Causal Inference for Extremes

Gnecco, Nicola

Causal Inference for Extremes

by

Nicola Gnecco

A thesis submitted to the
Geneva School of Economics and Management,
University of Geneva, Switzerland,
in fulfillment of the requirements for the degree of
PhD in Statistics

Members of the thesis committee:
Prof. Sebastian Engelke, Adviser, University of Geneva
Prof. Davide La Vecchia, Chair, University of Geneva
Prof. Richard A. Davis, Columbia University
Prof. Nicolai Meinshausen, ETH Zurich

Thesis No. 113
October 2022

La Faculté d'économie et de management, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par-là, émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 19 octobre 2022

Dean

Markus MENZ

Impression d'après le manuscrit de l'auteur

Acknowledgements

I want to thank my supervisor Sebastian Engelke for his continuous support and encouragement over the last four years. If I enjoyed every moment of my PhD is primarily thanks to you. Thank you for the countless discussions about research, math, and life. And thank you for initiating me to the winter bathing tradition in the Lac Léman!

I also want to thank Jonas Peters for his constant presence during the path. Thanks for the beautiful research period in Copenhagen and for teaching me all there is to learn about causality.

Thanks to my co-authors, Nicolai Meinshausen, Rune Christiansen, Martin Emil Jakobsen, Niklas Pfister, and Edossa Merga Terefe, for the fruitful collaborations, the exciting discussions, and fantastic teamwork.

To all my colleagues at the Research Center of Statistics and the Copenaghen Causality Lab, thank you for creating a great atmosphere and making me feel at home.

To my friends, Cesare Miglioli and Alberto Quaini, who can understand the challenges and the joys of being a statistician, thank you for the fantastic time in Geneva.

And thank you to my father Fabio, my mother Francesca, my siblings Anna, Pietro and Allegra, and my grandmother Anna. Life is beautiful when we are together!

Abstract

This thesis develops novel statistical methodologies to bring closer the fields of extreme value theory and causality. It revolves around two independent axes of research.

The first axis studies causal discovery for extreme events, where one can infer the causal structure of a system by exploiting the signal in the tails of the variables. In the first project, we introduce a causal coefficient that identifies the causal relationship of heavy-tailed pairs of variables. Then, we propose a computationally highly efficient algorithm based on this causal tail coefficient to recover the causal order of a set of variables. Finally, we compare our method to other well-established and non-extremal approaches in causal discovery on synthetic and real data.

The second axis of research develops flexible predictive models for extremes and distribution generalization. The second project of this thesis develops a quantile regression method to estimate extreme quantiles given a large set of predictors. Our method combines the flexibility of the random forests with the extrapolation guarantees of the generalized Pareto distribution. In simulations, our method is competitive with both classical quantile regression methods and existing regression approaches from extreme value theory. Finally, we apply our methodology to extreme quantile prediction for U.S. wage data. The third project of this thesis studies the problem of distribution generalization from a causal perspective. We assume the data comes from different environments that shift the mean of the predictors so that the training and test distributions are different. We model distributional shifts with the concept of causal intervention. Here, we propose a method to learn a nonparametric function with invariant predictions across environments and as predictive as possible, defined as the invariant most predictive (IMP) function. We show identification of the IMP, provide minimax guarantees over unseen environments over the class of square-integrable functions, and propose an adaptation of the regression tree algorithm to learn the IMP function nonparametrically in large dimensions.

Résumé

Cette thèse développe de nouvelles méthodologies statistiques pour rapprocher les domaines de la théorie des valeurs extrêmes et de la causalité. Il s'articule autour de deux axes de recherche indépendants.

Le premier axe étudie la découverte causale pour les événements extrêmes, où l'on peut déduire la structure causale d'un système en exploitant le signal dans les queues des variables. Dans le premier projet, nous introduisons un coefficient causal qui identifie la relation causale des paires de variables à queue lourde. Ensuite, nous proposons un algorithme de calcul hautement efficace basé sur ce coefficient de queue causale pour récupérer l'ordre causal d'un ensemble de variables. Enfin, nous comparons notre méthode à d'autres approches bien établies et non extrêmes de découverte causale sur des données synthétiques et réelles.

Le deuxième axe de recherche développe des modèles prédictifs flexibles pour les extrêmes et la généralisation de la distribution. Le deuxième projet de cette thèse développe une méthode de régression quantile pour estimer les quantiles extrêmes étant donné un large ensemble de prédicteurs. Notre méthode combine la flexibilité des forêts aléatoires avec les garanties d'extrapolation de la distribution de Pareto généralisée. Dans les simulations, notre méthode est compétitive à la fois avec les méthodes classiques de régression quantile et avec les approches de régression existantes issues de la théorie des valeurs extrêmes. Enfin, nous appliquons notre méthodologie à la prédiction des quantiles extrêmes pour les données sur les salaires aux États-Unis. Le troisième projet de cette thèse étudie le problème de la généralisation de la distribution dans une perspective causale. Nous supposons que les données proviennent de différents environnements qui modifient la moyenne des prédicteurs de sorte que les distributions d'apprentissage et de test sont différentes. Nous modélisons les changements de distribution avec le concept d'intervention causale. Ici, nous proposons une méthode pour apprendre une fonction non paramétrique avec des prédictions invariantes à travers les environnements et aussi prédictive que possible, définie comme la fonction invariante la plus prédictive (IMP). Nous montrons l'identification de l'IMP, fournissons des garanties minimax sur différents environnements sur la classe des fonctions carrées intégrables et proposons une adaptation de l'algorithme d'arbre de régression pour apprendre la fonction IMP de manière non paramétrique en grandes dimensions.

