_{0\leq u\leq T}|(Z_u-\Pi_u)| \sup_{0\leq u\leq T}|(\lambda_u - \hat{\lambda}_u^{(n)})| \leq 4CC' \end{aligned}$$

by Assumption 3.1. Therefore, there exists $\nu>0$ such that $X^{s,t}$ is sub-Gaussian with variance factor ν for all $s<t$. As the bound on $|X^{s,t}|$ does not depend on s,t or $\tilde{\mathcal{G}}_T^c$, the variance factor ν can be chosen uniformly in s,t, and sub-Gaussianity holds conditionally on $\tilde{\mathcal{G}}_T^c$. □

Then we have the following regarding $R_1^{(n)}$.

Lemma A.4. *Under H_0 and Assumptions 3.1 and 3.2 it holds that $R_1^{(n)} \Longrightarrow 0$.*

Proof. The proof consists of two parts. First we show that for each $t \in [0,T]$ it holds that

$$R_{1,t}^{(n)} \xrightarrow{P} 0$$

for $n \to \infty$. Then we show *stochastic equicontinuity* of the process $R_1^{(n)}$, and by Theorem 2.1 in Newey (1991) it follows that

$$\sup_{t\in[0,T]} |R_{1,t}^{(n)}| \xrightarrow{P} 0.$$

The collection of random variables

$$\left((Z_{j,s} - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right)\right)_{j\in J_n}$$

are i.i.d. conditionally on $\tilde{\mathcal{G}}_T^c$. Therefore,

$$\begin{aligned}
E(R_{1,t} \mid \tilde{\mathcal{G}}_T^c) &= E\left(\frac{1}{\sqrt{|J_n|}}\sum_{j\in J_n}\int_0^t (Z_{j,s} - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right)\mathrm{d}s \mid \tilde{\mathcal{G}}_T^c\right) \\
&= \sqrt{|J_n|}\int_0^t E\left((Z_{j,s} - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right) \mid \tilde{\mathcal{G}}_T^c\right)\mathrm{d}s \\
&= \sqrt{|J_n|}\int_0^t E\left(E\left((Z_{j,s} - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right) \mid \mathcal{F}_{j,s-} \vee \tilde{\mathcal{G}}_T^c\right) \mid \tilde{\mathcal{G}}_T^c\right)\mathrm{d}s \\
&= \sqrt{|J_n|}\int_0^t E\left((E\left(Z_{j,s} \mid \mathcal{F}_{j,s-}\right) - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right) \mid \tilde{\mathcal{G}}_T^c\right) = 0,
\end{aligned}$$

where we have used that $\lambda_{j,t} - \hat{\lambda}_{j,t}^{(n)}$ is $\mathcal{F}_{j,t}$-predictable conditionally on $\tilde{\mathcal{G}}_T^c$, that $\Pi_{j,t}$ is $\mathcal{F}_{j,t}$-predictable, and that $E\left(Z_{j,s} \mid \mathcal{F}_{j,s-}\right) - \Pi_{j,s} = 0$ per definition. Whence $E(R_{1,t}^{(n)}) = 0$, and $V(R_{1,t}^{(n)}) = E(V(R_{1,t}^{(n)}) \mid \tilde{\mathcal{G}}_T^c)$, so

$$\begin{aligned}
V(R_{1,t}^{(n)}) &= E\left(\frac{1}{|J_n|}\sum_{j\in J_n} V\left(\int_0^t (Z_{j,s} - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right)\mathrm{d}s \mid \tilde{\mathcal{G}}_T^c\right)\right) \\
&= E\left(E\left(\left(\int_0^t (Z_{j,s} - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right)\mathrm{d}s\right)^2 \mid \tilde{\mathcal{G}}_T^c\right)\right) \\
&= E\left(\left(\int_0^t (Z_{j,s} - \Pi_{j,s})\left(\lambda_{j,s} - \hat{\lambda}_{j,s}^{(n)}\right)\mathrm{d}s\right)^2\right) \\
&\leq E\left(\sup_{0\leq t\leq T}(Z_{j,t} - \Pi_{j,t})^2 \sup_{0\leq t\leq T}(\lambda_{j,t} - \hat{\lambda}_{j,t}^{(n)})^2\right) \\
&\leq \left|\left|\left|(Z_j - \Pi_j)^2\right|\right|\right|_{\infty,T}\left|\left|\left|\left(\lambda_j - \hat{\lambda}_j^{(n)}\right)^2\right|\right|\right|_{\infty,T} \\
&\leq K\left|\left|\left|\left(\lambda_j - \hat{\lambda}_j^{(n)}\right)^2\right|\right|\right|_{\infty,T},
\end{aligned}$$

where we have used the Cauchy-Schwartz inequality and Assumption 3.1 (ii). Hence by Chebychev's inequality,

$$P(|R_{1,t}^{(n)}| > \varepsilon) \leq \frac{V(R_{1,t}^{(n)})}{\varepsilon^2} \leq \frac{K}{\varepsilon^2} \left\| \left(\lambda_j - \hat{\lambda}_j^{(n)} \right)^2 \right\|_{\infty,T} \to 0$$

as $n \to \infty$ for all $\varepsilon > 0$ by Assumption 3.2. This completes the first part of the proof. For the second part, we use a chaining argument based on the exponential inequality in Lemma A.3. Again we consider

$$X_j^{s,t} = \frac{1}{t-s} \int_s^t (Z_{j,u} - \Pi_{j,u}) \left(\lambda_{j,u} - \hat{\lambda}_{j,u}^{(n)} \right) \mathrm{d}u$$

such that

$$S = \frac{1}{\sqrt{|J_n|}} \sum_{j \in J_n} X_j^{s,t} = \frac{1}{t-s} (R_{1,t}^{(n)} - R_{1,s}^{(n)}).$$

Using that $(X_j^{s,t})_{j \in J_n}$ are i.i.d. conditionally on $\tilde{\mathcal{G}}_T^c$ we have by Lemma A.3 that

$$\begin{aligned}
\log E\left(e^{\lambda S}\right) &= \log E\left(E\left(e^{\lambda S} \mid \tilde{\mathcal{G}}_T^c\right)\right) \\
&= \log E\left(\prod_{j \in J_n} E\left(e^{\frac{\lambda}{\sqrt{|J_n|}} X_j^{s,t}} \mid \tilde{\mathcal{G}}_T^c\right)\right) \\
&\leq \log E\left(e^{\frac{\lambda^2 \nu}{2}}\right) \\
&= \frac{\lambda^2 \nu}{2}.
\end{aligned}$$

So S is also sub-Gaussian with variance factor ν. This implies that

$$P(|S| > \eta) \leq 2e^{-\frac{\eta^2 \nu}{2}}$$

for all $\eta > 0$. Rephrased in terms of $R_1^{(n)}$ this bound reads

$$P\left(|R_{1,t}^{(n)} - R_{1,s}^{(n)}| > \eta(t-s)\right) \leq 2e^{-\frac{\eta^2 \nu}{2}}$$

for all $\eta > 0$ and $s < t$. It now follows from the chaining lemma, Pollard (1984) Lemma VII.9, that $R_1^{(n)}$ is stochastic equicontinuous (the covering numbers and integral are bounded by standard arguments for $[0, T]$). This completes the second part of the proof. □

Proposition 3.4 now follows from combining the three lemmas above.