Contents

To my family

Introduction

This thesis develops novel statistical methodologies to build a bridge between the fields of extreme value theory and causality. The goal of causal inference is to infer causal relationships from data. Causal models describe a system's behaviour under interventions providing a richer understanding of the data-generating process (Pearl, 2009b; Peters et al., 2017). Understanding how a system reacts under interventions is crucial whenever one needs to predict the effect of treatments (e.g., in medicine) or policy changes (e.g., in social sciences). Moreover, causal models can be helpful even in pure prediction problems since they provide invariant predictions when the training and test data do not follow the same distribution. Extreme value theory (EVT) is the branch of statistics dealing with the modelling and inference of rare events. On the one hand, the study of univariate extremes is well-understood (de Haan and Ferreira, 2006, Resnick, 2008). On the other hand, several applications require an understanding of extreme joint events (Coles and Tawn, 1996; de Haan et al., 1999; Schlather and Tawn, 2003; Engelke and Hitz, 2020) and extreme events conditionally on a set of predictors (Chernozhukov, 2005; Wang and Tsai, 2009).

This thesis revolves around two independent axes of research. The first axis studies causal discovery for extreme events. While much progress has been made in the formalization of causal language (Spirtes et al., 2000; Pearl, 2009b; Imbens and Rubin, 2015; Peters et al., 2017), there are several situations where causal relationships manifest themselves only in extreme events. As stated by (Cox and Wermuth, 1996, Sec. 8.7), one can view extreme events as 'natural interventions' that convey causal information. From this perspective, one can infer the causal structure of a system by exploiting the signal in the tails of the variables. For example, in hydrology, one can reconstruct the topology of a river network by observing extreme river discharges at different points along the basin (Asadi et al., 2015). From a different angle, one can argue that the causal mechanisms during extreme events differ from those in the tails of the distribution. For example, in financial markets, we often observe regime changes between quiet periods and turmoil. Forbes and Rigobon (2002) describe these extremal causal mechanisms in terms of contagion, i.e., a single significant shock propagating through a given system. In Earth system science, there are also examples of different causal structures between the bulk of the data and the tails. For instance, Seneviratne et al. (2010) study the causal mechanism between air temperature and evapotranspiration of the soil moisture and show a regime change between low and high temperatures. Engelke and Hitz (2020) recently connected the fields of graphical models (Lauritzen, 1996) and extremes. They defined a notion of conditional independence for extremes to build sparse and parsimonious models in high-dimensions (Engelke and Ivanovs, 2021; Engelke and Volgushev, 2020). While graphical models do not necessarily have a causal meaning, future research in causality for extremes will benefit from these novel results.

The second axis of research develops flexible predictive models for extremes and dis-

tribution generalization. The first work in this direction develops a quantile regression method, named Extremal Random Forests (ERF), to estimate extreme quantiles given a large set of predictors. The proposed method could be adapted to estimate treatment effects (Athey et al., 2019) at extreme quantile levels (Deuber et al., 2021) when the quantile regression surface is nonlinear and the treatment has large dimensions. Relevant applications in this direction are in climate science when one is interested in measuring the impact of climate change on extreme weather events.

The second work along this axis introduces a regression method when the training and test distributions are different. Most machine learning algorithms assume that the training and test data are independent and identically distributed (i.i.d.). However, in several applications, the data comes from heterogeneous environments, and it is hard to justify the i.i.d. assumption. Here, we frame the problem of distribution generalization from a causal perspective by modelling distributional shifts using the concept of causal interventions (Meinshausen, 2018; Rothenhäusler et al., 2021). While this work considers the regression setting, we consider extending it to extreme quantile regression as a next step.

Axis 1 – causal discovery for extreme events

The first work of this thesis studies causal discovery in the presence of extreme events. We can look at the goal from two angles. On the one hand, we would like to define a notion of causality for extreme events, where the causal mechanism in the tail may differ from the one in the bulk of the distributions. On the other hand, causal relationships manifest themselves more clearly during extreme events, and thus, we exploit the signal in the distribution's tails to perform classical causal discovery.

Prior to this work, the literature combining causality and extremes was sparse. Notable mentions are the works from Gissibl and Klüppelberg (2018) and Gissibl et al. (2020) about max-linear models, from Naveau et al. (2018) about causal analysis in climate science using extreme events, and Mhalla et al. (2020) who develop a causal discovery method based on the concept of the Kolmogorov complexity of extreme conditional quantiles.

We consider a linear structural causal model (SCM) (Pearl, 2009b; Peters et al., 2017) with independent *heavy-tailed* noise terms that share the same tail index. The goal is to recover the causal order of the variables from observational data. Linear non-Gaussian acyclic models (LiNGAM) exploit non-Gaussian errors to recover the causal structure of the SCM (Shimizu et al., 2006, 2011; Hyvärinen and Smith, 2013). Unlike these methods, which consider the whole data distribution, here we define a causal coefficient that focuses on the bivariate tails of the data. Our *causal tail coefficient* encodes causal relationships between pairs of variables by exploiting the signal in the bivariate tails of the distribution. We propose a nonparametric estimator for the causal tail coefficient and show its consistency. Based on the causal tail coefficient, we devise the Extremal Ancestral Search (EASE) algorithm that recovers a valid causal from data and is consistent as the sample size tends to infinity. After comparing the EASE algorithm to well-established methods in causality, we apply it to river discharges and financial datasets.