A.4. **Proof of Corollary 3.6.** Under H_0 and Assumptions 3.1 and 3.2 we know by Theorem 3.5 that

$$\sqrt{|J_n|} \hat{I}^{(n)} \Longrightarrow U$$

as $n \to \infty$, and therefore we have convergence of all finite dimensional distributions. In particular, we have that

$$\sqrt{|J_n|} \hat{I}_T^{(n)} \xrightarrow{\mathcal{D}} U_T \sim \mathcal{N}(0, \sigma^2(T))$$

as $n \to \infty$. As $\hat{\sigma}_n^2(T) \xrightarrow{P} \sigma^2(T)$, Slutsky's lemma now implies that

$$\sqrt{\frac{|J_n|}{\hat{\sigma}_n^2(T)}} \hat{I}_T^{(n)} = \sqrt{\frac{\sigma^2(T)}{\hat{\sigma}_n^2(T)}} \sqrt{\frac{|J_n|}{\sigma^2(T)}} \hat{I}_T^{(n)} \xrightarrow{\mathcal{D}} 1 \cdot \frac{U_T}{\sqrt{\sigma^2(T)}} \sim \mathcal{N}(0,1)$$

as $n \to \infty$. □

A.5. **Proof of Theorem 3.7.** According to Theorem 3.5 we have that

$$\sqrt{|J_n|}\hat{I}^{(n)} \Longrightarrow U$$

where U is a Gaussian process with variance function σ^2. By the Dubins-Schwarz theorem, we can represent

$$U_t = B_{\sigma^2(t)} \tag{22}$$

where B_u is a Brownian motion on $[0, \sigma^2(T)]$. See Revuz & Yor (2013) Chapter V Theorems 1.6 and 1.7 and the surrounding discussion. Therefore, we have that

$$\sqrt{|J_n|}\hat{T}_n \xrightarrow{\mathcal{D}} \sup_{0\leq t\leq T} |U_t| = \sup_{0\leq t\leq T} |B_{\sigma^2(t)}| = \sup_{0\leq u\leq \sigma^2(T)} |B_u|$$

where we have used that $t \mapsto \sigma^2(t)$ is increasing. □

Appendix B. Martingale functional central limit theorem

In this appendix we provide a functional central limit theorem for martingales following Chapter 5 of Fleming & Harrington (2011).

Let $N_j^{(n)}$ be a counting process on $(\Omega, \mathcal{F}, P)$ adapted to the filtration $(\mathcal{F}_t)$ with continuous compensator $\Lambda_j^{(n)}$ and associated locally square integrable martingale $M_j^{(n)} = N_j^{(n)} - \Lambda_j^{(n)}$ for $j = 1, \ldots, n$. Also let $H_j^{(n)}$ be a locally bounded $\mathcal{F}_t$-predictable process for each $j = 1, \ldots, n$. Define the process $U^{(n)}$ by

$$U_t^{(n)} = \sum_{j=1}^{n} \int_0^t H_{j,s}^{(n)} \mathrm{d}M_{j,s}^{(n)}$$

and for $\varepsilon > 0$ define the process $U_\varepsilon^{(n)}$,

$$U_{\varepsilon,t}^{(n)} = \sum_{j=1}^{n} \int_0^t H_{j,s}^{(n)} 1\left(|H_{j,s}^{(n)}| \geq \varepsilon\right) \mathrm{d}M_{j,s}^{(n)},$$

which contains all the jumps of $U^{(n)}$ of size at least ε. It can then be noted that $U^{(n)}$ and $U_\varepsilon^{(n)}$ are themselves locally square integrable martingales and that

$$\left\langle U^{(n)}, U^{(n)} \right\rangle (t) = \sum_{j=1}^{n} \int_0^t \left(H_{j,s}^{(n)}\right)^2 \mathrm{d}\Lambda_{j,s}^{(n)}$$

and similarly that

$$\left\langle U_\varepsilon^{(n)}, U_\varepsilon^{(n)} \right\rangle (t) = \sum_{j=1}^{n} \int_0^t \left(H_{j,s}^{(n)}\right)^2 1\left(|H_{j,s}^{(n)}| \geq \varepsilon\right) \mathrm{d}\Lambda_{j,s}^{(n)}$$

according to Theorem 2.4.1 and Theorem 2.4.3 of Fleming & Harrington (2011). We then have the following functional martingale central limit theorem (Fleming & Harrington 2011, Theorem 5.3.3), where we use $\Longrightarrow$ to denote convergence in $D[0, \infty)$, the space of cadlag functions $[0, \infty) \to \mathbb{R}$, endowed with the Skorokhod topology.

Theorem B.1. *Let $t \mapsto \sigma^2(t)$ be a nonnegative measurable function. Assume that*

$$\left\langle U^{(n)}, U^{(n)} \right\rangle (t) \xrightarrow{P} \sigma^2(t) \quad \text{and} \quad \left\langle U_\varepsilon^{(n)}, U_\varepsilon^{(n)} \right\rangle (t) \xrightarrow{P} 0$$

as $n \to \infty$ for each $t > 0$. Then it holds that

$$U^{(n)} \Longrightarrow U$$

as $n \to \infty$ where U is a Gaussian martingale with variance function σ^2.

References

Aalen, O. O. (1987), 'Dynamic modelling and causality', *Scandinavian Actuarial Journal* pp. 177–190.

Aalen, O. O., Røysland, K., Gran, J. M. & Ledergerber, B. (2012), 'Causality, mediation and time: a dynamic viewpoint', *Journal of the Royal Statistical Society, Series A* **175**(4), 831–861.

Achab, M., Bacry, E., Gaïffas, S., Mastromatteo, I. & Muzy, J.-F. (2017), Uncovering causality from multivariate Hawkes integrated cumulants, *in* D. Precup & Y. W. Teh, eds, 'Proceedings of the 34th International Conference on Machine Learning', Vol. 70 of *Proceedings of Machine Learning Research*, PMLR, International Convention Centre, Sydney, Australia, pp. 1–10.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993), *Statistical models based on counting processes*, Springer Series in Statistics, Springer-Verlag, New York.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press.

Brockhaus, S., Melcher, M., Leisch, F. & Greven, S. (2017), 'Boosting flexible functional regression models with a high number of functional historical effects', *Statistics and Computing* **27**(4), 913–926.

Brockhaus, S., Rügamer, D. & Greven, S. (2020), 'Boosting functional regression models with fdboost', *Journal of Statistical Software, Articles* **94**(10), 1–50.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**(1), C1–C68.

Commenges, D. & Gégout-Petit, A. (2009), 'A general dynamical statistical model with causal interpretation', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **71**(3), 719–736.

Didelez, V. (2006), 'Graphical models for composable finite Markov processes', *Scandinavian Journal of Statistics* **34**(1), 169–185.

Didelez, V. (2008), 'Graphical models for marked point processes based on local independence', *Journal of the Royal Statistical Society, Series B* **70**(1), 245–264.

Didelez, V. (2015), Causal reasoning for events in continuous time: A decision-theoretic approach, *in* 'Proceedings of the UAI 2015 Workshop on Advances in Causal Inference'.

Fleming, T. R. & Harrington, D. P. (2011), *Counting processes and survival analysis*, Vol. 169, John Wiley & Sons.

Granger, C. W. J. (1969), 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica* **37**(3), 424–438.

Malfait, N. & Ramsay, J. O. (2003), 'The historical functional linear model', *The Canadian Journal of Statistics* **31**(2), 115–128.

Martinussen, T. & Vansteelandt, S. (2013), 'On collapsibility and confounding bias in Cox and Aalen regression models', *Lifetime Data Anal.* **19**(3), 279–296.
URL: *https://doi.org/10.1007/s10985-013-9242-z*

Mogensen, S. W. & Hansen, N. R. (2020), 'Markov equivalence of marginalized local independence graphs', *Ann. Statist.* **48**(1), 539–559.

Mogensen, S. W., Malinsky, D. & Hansen, N. R. (2018), Causal learning for partially observed stochastic dynamical systems, *in* 'Proceedings of the 34th conference on Uncertainty in Artificial Intelligence', pp. 350–360.

Newey, W. K. (1991), 'Uniform convergence in probability and stochastic equicontinuity', *Econometrica* **59**(4), 1161–1167.

Pollard, D. (1984), *Convergence of stochastic processes*, Springer Series in Statistics, Springer-Verlag, New York.

Revuz, D. & Yor, M. (2013), *Continuous martingales and Brownian motion*, Vol. 293, Springer Science & Business Media.

Rogers, L. C. G. & Williams, D. (2000), *Diffusions, Markov processes, and martingales*, Vol. 2 of *Cambridge Mathematical Library*, Cambridge University Press, Cambridge. Itô calculus, Reprint of the second (1994) edition.

Schweder, T. (1970), 'Composable Markov processes', *Journal of Applied Probability* **7**(2), 400–410.

Xiao, S., Yan, J., Farajtabar, M., Song, L., Yang, X. & Zha, H. (2019), 'Learning time series associated event sequences with recurrent point process networks', *IEEE Transactions on Neural Networks and Learning Systems* **30**(10), 3124–3136.

Xu, H., Farajtabar, M. & Zha, H. (2016), Learning Granger causality for Hawkes processes, *in* M. F. Balcan & K. Q. Weinberger, eds, 'Proceedings of The 33rd International Conference on Machine Learning', Vol. 48, pp. 1717–1726.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005), 'Functional linear regression analysis for longitudinal data', *The Annals of Statistics* **33**(6), 2873 – 2903.

5.1 Regularity and rate assumptions

Let us discuss the regularity and rate assumptions that are required for the results of the manuscript in its current form. The main regularity assumptions are Assumption 3.1 (i) and (ii). Note that both Assumption 3.1 (iii) and (iv) are redundant, since they follow from the boundedness of Z and λ in (i) and (ii). We have included them to make it explicit what is required for the results to hold, since we need these to hold even if we relax (i) and (ii) at a future point.

The assumption that the intensity λ is caglad is required in two aspects. Firstly, we need λ to be $\mathcal{F}_t$-predictable, and since an $\mathcal{F}$-adapted and left continuous process is $\mathcal{F}$-predictable, it is natural to assume λ to be caglad. Secondly, a requirement for the functional martingale central limit theorem in Theorem B.1 is that the compensators are continuous, and this is ensured by the intensity λ being caglad, which is therefore a natural smoothness condition. The fact that Z is assumed caglad is also natural, since the stochastic integral construction of I is only a $\mathcal{G}$-martingale under conditional local independence if the integrand is $\mathcal{G}$-predictable, which is ensured by Z being caglad since it is $\mathcal{G}$-adapted by construction. For these reasons we consider the caglad assumptions on λ and Z to be natural and unproblematic.

The more restrictive assumptions are the almost sure uniform boundedness of λ and Z. The importance of Z being bounded is partly due to the functional martingale central limit theorem in Theorem B.1, where it is required that the integrand is locally bounded. This suggests that we might be able to relax the assumption to Z being locally bounded, however, we do not consider the boundedness assumption to be problematic, since we are typically free to choose Z ourselves in practice. Let us explain.

An intended use case of the hypothesis test is when $(N^d)_{d=1,\dots,p}$ is assumed to be a multivariate counting process, and we wish to test whether N^a is conditionally locally independent of N^b given N^C for some $a \in \{1,\dots,p\}$ and $C \subset \{1,\dots,p\}$ with $b \notin C$. This can be cast in the setup of our manuscript as follows. We let $\mathcal{F}$ be the filtration generated by $N^{\{a\}\cup C}$ such that N^a is $\mathcal{F}$-adapted, and then we choose Z to be a bounded, caglad process that depends on N^b. For example, we could choose $Z_t = \varphi(N^b_{t-})$ where φ is continuous and bounded, or we could choose

$$Z_t = \varphi\left(\int_0^{t-} k(t-s)\mathrm{d}N^b_s\right) \tag{5.1}$$

where φ is bounded and continuous, and k is some kernel function. Thus, the boundedness of Z is unproblematic, since we are free to choose it to be bounded as long as it depends on N^b such that the filtration $\mathcal{G}$ is strictly larger than $\mathcal{F}$.

The importance of the boundedness of λ is (together with boundedness of Z) to obtain the sub-Gaussianity in Lemma A.3, which is crucial for proving Lemma A.4, which states that the remainder term $R_1^{(n)}$ converges weakly to the zero process. Here we use Pollard's chaining lemma to establish stochastic equicontinuity, which is a rather strong result which requires an exponential tail bound. It remains an open question whether the boundedness of λ can be relaxed to, e.g., local boundedness, and whether stochastic equicontinuity of $R_1^{(n)}$ can be shown directly without using the chaining lemma.

Besides the regularity conditions of Assumption 3.1, we also assume consistency of nonparametric estimators of λ and Π according to the norm

$$|||X|||^2_{\infty,T} = E\left(\left(\sup_{0\le t\le T}|X_t|\right)^2\right).$$

in Assumption 3.2. The strongest assumption is that

$$k(n) = \left|\left|\left|(\lambda - \hat{\lambda}^{(n)})^2\right|\right|\right|_{\infty,T} \to 0$$

as $n \to \infty$ which is used in the proof of Lemma A.4. We hypothesize that this can be relaxed to the assumption that

$$k_\alpha(n) = \left|\!\left|\!\left|(\lambda - \hat{\lambda}^{(n)})^\alpha\right|\!\right|\!\right|_{\infty,T} \to 0$$

as $n \to \infty$ for some $0 < \alpha \leq 2$ by the use of Hölder's inequality. However, the details are the topic of further research. The requirement that $\sqrt{|J_n|}g(n)h(n) \to 0$ as $n \to \infty$ is standard in the context of double machine learning, and we only need a $n^{-\frac{1+\varepsilon}{4}}$-consistency rate for some $\varepsilon > 0$ for both $\hat{\lambda}^{(n)}$ and $\hat{\Pi}^{(n)}$ for this to hold. If this rate condition is not satisfied, a possible solution is to dedicate a larger portion of data to estimation of $\hat{\lambda}^{(n)}$ and $\hat{\Pi}^{(n)}$ than the computation of $\hat{I}^{(n)}$ by choosing $J_n \subset \{1, \ldots, n\}$ such that $|J_n|$ grows at a slower than linear rate, e.g., $|J_n| = \sqrt{n}$. Of course this comes at the expense of slower weak convergence of $\sqrt{|J_n|}\hat{I}^{(n)}$ to the Gaussian martingale U.

Note that we do not postulate that these consistency requirements are always satisfied for non-parametric models of λ and Π. The purpose of the manuscript is to provide a nonparametric test for conditional local independence in the case where the consistencies are believed to be true, and this must be justified or assumed for a specific application.

5.2 Directions of further research

Besides exploring the opportunity to relax the regularity conditions as stated above, there are several directions of further research regarding both theory and implementation of the proposed conditional local independence test. Firstly, in order to carry out the test using the finite dimensional functional $\hat{T}_n$ used in Theorem 3.7, we need to be able to estimate the variance function σ^2 of the limiting Gaussian martingale. In the manuscript we propose an estimator $\hat{\sigma}_n^2$, but do not show consistency. We believe that it is possible to show consistency by expanding $\hat{\sigma}_n^2$ as

$$\hat{\sigma}_n^2(t) = \frac{1}{|J_n|} \sum_{j \in J_n} \int_0^t \left(Z_{j,s} - \Pi_{j,s} + \Pi_{j,s} - \hat{\Pi}_{j,s}^{(n)}\right)^2 \mathrm{d}N_{j,s}$$

and separate the integral into a term not involving estimation uncertainty and a remainder term. From this we believe that it will be possible to use Assumption 3.2 to show consistency.

Secondly, we have not carried out an analysis of power of the tests, i.e., it is an open question which alternatives of conditional local dependence the test has power against. One possible starting point of such an analysis, would be to consider Z-processes given by (5.1), and then study the power of the test under different parametrized kernel functions by simulation.

Thirdly, we consider our testing procedure to be an example of double machine learning, where both the nuisances Λ and Π and the target parameter $t \mapsto \gamma_t$ are infinite dimensional. We believe that the subtraction of the predictable projection Π in the integrand of I can be seen as a Neyman-orthogonalization. However, the Neyman-orthogonal score *functions* described in Chernozhukov et al. [2018] only consider finite dimensional target parameters. It would be of independent theoretical interest to generalize this notion to infinite dimensional target parameters and develop a theory of Neyman-orthogonal score *processes*, which, to the best of our knowledge, is unexplored.