Axis 2 – flexible methods for extremes and distribution generalization

The second and third works develop novel, flexible predictive methods for extremes and distribution generalization.

In Chapter 2, we introduce the extremal random forests (ERF) algorithm to predict conditional extreme quantiles when the predictor space has a large dimension. On the one hand, the literature on flexible quantile regression is not well suited when the quantile levels are beyond the range of the observed data (Meinshausen, 2006; Athey et al., 2019). On the other hand, the approaches that estimate extreme conditional quantiles based on the GPD do not scale well with large dimensional predictor spaces (Chernozhukov, 2005; Wang et al., 2012; Youngman, 2019). Here, we propose to combine the GRF from Athey et al. (2019) with the tail extrapolation of the GPD. GRF is an ensemble method that grows trees that split the predictor space according to custom losses, in this case, quantile loss (Athey et al., 2019). The GPD is a limit distribution that describes observations exceeding an increasing threshold and applies to the most common densities (Balkema and de Haan, 1974). It is a parametric distribution indexed by the shape and scale parameter. The shape parameter determines the decay of the tail, differentiating heavy, light, and short-tailed distributions. For a given predictor point in the sample space, we propose to fit a weighted GPD log-likelihood, using the similarity weights from a quantile GRF. The weights from GRF take care of the dimensionality of the predictor space, whereas the GPD deals with the sparse observations at high quantiles. Under simplifying assumptions, we show that the estimated parameters of the GPD are consistent. In practice, the shape of the quantile function is most sensitive to the shape parameter of the GPD. Therefore, we further introduce penalization while fitting the weighted log-likelihood. We compare our ERF algorithm to other quantile regression methods and apply it to the U.S. wage data set (Angrist et al., 2009).

In Chapter 3 we consider the problem of distribution generalization in a regression setup (Quiñonero-Candela et al., 2009) when the data comes from heterogeneous environments. The goal is to develop a predictive method that minimizes the worst-case mean squared prediction error (MSPE) over unseen environments, i.e., it is minimax. We assume that different environments induce shifts in the predictors' means, we allow for hidden confounders, and we consider a possibly large dimensional predictor space.

Rothenhäusler et al. (2021) first introduced this problem in a linear setup. They consider an instrumental variable (IV) model where the causal function is under-identified, i.e., there are more predictors than instruments, and where the instruments are invalid, i.e., they can directly affect the response (Angrist et al., 1996; Imbens, 2014). They cast the problem of distribution generalization from a causal perspective, where the instruments encode different environments and causal interventions describe shifts in such environments. They develop the anchor regression method that interpolates between ordinary least squares (OLS) and IV solution and minimizes the worst-case MSPE. Bühlmann (2020) extends anchor regression to a nonlinear setup and proposes a predictive method based on the gradient boosting algorithm (Friedman, 2001a). However, from Bühlmann (2020) it is unclear whether the proposed method minimizes the worst-case MSPE over unseen environments. Christiansen et al. (2021) develop the NILE algorithm, a predictive method to achieve distribution generalization in the nonlinear anchor regression setup. On the one hand, the NILE algorithm outperforms other well-established methods in experiments. However, on the other hand, it has no minimax guarantees, and, in practice,

it works well only with a few predictors.

Our work extends the anchor regression setup when the predictive functions are nonlinear, and the environments act linearly on the predictors. We develop a flexible method that minimizes the worst-case MSPE over unseen environments and scales well with large dimensional predictor space. First, we define the *invariant most predictive* (IMP) function. The IMP function (i) achieves invariant performance across the environments, and (ii) is as predictive as possible. Using the literature on control functions (Ng and Pinkse, 1995; Newey et al., 1999), we provide identification of the IMP function. Moreover, we prove that the IMP is minimax over the class of square-integrable functions. Finally, we propose an adaptation of the regression trees algorithm from Breiman et al. (1984) to learn the IMP nonparametrically and in large dimensions. In the following steps, we intend to extend the algorithm to random forests and show its consistency. Moreover, we plan to implement and apply the algorithm to real-world datasets in Earth system science and medicine.

Structure of the thesis

This thesis consists of three chapters corresponding to the following projects.

Chapter 1: N. Gnecco, N. Meinshausen, J. Peters, S. Engelke. Causal discovery in heavy-tailed models. *Annals of Statistics, 49(3): 1755 – 1778, 2021.*

Chapter 2: N. Gnecco, E. M. Terefe, S. Engelke. Extremal Random Forests. *Under revision for the Journal of the American Statistical Association, Theory and Methods, https://arxiv.org/abs/2201.12865*.

Chapter 3: N. Gnecco, N. Pfister, J. Peters, S. Engelke. Distribution generalization in semi-parametric models: A control function approach. *Manuscript.*

Supplementary information for each article is in the corresponding appendix at the end of the thesis.

Chapter 1

Causal discovery in heavy-tailed data

Joint work with

Nicolai Meinshausen, Jonas Peters, and Sebastian Engelke

Abstract

Causal questions are omnipresent in many scientific problems. While much progress has been made in the analysis of causal relationships between random variables, these methods are not well suited if the causal mechanisms only manifest themselves in extremes. This work aims to connect the two fields of causal inference and extreme value theory. We define the causal tail coefficient that captures asymmetries in the extremal dependence of two random variables. In the population case, the causal tail coefficient is shown to reveal the causal structure if the distribution follows a linear structural causal model. This holds even in the presence of latent common causes that have the same tail index as the observed variables. Based on a consistent estimator of the causal tail coefficient, we propose a computationally highly efficient algorithm that estimates the causal structure. We prove that our method consistently recovers the causal order and we compare it to other well-established and non-extremal approaches in causal discovery on synthetic and real data. The code is available as an open-access `R` package.