Lastly, in order to carry out the test in practice we need nonparametric estimators of λ and Π. In Chapter 6 we present the current status of a nonparametric estimator of intensity functions by using recurrent neural networks.

Chapter 6

Intensity Estimation using Neural Networks

6.1 Introduction

This chapter introduces a Python library for modelling intensities of marked point processes using recurrent neural networks. More specifically, we present an efficient implementation of the marked point process negative log-likelihood function as a loss function for training neural network models of intensities in the Keras [Chollet et al., 2015] high level API to the TensorFlow [Abadi et al., 2015] library for building and training neural network models.

Modeling intensities using neural networks has been considered by e.g. Xiao et al. [2019] and Zhang et al. [2020]. However, the focus of previous work has been on a fixed architecture of the neural network, and there are no publicly available implementations. The purpose of this work is to provide a toolbox for experimenting with neural networks for intensity estimation leveraging the modularity of Keras for specifying the network architecture and the automatic differentiation capabilities of TensorFlow for training the models. The implementation and code for reproducing the simulations of this chapter is available at `https://github.com/lassepetersen/IntensityRNN`.

The implementation presented here is ongoing work, and its motivation is for nonparametric intensity estimation in connection with the test for conditional local independence presented in Chapter 5. There we were given a counting process N adapted to a filtration $\mathcal{F}$, and for our test statistic process we needed to be able to estimate the $\mathcal{F}$-intensity of N. Here we consider the situation where the filtration $\mathcal{F}$ is generated by a multivariate counting process with N being one of its coordinates.

The organization of the chapter is as follows. In Section 6.2 we describe the setup more precisely in terms of marked point processes. In Section 6.3 we describe the proposed recurrent neural network modeling framework, and the loss function we will use to train the models. In Section 6.4 we describe in details how to use the module **intensitymodel**, which contains a model class for fitting intensities with neural network models. In Section 6.5 we show the performance of the model on renewal processes. Lastly, in Section 6.6 we discuss the future development of the module.

6.2 Setup

Let $(\tau_j, z_j)_{j\geq 1}$ be a marked point process with finite mark space $\mathcal{M}$ such that τ_j denotes the j'th event time and $z_j \in \mathcal{M}$ denotes the event type or mark. We assume that $0 < \tau_1 < \cdots < \tau_k < T$ where T is an independent censoring time. For each mark $d \in \mathcal{M}$ we associate a counting process

$$N_t^d = \sum_{j=1}^{\infty} 1(\tau_j \leq t) 1(z_j = d).$$

For a subset of marks $D \subset \mathcal{M}$ we let $\mathcal{F}^D = (\mathcal{F}_t^D)_{t\geq 0}$ be the right continuous and completed filtration generated by the counting processes $(N^d)_{d\in D}$. Given a fixed mark $d \in \mathcal{M}$ and a subset $D \subset \mathcal{M}$ with $d \in D$, we assume that N^d has an $\mathcal{F}^D$-intensity $\lambda^{d,D}$, that is, $\lambda^{d,D}$ is an $\mathcal{F}^D$-predictable, non-negative stochastic process such that, with $\Lambda_t = \int_0^t \lambda_s \mathrm{d}s$ being the $\mathcal{F}^D$-compensator of N^d, then

$$M_t^{d,D} := N_t^d - \Lambda_t^{d,D}$$

is an $\mathcal{F}^D$-martingale. The $\mathcal{F}^D$-intensity $\lambda^{d,D}$ of N^d is interpreted as the instantaneous rate of an event with mark $d \in D$ occurring given the D-history, i.e.,

$$\lambda_t^{d,D} = \lim_{\delta \to 0^+} \frac{P(N_{t+\delta}^d - N_t^d = 1 \mid \mathcal{F}_{t-}^D)}{\delta}. \tag{6.1}$$

In what follows we will assume that $d \in D \subset \mathcal{M}$ are fixed, and will omit the superscripts from the notation when it eases notation without causing confusion.

The goal is to estimate the intensity λ from data. Let $(\tau_{i,j}, z_{i,j})_{j\geq 1, i=1,\ldots,n}$ denote i.i.d. copies of the marked point process, and let $(T_i)_{i=1,\ldots,n}$ be the corresponding i.i.d. censoring times such that $0 < \tau_{i,1} < \cdots < \tau_{i,k_i} < T_i$ for each $i = 1, \ldots, n$. The negative log-likelihood of λ given the observations is then given by

$$\ell(\lambda) = \sum_{i=1}^{n} \left(\int_0^{T_i} \lambda_s \mathrm{d}s - \sum_{j=1}^{k_i} I_{i,j} \cdot \log \lambda_{\tau_{i,j}} \right) \tag{6.2}$$

where $I_{i,j} = 1(z_{i,j} = d)$. A model of λ can then be fitted to data using maximum likelihood.

6.3 Recurrent neural network model

Note that λ needs to be $\mathcal{F}$-predictable, that is, for each $t \geq 0$, λ_t should be $\mathcal{F}_{t-}$-measurable. The predictable σ-algebra $\mathcal{F}_{t-}$ is generated by the events strictly prior to t,

$$\mathcal{F}_{t-} = \sigma((\tau_j, z_j) \mid \tau_j < t, z_j \in D),$$

thus λ_t is a function of $(\tau_1, z_1), \ldots, (\tau_{m_t}, z_{m_t})$ where $m_t = \max\{j \in \mathbb{N} \mid \tau_j < t\}$ and $z_j \in D$. Since the input sequence length m_t dynamically varies with t, and the inputs are temporally ordered, it is natural to consider modelling λ using recurrent neural networks. To this end, we consider modelling λ as

$$\lambda_t = 1(t < T) \varphi_\theta(t, (\tau_j, z_j)_{j=1}^{m_t}) \tag{6.3}$$

where $\varphi_\theta(t, (\tau_j, z_j)_{j=1}^{m_t})$ is the output of a recurrent neural network parametrized by a parameter $\theta \in \Theta$. To ease notation we will write $\varphi_\theta(t) := \varphi_\theta(t, (\tau_j, z_j)_{j=1}^{m_t})$, but keep in mind its dependence on the

history. In (6.3) we have chosen to explicitly model the censoring mechanism using an at-risk-indicator $Y_t = 1(t < T)$, while we nonparametrically model the intensity prior to censoring.

The only restriction for the architecture of the neural network is that the first layer is compatible with the input dimension after a potential pre-transformation of $(\tau_j, z_j)_{j=1}^{m_t}$, and that the output activation needs to be non-negative, since the intensity needs to be non-negative. Once an architecture of the neural network is chosen, we can consider the negative log-likelihood as a function of $\theta \in \Theta$,

$$\ell(\theta) = \sum_{i=1}^{n} \left(\int_0^{T_i} \varphi_\theta(s) \mathrm{d}s - \sum_{j=1}^{k_i} I_{i,j} \cdot \log \varphi_\theta(\tau_{i,j}) \right). \tag{6.4}$$

In order to leverage the automatic differentiation capabilities of TensorFlow, the loss function needs to be expressed in terms of differentiable functions. The integral term of (6.4) does not in general have an analytically closed form, and therefore cannot be automatically differentiated. Therefore, we consider a discretization of the integral in the following way. For each $i = 1, \ldots, n$, let $(\delta_{i,\ell})_{\ell=0}^{L}$ be an equidistant grid $0 = \delta_{i,0} < \delta_{i,1} < \cdots < \delta_{i,L} = T_i$ with grid coarseness $\Delta_i = \delta_{i,1} - \delta_{i,0}$. Then we define the approximated negative log-likelihood function as

$$\tilde{\ell}(\theta) = \sum_{i=1}^{n} \left(\Delta_i \cdot \sum_{\ell=1}^{L} \varphi_\theta(\delta_{i,\ell}) - \sum_{j=1}^{k_i} I_{i,j} \cdot \log \varphi_\theta(\tau_{i,j}) \right). \tag{6.5}$$

This function approximates (6.4) for large values of L and is expressed solely in terms of differentiable functions of θ. Hence, we will consider training the model (6.3) using $\tilde{\ell}$ as loss function for a user specified value of L. Computation of $\tilde{\ell}$ for a batch $I \subseteq \{1, \ldots, n\}$ requires $|I| \cdot L + \sum_{i \in I} k_i$ forward passes from the network φ_θ.