Keywords: causality, extreme value theory, heavy-tailed distributions, non-parametric estimation.

1.1 Introduction and background

Reasoning about the causal structure underlying a data generating process is a key scientific question in many disciplines. In recent years, much progress has been made in the formalisation of causal language (Pearl, 2009b; Spirtes et al., 2000; Imbens and Rubin, 2015). In several situations, causal relationships manifest themselves only in extreme events. As stated by Cox and Wermuth (1996, Sec. 8.7), large interventions (also named natural experiments) often carry information that is likely to be causal. In this light, one can view extreme observations as natural experiments that perturb a given system and facilitate causal analysis. On the one hand, existing causal methodology, focuses on

moment related quantities of the distribution and is not tailored to estimate causal relationships from extreme events. On the other hand, the statistics of univariate extremes is relatively well understood (Resnick, 1987) and there is a large set of tools for the analysis of heavy-tailed distributions. This work attempts to bring the fields of causality and extremes closer.

Let us first consider a bivariate random vector (X_1, X_2) and assume that we are interested in the causal relationship between the two random variables, X_1 and X_2. We consider a linear structural causal model (Pearl, 2009b, Sec. 1.4) over variables including (X_1, X_2) (without feedback mechanisms). We can then distinguish between the six scenarios of causal configurations shown in Figure 1.1 that include X_1, X_2 and possibly a third unobserved random variable X_0. This collection is complete in the following sense. Any structural causal model including X_1 and X_2 is interventionally equivalent (e.g., Peters et al., 2017, Sec. 6.8) to one of the examples shown in Figure 1.1 when taking into account interventions on X_1 or X_2 only. The dashed edges can be interpreted as directed paths induced by a linear structural causal model (SCM); see Section 1.1.1 for a formal definition. Here and in the sequel, we say that "X_1 is the cause of X_2" or "X_1 causes X_2"

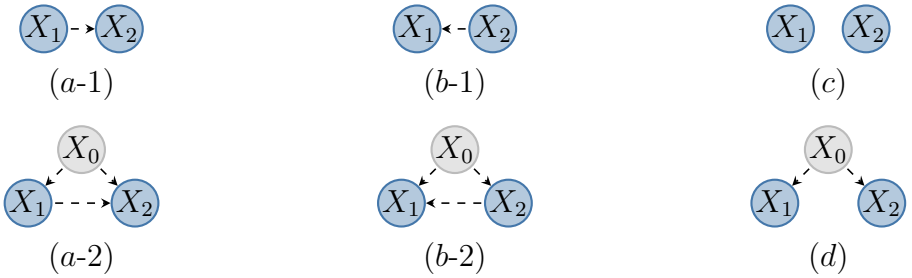


Figure 1.1: The six possible causal configurations between X_1 and X_2, and possibly a third unobserved variable X_0. The variable X_0 will be referred to as a hidden confounder. Formal definitions are included in Section 1.1.1. We will see in Section 1.2.2 that both configurations (a-1) and (a-2), for example, show the same tail coefficient behaviour. The enumeration letters (a)–(d) visualise the cases in Table 1.2.1.

if there is a directed path from X_1 to X_2 in the SCM's underlying directed acyclic graph (DAG). Assume that X_1 is the cause of X_2, and that both variables are heavy-tailed. Intuitively, if the causal relationship is monotonic, then an extremely large value of X_1 should cause an extreme value of X_2. The causal direction should, therefore, be strongly visible in the largest absolute values of the random variables. Here, "extreme" is to be seen in the respective scale of each variable, so it will make sense to consider the copula $\{F_1(X_1), F_2(X_2)\}$, where F_j is the marginal distribution of X_j, $j = 1, 2$. To exploit this intuition, we define the *causal tail coefficient* between variables X_1 and X_2 as

$$\Gamma_{12} := \lim_{u \to 1^-} \mathbb{E}\left[F_2(X_2) \mid F_1(X_1) > u\right] \in [0, 1], \tag{1.1.1}$$

if the limit exists. It reflects the causal relationship between X_1 and X_2 since, intuitively, if X_1 has a monotonically increasing causal influence on X_2, we expect Γ_{12} to be close to one. Conversely, extremes of X_2 will not necessarily lead to extremes of X_1 and therefore, the coefficient Γ_{21}, where the roles of X_1 and X_2 in (1.1.1) are reversed, may be strictly smaller than one. This asymmetry will be made precise in Theorem 1.3 for linear structural causal models with heavy-tailed noise variables, and it forms the basis of our causal discovery algorithm.

A different perspective of our approach goes beyond the usual way of defining causality in the bulk of distribution. Namely, by looking at the signal in the tails, we might recover an extremal causal mechanism that is not necessarily present in the central part of the distribution. An example of this can be observed in financial markets. During calm periods, it is not clear whether any causal relationship exists among the financial variables. However, during turmoil, it is common to observe one specific stock or sector causing very negative (or positive) returns of other stocks (or sectors), displaying an extremal causal mechanism. In the finance literature, the concept of extremal causal mechanism is explained in terms of contagion, i.e., the spread of shocks across different markets (see Forbes and Rigobon, 2002). For example, Rodriguez (2007) uses a copula model with time-varying parameters to explain such contagion phenomena within countries in Latin America and Asia. On the other hand, there are also applications where the causal mechanism is present in the bulk of the distribution but absent in the tails. Seneviratne et al. (2010) present this type of causal relationship between air temperature and the evapotranspiration of soil moisture. As the air temperature increases, the evapotranspiration process increases, too. This continues until the soil moisture resources are reduced to the point that a further increase in the temperature has no causal effect on the evapotranspiration.

Heavy-tailed distributions are an example of non-Gaussian models, which have received some attention in the causal literature. The LiNGAM algorithm (Shimizu et al., 2006) exploits non-Gaussianity through independent component analysis (Comon, 1994) to estimate an underlying causal structure. Misra and Kuruoglu (2016) consider stable noise variables in a Bayesian network and develop a structure learning algorithm based on BIC. The work of Gissibl et al. (2020) also studies causal questions related to extreme events. They consider max-linear models (Gissibl and Klüppelberg, 2018) where only the largest effect propagates to the descendants in a Bayesian network. The work by Naveau et al. (2018) falls within the domain of attribution science. Namely, by studying extreme climate events, they try to answer counterfactual questions such as "what the Earth's climate might have been without anthropogenic interventions". Mhalla et al. (2020) develop a method to estimate the causal relationships between bivariate extreme observations. It relies on the Kolmogorov complexity concept (see Kolmogorov, 1968) adapted to high conditional quantiles. Engelke and Ivanovs (2021) reviews recent work on causality and sparsity in extremes.

The rest of the paper is organised as follows. Sections 1.1.1 and 1.1.2 briefly review structural causal models and some important concepts from extreme value theory. Section 1.2 contains a causal model for heavy-tailed distributions with positive coefficients. We prove that, given the underlying distribution, the causal tail coefficient allows us to distinguish between the causal scenarios shown in Figure 1.1. In Section 1.3, we introduce an algorithm named *extremal ancestral search* (EASE) that can be applied to a matrix of causal tail coefficients and that retrieves the causal order of the true graph, in the population case. We prove that our algorithm estimates a causal order even in the case where the causal tail coefficients are estimated empirically from data, as the sample size tends to infinity. In Section 1.4, we first generalise the results of the previous sections to the case of a structural causal model with real-valued coefficients. To do that, we introduce a more general causal tail coefficient that is sensitive to both the upper and lower tail of the variables. Second, we discuss the robustness properties of EASE in the presence of hidden confounders. Third, we analyse the properties of the causal tail coefficient when the noise variables have different tail indices. Section 1.5 contains experiments on simulated data

and real-world applications. The Appendix consists of six sections. Section A.1 summarises important facts about regularly varying random variables. Section A.2 contains the proofs of the results of the paper. Section A.3 illustrates how the EASE algorithm retrieves a causal order of a DAG. Section A.4 describes the settings used in the simulation study. Section A.5 contains additional figures and tables. Section A.6 presents further results for Section 1.5.2.

1.1.1 Structural causal models

A linear structural causal model, or linear SCM, (Bollen, 1989; Pearl, 2009b, Sec. 1.4) over variables $X_1, \dots, X_p$ is a collection of p assignments

$$X_j := \sum_{k \in \mathrm{pa}(j)} \beta_{jk} X_k + \varepsilon_j, \quad j \in V, \tag{1.1.2}$$

where $\mathrm{pa}(j) \subseteq V = \{1, \dots, p\}$ and $\beta_{jk} \in \mathbb{R} \setminus \{0\}$, together with a joint distribution over the noise variables $\varepsilon_1, \dots, \varepsilon_p$. Here, we assume that the noise variables are jointly independent and that the induced graph $G = (V, E)$, obtained by adding directed edges from the parents $\mathrm{pa}(j)$ to j, is a directed acyclic graph (DAG) with nodes V and (directed) edges $E \subset V \times V$. To ease notation, we adopt the convention to sometimes identify a node with its corresponding random variable. To highlight the fact that $\mathrm{pa}(j)$ depends on a specific DAG G, we write $\mathrm{pa}(j, G)$, $j \in V$.

Structural causal models describe not only observational distributions but also interventional distributions. An intervention on X_j, for example, is defined as replacing the corresponding assignment (1.1.2) while leaving the other equations as they were. In practice, a causal model can be falsified via randomised experiments (see, e.g., Peters et al., 2017, Sec. 6.8).

We define a directed path between node j and k as a sequence of distinct vertices such that successive pairs of vertices belong to the edge set E of G. If there is a directed path from j to k, we say that j is an ancestor of k in G. The set of ancestors of j is denoted by $\mathrm{An}(j, G)$, and we write $\mathrm{an}(j, G) = \mathrm{An}(j, G) \setminus \{j\}$ when we consider the ancestors of j except itself. A node j that has no ancestors, i.e., $\mathrm{an}(j, G) = \varnothing$, is called a source node (or root node). Given two nodes $j, k \in V$, we say that X_j *causes* X_k if there is a directed path from j to k in G. Furthermore, given nodes $i, j, k \in V$, we say that X_i is a *confounder* (or *common cause*) of X_j and X_k if there is a directed path from node i to node j and k that does not include k and j, respectively. Whenever a confounder is unobserved, we say it is a *hidden* confounder or *hidden* common cause. Finally, if $\mathrm{An}(j, G) \cap \mathrm{An}(k, G) = \varnothing$, then we say that there is *no causal link* between X_j and X_k. A graph $G_1 = (V_1, E_1)$ is called a *subgraph* of G if $V_1 \subseteq V$ and $E_1 \subseteq (V_1 \times V_1) \cap E$. Recall that any subgraph of a DAG G is also a DAG. For details on graphical models, we refer to Lauritzen (1996).