6.4 Module usage

The module **intensitymodel** contains a function for transforming raw sequence data to a Keras compatible data set, two custom preprocessing layers, and a model class for training neural network models for intensities. We will now showcase the functionality of the module. We will spend some time on explaining the data preparation and data input form, since this knowledge is required for a practitioner to use our module for specifying custom neural network architectures.

Data preparation

As an example we will consider simulated data from a multivariate Hawkes process with exponentially decreasing kernel functions, that is, the $\mathcal{F}^D$-intensity of the counting process $N^d, d \in D$, is given by

$$\lambda_t^{d,D} = \mu_t^d + \sum_{h \in D} \int_0^t f_{dh}(t-s) \mathrm{d}N_s^h = \mu_t^d + \sum_{h \in D} \sum_{\tau_j < t, z_j = h} f_{dh}(t - \tau_j) \tag{6.6}$$

where μ_t^d are baseline intensities and f_{dh} are kernel functions describing how N^d depends on N^h. Here we consider simulating from a 2-dimensional Hawkes process with

$$\mu_t^1 = \mu_t^2 = 0.25, \quad f_{11}(t) = f_{12}(t) = f_{22}(t) = 0.5 \cdot \exp(-3t) \quad \text{and} \quad f_{21}(t) = 0 \tag{6.7}$$

such that N^1 depends on itself and N^2, and N^2 only depends on itself. We simulate the data using the Python library **tick** [Bacry et al., 2017].

```
import numpy as np
from tick.hawkes import SimuHawkesExpKernels

n_nodes = 2
adjacency = 0.5 * np.ones((n_nodes, n_nodes))
adjacency[1, 0] = 0
decays = 3 * np.ones((n_nodes, n_nodes))
baseline = 0.25 * np.ones(n_nodes)
hawkes = SimuHawkesExpKernels(adjacency=adjacency,
                              decays=decays,
                              baseline=baseline,
                              verbose=False,
                              seed=123)
hawkes.end_time = 100
hawkes.track_intensity(0.01)

N = 100
raw_data = []
for i in range(N):
    hawkes.max_jumps = int(np.random.randint(5, 10, size=1))
    hawkes.simulate()
    events = hawkes.timestamps
    T = 5
    raw_data.append([T, events])
    hawkes.reset()
```

The raw data format is a list of observation. Each observation is itself a list, where the first element is the censoring time, and the subsequent elements are 1-dimensional numpy arrays of event times. The raw data can be turned into a **tf.data.Dataset** using **intensitymodel.create_dataset**, which contains functionality for splitting into training and validation data.

```
from intensitymodel import create_dataset

TRAIN_SIZE = 0.8
data = create_dataset(raw_data)
data_train = data.take(int(TRAIN_SIZE * N))
data_valid = data.skip(int(TRAIN_SIZE * N))
```

A single sample returned by the data object is a rank 2 tensor, where the first column are the event times and the second column are event types. The first row is reserved for the censoring time, which is not associated with a mark.

```
x = next(data.batch(1).as_numpy_iterator())[0]
print(x)
```

```
--------------------------------------------------------------------------------
[[5.        nan]
```

```
 [0.75142485 2.        ]
 [1.822723   1.        ]
 [6.5150847  2.        ]
 [6.6051927  2.        ]
 [6.6147637  2.        ]
 [7.234427   2.        ]
 [7.2759256  1.        ]
 [       nan        nan]
 [       nan        nan]]
```

Note that all samples are padded to be the same length. This is in order to optimize the data input pipeline, and the padding is removed when the sample is passed through the network by using our pre-processing layers.

Pre-processing layers

The module contains two pre-processing layers that transforms samples as part of the forward pass of the neural network. Let us first describe the structure of data used for the forward pass. The censoring time T is used for computing the loss function (6.5), but is not a part of the forward pass, and is therefore removed. In addition to the event times and marks, we need to provide the time point at which the intensity should be evaluated, which is appended to the end of the data. This is done automatically during training, but here we do it manually to make things explicit. A generic data point has the following structure, where we intend to evaluate the intensity at $t = 7$:

```
x = np.vstack((x, np.array([7.0, 0.0])))
print(x)
```

```
--------------------------------------------------------------------------------
[[0.75142485 2.        ]
 [1.82272303 1.        ]
 [6.51508474 2.        ]
 [6.60519266 2.        ]
 [6.61476374 2.        ]
 [7.23442698 2.        ]
 [7.27592564 1.        ]
 [       nan        nan]
 [       nan        nan]
 [7.         0.        ]]
```

The pre-processing layer **intensitymodel.LaggedSequence** creates lags of the event times prior to t and removes the padding, i.e.,

$$(t, (\tau_j)_{j \geq 1}) \mapsto (t - 0, t - \tau_1, t - \tau_2, \ldots, t - \tau_{m_t}).$$

The first element $t - 0$ is then associated with the mark 0, which represent the time since trial start. By adding the event $\tau_0 = 0$ with mark $z_0 = 0$ we ensure that the forward pass is well-defined for $t \leq \tau_1$.

```
from intensitymodel import LaggedSequence

lagged_sequence = LaggedSequence()
print(lagged_sequence(x))
```

```
tf.Tensor(
[[[7.        ]
  [6.248575  ]
  [5.177277  ]
  [0.48491526]
  [0.39480734]
  [0.38523626]]], shape=(1, 6, 1), dtype=float32)
```

The output is a tensor of shape `(batch_size, time_steps, num_features)`, which can then be processed by any standard recurrent layer.

We also provide a pre-processing layer **intensitymodel.EmbeddingWithNAN**, which does a standard embedding of the marks into $\mathbb{R}^M$, where M is user specified, while removing the padding and only considering events prior to t. In other words each mark z_j is one-hot-encoded using a dummy variable $z_j^* \in \{0,1\}^{|\mathcal{M}|+1}$, and then linearly transformed by a matrix B with trainable parameters:

$$(t, (\tau_j, z_j)_{j \geq 1}) \mapsto (Bz_0^*, Bz_1^*, Bz_2^*, \ldots, Bz_{m_t}^*)$$

In this way each column of B corresponds to a dense representation of the event marks, and this dense representation is trainable.

```
from intensitymodel import EmbeddingWithNAN

embedding_layer = EmbeddingWithNAN(input_dim=3, output_dim=4)
print(embedding_layer(x))
```

```
tf.Tensor(
[[[-0.02994033  0.01735939 -0.01852044  0.04576658]
  [ 0.03496361 -0.01699029  0.04163836  0.03269113]
  [ 0.02381103 -0.01881518  0.00110693 -0.00490152]
  [ 0.03496361 -0.01699029  0.04163836  0.03269113]
  [ 0.03496361 -0.01699029  0.04163836  0.03269113]
  [ 0.03496361 -0.01699029  0.04163836  0.03269113]]], shape=(1, 6, 4), dtype=float32)
```

Both pre-processing layers are subclasses of **keras.layers.Layer**, and inherits the functionality of this class.