1.1.2 Regularly varying functions and random variables

A positive, measurable function f is said to be *regularly varying* with index $\alpha \in \mathbb{R}$, $f \in \mathrm{RV}_\alpha$, if it is defined on some neighbourhood of infinity $[x_0, \infty)$, $x_0 > 0$, and if for all $c > 0$, $\lim_{x \to \infty} f(cx)/f(x) = c^\alpha$. If $\alpha = 0$, f is said to be slowly varying, $f \in \mathrm{RV}_0$.

A random variable X is said to be *regularly varying* with index α if

$$\mathbb{P}(X > x) \sim \ell(x) x^{-\alpha}, \quad x \to \infty,$$

for some $\ell \in \mathrm{RV}_0$, where for any function f and g, we write $f \sim g$ if $f(x)/g(x) \to 1$ as $x \to \infty$. If X is regularly varying with index α then cX is also regularly varying with the same index, for any $c > 0$. For example, random variables with a Student's-t, Pareto, Cauchy, or Fréchet distribution are regularly varying.

A characteristic property of regularly varying random variables is the *max-sum-equivalence*. The idea is that large sums of independent random variables tend to be driven by only one single large value. For this reason, the tail of the distribution of the maximum is equal to the tail of the distribution of the sum. For a rigorous formulation see Appendix A.1. We refer to Embrechts et al. (1997, Sec. A3) for further details on regular variation and max-sum-equivalence.

1.2 The causal tail coefficient

To measure the causal effects in the extremes, we define the following parameter.

Definition 1.1. *Given two random variables X_1 and X_2, we define the* causal tail coefficient

$$\Gamma_{jk} = \lim_{u \to 1^-} \mathbb{E}\left[F_k(X_k) \mid F_j(X_j) > u\right], \tag{1.2.1}$$

if the limit exists, for $j, k = 1, 2$ and $j \neq k$.

The coefficient Γ_{jk} lies between zero and one and is invariant under any marginal strictly increasing transformation since it depends on the rescaled margins $F_j(X_j)$, for $j = 1, 2$. Below, we lay down the setup.

1.2.1 Setup

Consider a linear structural causal model (SCM) with an induced directed acyclic graph (DAG) G,

$$X_j := \sum_{k \in \mathrm{pa}(j,G)} \beta_{jk} X_k + \varepsilon_j, \quad j \in V,$$

where we assume that the coefficients β_{jk} are strictly positive, $j, k \in V$. Let the independent noise variables $\varepsilon_1, \ldots, \varepsilon_p$ be real-valued and regularly varying with comparable tails, i.e., there exists a tail-index $\alpha > 0$ and $\ell \in \mathrm{RV}_0$ such that for all $j \in V$, there exists $c_j > 0$ that satisfies

$$\mathbb{P}(\varepsilon_j > x) \sim c_j \ell(x) x^{-\alpha}, \quad x \to \infty. \tag{1.2.2}$$

To simplify the notation, we rescale the variables X_j such that $c_j = 1$, $j \in V$. Furthermore, denote by $\beta_{k \to j}$ the sum of distinct weighted directed paths from node k to node j, with $\beta_{j \to j} := 1$. Since G is acyclic, we can express recursively each variable X_j, $j \in V$, as a *weighted* sum of the noise terms $\varepsilon_1, \ldots, \varepsilon_k$ that belong to the ancestors of X_j, that is,

$$X_j = \sum_{h \in \mathrm{An}(j,G)} \beta_{h \to j} \varepsilon_h. \tag{1.2.3}$$

The noise terms in (1.2.3) are independent and regularly varying with comparable tails as in (1.2.2). Therefore, by using Lemma A.1 of Appendix A.1, we can write

$$\mathbb{P}(X_j > x) \sim \sum_{h \in \mathrm{An}(j,G)} \beta_{h \to j}^{\alpha} \ell(x) x^{-\alpha}, \quad x \to \infty.$$

Any probability distribution induced by an SCM is Markov with respect to the induced DAG G, and thus we can read off statistical independencies from it by d-separation (Lauritzen et al., 1990; Pearl, 2009b, Sec. 1.2.3). Conversely, to infer dependencies directly from the graph, one needs to assume that the distribution is *faithful* to the DAG G (see Spirtes et al., 2000, Sec. 2.3.3). Most causal methods based on restricted SCMs, such as LiNGAM (see Shimizu et al., 2006), RESIT (Peters et al., 2014), and Peters and Bühlmann (2014), do not assume faithfulness. Similarly, in this work we require the milder assumption that $\beta_{j\to k}$ is non-zero if j is an ancestor of k, i.e., X_j causes X_k. This is automatically satisfied if the SCM has positive coefficients. In the sequel, we refer to this model as a *heavy-tailed linear SCM*.

1.2.2 Causal structure and the causal tail coefficient

In the setting of Section 1.2.1, the causal tail coefficient always exists and carries information about the underlying causal structure. In particular, it can be expressed in closed form.

Lemma 1.2. *Consider a heavy-tailed linear SCM over p variables. Then, for $j, k \in V$ and $j \neq k$,*

$$\Gamma_{jk} = \frac{1}{2} + \frac{1}{2}\frac{\sum_{h\in A_{jk}} \beta_{h\to j}^{\alpha}}{\sum_{h\in \mathrm{An}(j,G)} \beta_{h\to j}^{\alpha}},$$

where $A_{jk} = \mathrm{An}(j,G) \cap \mathrm{An}(k,G)$, and the sum over an empty index set equals zero.

For a proof see Appendix A.2.1. Lemma 1.2 provides a closed form expression for the causal tail coefficient Γ_{jk}, which can be written as a sum of two terms. The first term corresponds to the case when X_j and X_k are independent. The second term is non-negative and depends on the coefficients of the SCM and the tail index $\alpha > 0$. By using matrix notation, it is possible to express the coefficient Γ_{jk} more compactly. Consider the matrix of coefficients $\mathbf{B}$ of the DAG G, where $\mathbf{B}_{jk} = \beta_{jk}$, $j, k \in V$, and let $\mathbf{I}$ be the identity matrix. Furthermore, for any $\mathbf{M} \in \mathbb{R}^{p\times p}$, denote by $\mathbf{M}_\alpha$ the matrix where each entry of $\mathbf{M}$ is raised to the power α. By applying the Neumann series, we obtain $\mathbf{H} = (\mathbf{I} - \mathbf{B})^{-1}$ where $\mathbf{H}_{jk} = \beta_{k\to j}$ for $j, k \in V$. Therefore, for $j, k \in V$ and $j \neq k$, we can write the causal tail coefficient as

$$\Gamma_{jk} = \frac{1}{2} + \frac{1}{2}\frac{e_j^T \mathbf{H}_\alpha e_{A_{jk}}}{e_j^T \mathbf{H}_\alpha e_{\mathrm{An}(j,G)}}, \tag{1.2.4}$$

where $e_j \in \mathbb{R}^p$ is the j-th standard basis vector, and $e_C = \sum_{j\in C} e_j \in \mathbb{R}^p$ for any set $C \subseteq \{1, \dots, p\}$.

Example 1.1. Consider the "diamond" graph $G = (V, E)$ in Figure 1.2, with $V = \{1, \dots, 4\}$. In this graph, for instance, it is easy to see that $\Gamma_{14} = 1$. To compute Γ_{41}, we list the weighted directed paths from the ancestors of node 4 to node 4 itself, i.e.,

$$\beta_{1\to 4} = \beta_{42}\beta_{21} + \beta_{43}\beta_{31},\ \beta_{2\to 4} = \beta_{42},\ \beta_{3\to 4} = \beta_{43},\ \beta_{4\to 4} = 1.$$

Additionally, the set of common ancestors of node 1 and 4 is $A_{14} = \{1\}$. Putting everything together, by using Lemma 1.2, or formula (1.2.4), we obtain

$$\Gamma_{41} = \frac{1}{2} + \frac{1}{2}\frac{\beta_{1\to 4}^{\alpha}}{\beta_{1\to 4}^{\alpha} + \beta_{2\to 4}^{\alpha} + \beta_{3\to 4}^{\alpha} + \beta_{4\to 4}^{\alpha}} < 1.$$

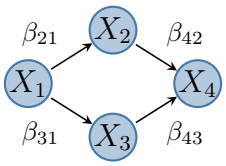


Figure 1.2: Graphical representation of an SCM with an underlying "diamond" DAG G.

In this example, we see that it is possible to infer the causal relationship between X_1 and X_4 because $\Gamma_{14} > \Gamma_{41}$. ◁

Consider now a general, heavy-tailed linear SCM over p variables including X_1 and X_2, and inducing graph G. The following theorem shows that the *causal tail coefficient*, which is computable from the bivariate distribution of X_1 and X_2, see Equation (1.2.1), encodes the causal relationship between the two variables.

Theorem 1.3. *Consider a heavy-tailed linear SCM over p variables including X_1 and X_2, as defined in Section 1.2.1. Then, knowledge of Γ_{12} and Γ_{21} allows us to distinguish the following cases: (a) X_1 causes X_2, (b) X_2 causes X_1, (c) there is no causal link between X_1 and X_2 (i.e., $\mathrm{An}(1, G) \cap \mathrm{An}(2, G) = \varnothing$), (d) there is a node $j \notin \{1, 2\}$, such that X_j is a common cause of X_1 and X_2 and neither X_1 causes X_2 nor X_2 causes X_1. The corresponding values for Γ_{12} and Γ_{21} are depicted in Table 1.2.1.*

Table 1.2.1: Summary of the possible values of Γ_{12} and Γ_{21} and the implications for causality.

	$\Gamma_{21} = 1$	$\Gamma_{21} \in (1/2, 1)$	$\Gamma_{21} = 1/2$
$\Gamma_{12} = 1$		(*a*) X_1 causes X_2	
$\Gamma_{12} \in (1/2, 1)$	(*b*) X_2 causes X_1	(*d*) common cause	
$\Gamma_{12} = 1/2$			(*c*) no causal link

For a proof see Appendix A.2.2. This result will also play a key role when estimating causal relationships from finitely many data. As a first remark, condition (a) and (b) might also include the presence of a common cause X_j. As a second remark, the empty entries in Table 1.2.1 cannot occur under the assumptions made in Section 1.2.1. For example, $\Gamma_{12} = \Gamma_{21} = 1$ can only happen if some variables have different tail indices. One possibility is when the *cause* has a heavier tail than the *effect*. Another scenario is when a common cause X_j, for some $j \neq 1, 2$, has heavier tails than the confounded variables X_1 and X_2. For further discussion on different tail indices see Section 1.4.3.