Specifying the network architecture

We will now show how the network architecture can be specified by using the Keras functional model API. Here one build an acyclic directed graph of computational layers that are compiled into a computational graph by specifying the leaf nodes. Below we see an example of this with our **intensitymodel.IntensityRNN** class together with our custom pre-processing layers.

```
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from intensitymodel import IntensityRNN, LaggedSequence, EmbeddingWithNAN

input = layers.Input(shape = (None, 2))
x1 = LaggedSequence()(input)
x2 = EmbeddingWithNAN(input_dim = 3, output_dim = 4)(input)
x = tf.concat([x1, x2], axis = -1)
x = layers.LSTM(units = 64, activation = 'tanh', return_sequences = True)(x)
x = layers.LSTM(units = 32, activation = 'tanh')(x)
x = layers.Dense(units = 16, activation = 'tanh')(x)
x = layers.ActivityRegularization(l2 = 0.1)(x)
x = layers.Dense(units = 8, activation = 'tanh')(x)
x = layers.ActivityRegularization(l2 = 0.1)(x)
x = layers.Dense(1, activation = 'softplus')(x)
output = layers.ActivityRegularization(l2 = 0.1)(x)

model = IntensityRNN(inputs = input, outputs = output)
```

First we specify the input shape, which is `(None, 2)` since we allow input sequences of arbitrary length with 2 features – the events times and marks. The inputs are then pre-processed using our custom pre-processing layers, and the results are concatenated into a single tensor of shape `(None, 5)`, since the embedding dimension is set to 4. The processed sequences are then passed to a stacked pair of recurrent LSTM layers [Hochreiter and Schmidhuber, 1997], and the last recurrent output is passed to three stacked dense layers with $||\cdot||_2$-regularization of the outputs. The activation of the output layer is 1-dimensional, and we specify a non-negative activation function to ensure non-negativity of the estimated intensity function. Lastly, we build the computational graph from the directed acyclic graph of layers with our **intensitymodel.IntensityRNN** class, which is a subclass of **keras.Model**.

Naturally, the above is only an example of an architecture of the neural network, and one can experiment with different layers, units, activations, regularizers, etc. However, we recommend using the following template for specifying the architecture.

```
input = layers.Input(shape = (None, 2))
x1 = LaggedSequence()(input)
x2 = EmbeddingWithNAN(input_dim, output_dim)(input)
x = tf.concat([x1, x2], axis = -1)
...
...
...
output = layers.Dense(1, activation = 'softplus')(x)
model = IntensityRNN(inputs = input, outputs = output)
```

Model training

We use the built-in functionality of Keras for training neural network models, which has numerous optimizers and stopping criteria built in. Below is an example which uses the Adam optimizer with user specified learning rates, and an early stopping criterion based on the loss of the validation data.

```
opt = keras.optimizers.Adam(lr=0.002, decay=0.0005)
early_stop = tf.keras.callbacks.EarlyStopping(
    monitor = 'val_loss',
    patience = 3,
    min_delta = 2,
    restore_best_weights = True
    )

model.compile(optimizer = opt, L = 50, d = 1)

BATCH_SIZE = 20
MAX_EPOCHS = 5
model.fit(x = data_train.batch(BATCH_SIZE),
          validation_data = data_gen_valid.batch(int((1-TRAIN_SIZE) * N)),
          callbacks = [early_stop],
          epochs = MAX_EPOCHS)
-------------------------------------------------------------------------------------------
Epoch 1/5
4/4 [==================] - 18s 2s/step - train_loss: 89.0038 - val_loss: 3.5268
Epoch 2/5
4/4 [==================] - 6s 2s/step - train_loss: -52.3426 - val_loss: -1.2048
Epoch 3/5
4/4 [==================] - 6s 2s/step - train_loss: -130.3572 - val_loss: -3.9064
Epoch 4/5
4/4 [==================] - 6s 2s/step - train_loss: -176.5087 - val_loss: -5.4356
Epoch 5/5
4/4 [==================] - 6s 2s/step - train_loss: -205.7987 - val_loss: -6.3774
```

When compiling the models, we specify the approximation accuracy L of the approximated negative log-likelihood function (6.5), and we also specify the mark point of interest d, which in this case is 1, i.e., we are telling the model that we are interested in fitting $\lambda^{1,\{1,2\}}$. The loss function is directly implemented into the model class, so we do not have to pass a loss function to the compile method.

As mentioned previously, to compute the loss function (6.5) for a batch $I \subset \{1, \dots, n\}$ with approximation accuracy L requires $|I| \cdot L + \sum_{i \in I} k_i$ forward passes of the network. In the above example we used batches of size 20, where one forward pass of the batch and one gradient step of the Adam optimizer took 2 seconds on a standard laptop.

We have implemented the loss function such that all forward passes are computed in parallel for efficiency. Furthermore, the loss function is written entirely using TensorFlow, such that we can make use of TensorFlow's built-in tools for converting an eager computational graph to a static computational

graph for significant computational speedup.[1] In the training example above, the first epoch takes 12 seconds more than the following. This is the time used for making one forward pass of a batch in the eager graph. Based on this one eager forward pass, TensorFlow builds a static graph for the subsequent forward passes, which take 2 seconds on average. In other words, the time used for one forward pass of the network is reduced 83% in this concrete example due to the efficiency of the implementation.

Extracting results

The estimated intensity $\hat{\lambda}_t = 1(t < T)\varphi_{\hat{\theta}}(t)$ and compensator $\hat{\Lambda}_t = \int_0^t \hat{\lambda}_s \mathrm{d}s$ can be extracted from the **IntensityRNN** class by providing a history $(\tau_j, z_j)_{j \geq 1}$ and censoring time T.

```
intensity = model.intensity(history=data, T=T)
compensator = model.compensator(history=data, T=T, approx=1000)

print(intensity(3.2))
print(compensator(3.2))
```

```
--------------------------------------------------------------------------------
tf.Tensor(1.5482968, shape=(), dtype=float32)
tf.Tensor(3.0606036, shape=(), dtype=float32)
```

The compensator is computed using a time discretization of the integral, where the approximation accuracy is determined by the user. In this case we pre-compute 1000 values of the estimated intensity over a grid from 0 to T, which are used to compute the integral when calling the function.

6.5 Simulations

In this section we will consider applying our modeling framework to simulated data. We will consider three classes of point processes. First we will consider renewal processes, where the waiting time between events is i.i.d., and where the intensity only depends on the most recent event. Next we will consider 1-dimensional Hawkes processes, where the intensity depends on the entire event history. Lastly, we will consider the 2-dimensional Hawkes process that we described in Section 6.4.

Renewal processes

Renewal processes are point processes $(\tau_j)_{j \geq 0}$ where the waiting times $(\tau_{j+1} - \tau_j)_{j \geq 0}$ are i.i.d. In this case there is only one mark $\mathcal{M} = \{1\}$, so we are modeling the $\mathcal{F}^{\{1\}}$-intensity $\lambda^{1,\{1\}}$ of the counting process $N_t^1 = \sum_{j \geq 1} 1(\tau_j \leq t)$ We consider three different waiting time distributions:

- Exponentially distributed $\tau_{j+1} - \tau_j \sim \mathrm{Exp}(1)$.
- Log-normally distributed $\tau_{j+1} - \tau_j \sim \mathrm{Lognormal}(3, 0.2)$.
- Uniformly distributed $\tau_{j+1} - \tau_j \sim \mathcal{U}([3, 10])$.

In each case we will use the following architecture of the neural network, where we note that there is no use for an embedding layer, since we only have one mark.

[1] For details see `https://www.tensorflow.org/guide/function`.

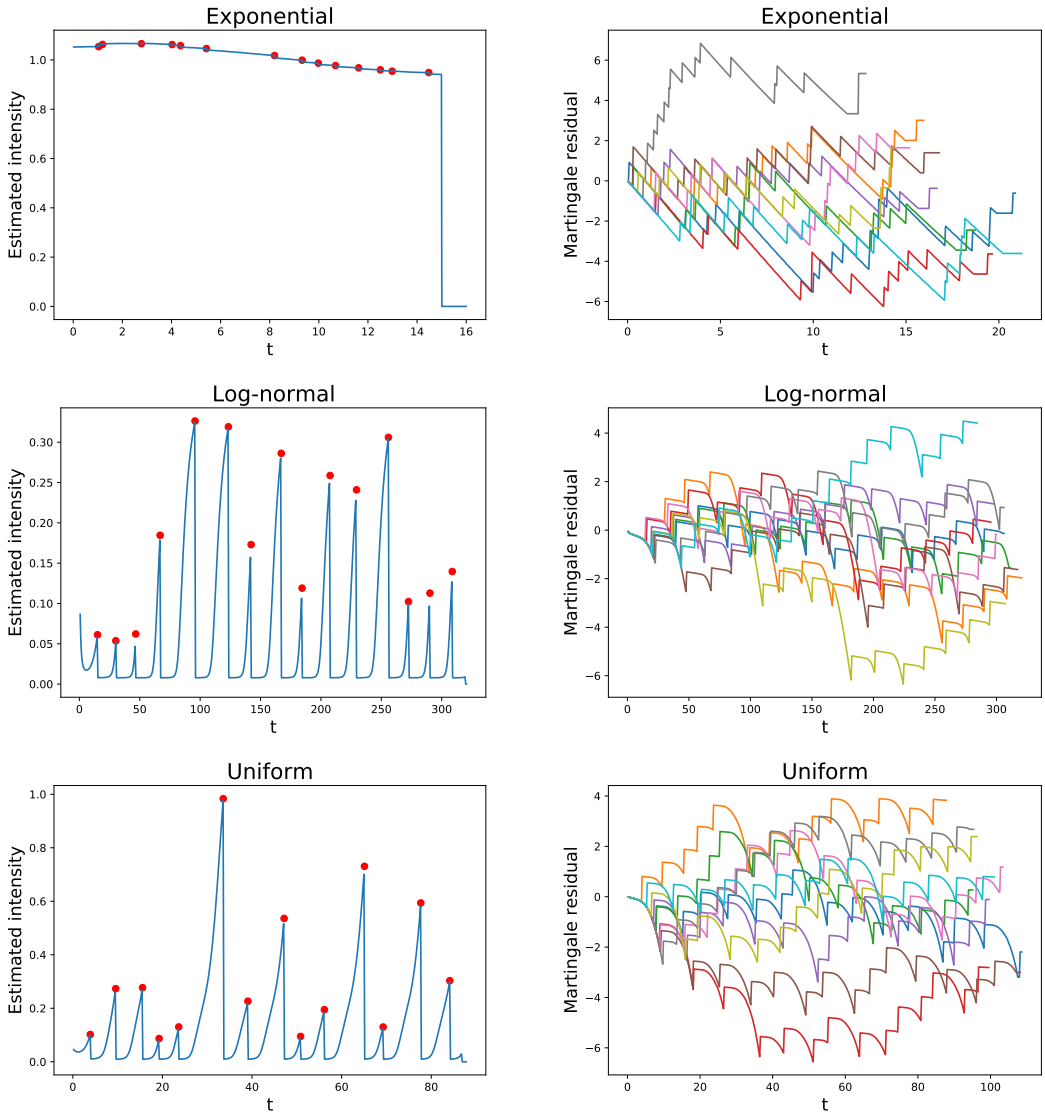

Figure 6.1: Left column: Predicted intensity functions for each of the three waiting time distributions evaluated on an independent test sample not used for training. Red points indicate event times. Right column: Martingale residuals for each of the three models for 10 test samples not used for training.

```
input = layers.Input(shape=(None, 2))
x = LaggedSequence()(input)
x = layers.LSTM(64, activation='tanh')(x)
x = layers.Dense(32, activation='tanh')(x)
x = layers.Dense(16, activation='tanh')(x)
x = layers.Dense(8, activation='tanh')(x)
output = layers.Dense(1, activation='softplus')(x)
model = IntensityRNN(inputs = input, outputs = output)
```

For each waiting time distribution we simulate $n = 500$ point processes that we use for training. The network is fitted using an Adam optimizer with learning rate 0.002 and decay 0.0005, and an approximation of the negative log-likelihood loss function with $L = 50$. We used a batch size of 20, and fitted the models using 10 epochs. Both the architecture of the network and the training related hyperparameters were chosen ad-hoc without any tuning.

We will evaluate the performance of the model on an independent test data set not used for training. By this we mean that the parameters of the neural network are obtained by fitting the model on a training data set to determine its functional form. When we evaluate the models on a test data set, we consider the parameters as fixed, but plug in the event history of the test point processes.

In Figure 6.1 we consider the predicted intensity on a test sample, and also the martingale residuals

$$\hat{M}_t = N_t - \hat{\Lambda}_t,$$

for 10 test samples, which are approximately zero-mean Gaussian processes if the model fits. Firstly, when the waiting time distribution is exponentially distributed with scale 1, we know that the intensity is $\lambda_t = 1(t < T)$. We see that the predicted intensity lies almost constantly around 1, so that the recurrent neural network has captured the functional form of the true intensity. We also see that the martingale residuals are centered around 0 with a slight tendency of underestimation. Secondly, for the log-normally distributed waiting times, we see that the predicted intensities has captured that the intensity drops almost to zero after an event, but increases after a while. The martingale residuals lie nicely around zero without any noticeable outliers. Lastly, for the uniformly distributed waiting times, the intensity drops to zero after an event, which is natural since the waiting times are $\mathcal{U}([3, 10])$-distributed, so an event cannot occur within 3 time units after the last event. In this case the martingale residuals also lie around zero without major outliers.

We conclude that our recurrent neural network model for the intensity fits the renewal process data well. Furthermore, we note that we used the same neural network architecture without any tailoring for each of the waiting time distributions. However, one might argue that renewal processes are too simple, since the intensity is independent of the history given the most recent event time. On the other hand, the recurrent neural network effectively learns to ignore the irrelevant information.

1-dimensional Hawkes process

Next we will consider 1-dimensional Hawkes processes, where the intensity depends on the entire event history. We simulate 500 point processes from the Hawkes process with intensity

$$\lambda_t^1 = 0.25 + \sum_{\tau_j < t} 0.5 \cdot e^{-3(t-\tau_j)}. \tag{6.8}$$

We use the same architecture, optimizer, batch size and number of training epochs as for the renewal processes.

Predicted intensity for an independent test sample, and martingale residuals for 10 independent test samples can be seen in Figure 6.2. First we note that the models captured the excitatory nature of the Hawkes process, where an event increases the intensity of a new event. Moreover, we see that the intensity exponentially decreases after an event, and that the baseline intensity of 0.25 is well predicted by our neural network model. We see that the martingale residuals are well centered around zero, and there is no sign of over- or underestimation of the intensity. We conclude that our recurrent neural network models captures the functional form of the intensity of the Hawkes process well. It also efficiently models

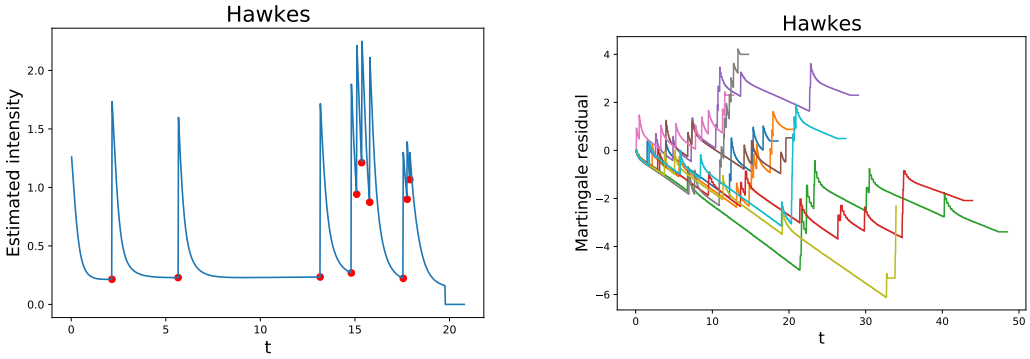

Figure 6.2: Left: Predicted intensity function on an independent sample from a 1-dimensional Hawkes process given by (6.8). Red points indicate event times. Right: Martingale residuals for 10 independent test samples.

the entire history of the process. Again we note that we have not tailored the architecture of the neural network to the specific functional form (6.8) of the Hawkes process.

2-dimensional Hawkes process

Finally, we consider the 2-dimensional Hawkes process (6.6) with parameters (6.7) where the process N^1 depends on its own history and the history of N^2, and N^2 only depends on its own history. Here the task is to estimate the $\mathcal{F}^{\{1,2\}}$-intensity $\lambda^{1,\{1,2\}}$ of N^1. In this case we have two marks, so we add an embedding layer to the model, which has the following architecture.

```
input = layers.Input(shape=(None, 2))
x1 = LaggedSequence()(input)
x2 = EmbeddingWithNAN(input_dim=3, output_dim=5)(input)
x = tf.concat([x1, x2], axis = -1)
x = layers.LSTM(128, activation='tanh')(x)
x = layers.Dense(64, activation='tanh')(x)
x = layers.Dense(32, activation='tanh')(x)
x = layers.Dense(16, activation='tanh')(x)
output = layers.Dense(1, activation='softplus')(x)
model = IntensityRNN(inputs = input, outputs = output)
```

Again we simulate $n = 500$ samples, and we use the same optimizer, batch size and number of epochs as in the two previous cases. The results can be seen in Figure 6.3.

We see that the predicted intensity correctly models the excitatory nature of the Hawkes process, where we have exhibition of both events of N^1 and N^2, with an exponentially decreasing intensity between events of either process. The baseline is also correctly estimated to approximately 0.25. From the martingale residual plot we see that the intensity function is fairly well estimated with a single outlier in this particular case.

In conclusion our neural network model correctly captures the functional form of the intensity function in the presence in the case of a 2-dimensional Hawkes process, where the coordinate of interest

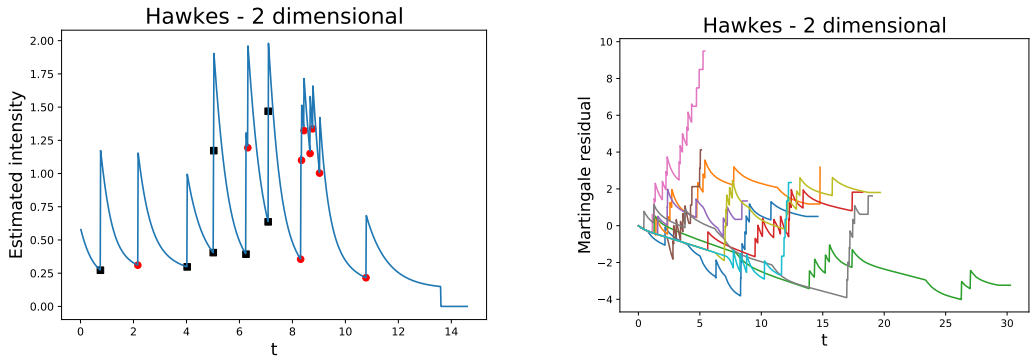

Figure 6.3: Left: Predicted intensity function on an independent sample from a 1-dimensional Hawkes process given by (6.8). Red points indicate event of N^1, while black squares indicate events of N^2. times. Right: Martingale residuals for 10 independent test samples.

depends on the second coordinate. As before, we emphasize that the network architecture has not been chosen according to the functional form of the Hawkes process.

6.6 Discussion

In this chapter we have presented a Python module for estimating intensity functions of marked point processes using recurrent neural networks. Our implementation uses the Keras high level API to the TensorFlow neural network library. We have made a highly efficient implementation of the marked point process negative log-likelihood function as a loss function for training recurrent neural network models. The advantage of the implementation is that it gives the practitioner full freedom to choose the architecture of the neural network as well as the training procedure via the Keras API.

Recurrent neural networks have previously been considered for estimating intensity functions, but to the best of our knowledge, this implementation is the first that does not rely on a fixed architecture and training procedure. In this regard we consider our contribution to be novel.

Our simulation study suggests that a fairly standard and non-specialized architecture is able to flexibly model the intensity function of various point processes without tailoring it to a known functional form. Here we have showcased our implementation on renewal processes, and 1- and 2-dimensional Hawkes processes with exponential kernels.

We note that this framework for modeling intensity functions is limited in the sense that is does not give interpretability of the influence of past events on the future. In this sense it is a black box model. However, for our intended usage of the model in connection with nonparametric intensity estimation for our conditional local independence test in Chapter 5 a black box model is sufficient.

Future development of the implementation includes the opportunity to perform a more principled architecture selection based on a validation data set, and automatic procedures for doing hyperparameter calibration. Furthermore, even though the martingale residuals give an impression of whether the intensity is under- or overestimated, it would be beneficial to be able to compare with bootstrapped martingale residuals under the true model in a simulation setting.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Steen A. Andersson, David Madigan, and Michael D. Perlman. Alternative Markov properties for chain graphs. *Scand. J. Statist.*, 28(1):33–85, 2001. ISSN 0303-6898.

Signorell Andri et mult. al. *DescTools: Tools for Descriptive Statistics*, 2021. URL `https://cran.r-project.org/package=DescTools`. R package version 0.99.41.

E. Bacry, M. Bompaire, S. Gaïffas, and S. Poulsen. tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*, July 2017.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Wicher Bergsma. *Testing conditional independence for continuous random variables*. Eurandom, 2004.

Wicher Bergsma. Nonparametric testing of conditional independence by means of the partial copula. *SSRN Electronic Journal*, 01 2011.

J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution Free Tests of Independence Based on the Sample Distribution Function. *The Annals of Mathematical Statistics*, 32(2):485 – 498, 1961. doi: 10.1214/aoms/1177705055. URL `https://doi.org/10.1214/aoms/1177705055`.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

François Chollet et al. Keras. `https://keras.io`, 2015.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Vanessa Didelez. Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics*, 34(1):169–185, 2006.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society, Series B*, 70(1):245–264, 2008.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Morten Frydenberg. The chain graph Markov property. *Scand. J. Statist.*, 17(4):333–353, 1990. ISSN 0303-6898.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Wassily Hoeffding. A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics*, 19 (4):546 – 557, 1948.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(22):613–636, 2007.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11): 1–26, 2012.

S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1):31–57, 1989. ISSN 0090-5364.

Steffen L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

Steffen L. Lauritzen and Thomas S Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133 – 3164, 2009.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410, 1995.

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *Ann. Statist.*, 48(1):539–559, 2020.

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th conference on Uncertainty in Artificial Intelligence*, pages 350–360, 2018.

Rohit K Patra, Bodhisattva Sen, and Gábor J Székely. On a nonparametric notion of residual and its applications. *Statistics & Probability Letters*, 109:208–213, 2016.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0934613737.

Judea Pearl. *Causality*. Cambridge University Press, second edition, 2009.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Lasse Petersen. Sparse Learning in Gaussian Chain Graphs for State Space Models. In *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pages 332–343, Prague, Czech Republic, 11–14 Sep 2018. PMLR.

Lasse Petersen and Niels Richard Hansen. Nonparametric conditional local independence testing. 2021a.

Lasse Petersen and Niels Richard Hansen. Testing Conditional Independence via Quantile Regression Based Partial Copulas. *Journal of Machine Learning Research*, 22(70):1–47, 2021b.

Niklas Pfister and Jonas Peters. *dHSIC: Independence Testing via Hilbert Schmidt Independence Criterion*, 2019. URL `https://CRAN.R-project.org/package=dHSIC`. R package version 2.1.

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

James M. Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

Tore Schweder. Composable Markov processes. *Journal of Applied Probability*, 7(2):400–410, 1970.

Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Thomas Verma and Judea Pearl. Causal networks: semantics and expressiveness. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, pages 69–78, 1990.

Sebastian Weichwald, Martin E. Jakobsen, Phillip B. Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pages 27–36, Vancouver, CA, 08–14 Dec 2020. PMLR.

Shuai Xiao, Junchi Yan, Mehrdad Farajtabar, Le Song, Xiaokang Yang, and Hongyuan Zha. Learning time series associated event sequences with recurrent point process networks. *IEEE transactions on neural networks and learning systems*, 30(10):3124–3136, 2019.

Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 632–639, 2002.

Wei Zhang, Thomas Panum, Somesh Jha, Prasad Chalasani, and David Page. Cause: Learning granger causality from event sequences using attribution methods. In *International Conference on Machine Learning*, pages 11235–11245. PMLR, 2020.