1.2.3 A non-parametric estimator

Consider a heavy-tailed linear SCM over p variables including X_1 and X_2, with distributions F_1 and F_2, as described in Section 1.2.1. In order to construct a non-parametric estimator of Γ_{12} and Γ_{21} based on independent observations (X_{i1}, X_{i2}), $i = 1, \ldots, n$, of (X_1, X_2), we define the empirical distribution function of X_j as

$$\widehat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{X_{ij} \leq x\right\}, \quad x \in \mathbb{R}, \tag{1.2.5}$$

for $j = 1, 2$. Denote by $g^{\leftarrow}$ the left continuous generalised inverse

$$g^{\leftarrow}(y) = \inf\{x \in \mathbb{R} : \ g(x) \geq y\}, \quad y \in \mathbb{R}.$$

In addition, let the $(n-k)$-th order statistics be denoted by $X_{(n-k),j} = \widehat{F}_j^{\leftarrow}(1-k/n)$, for all $k = 0, \ldots, n-1$ and $j = 1, 2$, such that $X_{(1),j} \leq \cdots \leq X_{(n),j}$. Replacing F_1 and F_2 in the definition of Γ_{12} in (1.2.1) by the empirical counterparts, and the threshold u by $u_n = 1 - k/n$, for some integer $0 < k \leq n-1$, we define the estimator

$$\widehat{\Gamma}_{12} = \widehat{\Gamma}_{12}^{(n)} = \frac{1}{k} \sum_{i=1}^{n} \widehat{F}_2(X_{i2}) \mathbf{1}\{X_{i1} > X_{(n-k),1}\}. \tag{1.2.6}$$

For this estimator to be consistent, a classical assumption in extreme value theory is that the number of upper order statistics $k = k_n$ depends on the sample size n such that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. The first condition is needed to increase the effective sample size, whereas the second condition eliminates the approximation bias. The estimator $\widehat{\Gamma}_{21} = \widehat{\Gamma}_{21}^{(n)}$ is defined in an analogous way as (1.2.6).

Theorem 1.4. *Let X_{i1} and X_{i2}, $i = 1, \ldots, n$, be independent copies of X_1 and X_2, respectively, where X_1 and X_2 are two of the p variables of a heavy-tailed linear SCM.*

(A1) Assume that the density functions $f_j = F_j'$, $j = 1, 2$, exist and satisfy the von Mises' condition

$$\lim_{x \to \infty} \frac{x f_j(x)}{1 - F_j(x)} = \frac{1}{\gamma}, \quad \text{for some } \gamma > 0. \tag{1.2.7}$$

(A2) Let $k_n \in \mathbb{N}$ be an intermediate sequence with

$$k_n \to \infty \quad \text{and} \quad k_n/n \to 0, \quad n \to \infty.$$

Then the estimators $\widehat{\Gamma}_{12}$ and $\widehat{\Gamma}_{21}$ are consistent, as $n \to \infty$, i.e.,

$$\widehat{\Gamma}_{12} \xrightarrow{P} \Gamma_{12} \text{ and } \widehat{\Gamma}_{21} \xrightarrow{P} \Gamma_{21}.$$

Remark 1.1. The von Mises' condition in (A1) is a very mild assumption that is satisfied by most univariate regularly varying distributions of interest. In our case $\gamma = 1/\alpha$, where α is the common tail index of the noise variables. ◁

For a proof of Theorem 1.4 see Appendix A.2.3. It uses several results from tail empirical process theory (e.g., de Haan and Ferreira, 2006, Sec. 2.2). The main challenge comes from the fact that the variables X_1 and X_2 are tail dependent, and that the use of the empirical distribution function $\widehat{F}_2$ in (1.2.6) introduces dependence between the terms corresponding to different observations $i = 1, \ldots, n$. A related problem is studied in Cai et al. (2015), where they derive asymptotic properties of the empirical estimator of the expected shortfall when another dependent variable is extreme. However, in contrast to Cai et al. (2015), in the proof of Theorem 1.4, we work with a more explicit model and we consider the variables scaled to uniform margins, i.e., $\widehat{F}_j(X_j)$ instead of X_j, $j = 1, 2$.

1.3 Causal discovery using extremes

We would like to recover the causal information from a dataset of p variables under the model specification of Section 1.2.1. We develop an algorithm named *extremal ancestral search* (EASE) based on the *causal tail coefficient* defined in (1.2.1). We show that EASE can recover the causal order of the underlying graph in the population case (Section 1.3.1), and that it is consistent (Section 1.3.2).

1.3.1 Learning the causal order

Our goal is to recover the causal order of a heavy-tailed linear SCM over p variables (as defined in Section 1.2.1) by observing n i.i.d. copies of the random vector $X \in \mathbb{R}^p$. Given a DAG $G = (V, E)$, a permutation $\pi : \{1, \dots, p\} \to \{1, \dots, p\}$ is said to be a *causal order* (or *topological order*) of G if $\pi(i) < \pi(j)$ for all i and j such that $i \in \mathrm{an}(j, G)$. We denote by Π_G the set of all causal orders of G. For a permutation π we sometimes use the notation $\pi = (\pi(1), \dots, \pi(p))$.

A given causal order π does not specify a unique DAG. As an example, the causal order $\pi = (1, 2)$ comprises two DAGs: one where there is a directed edge between node 1 and node 2, and one where the two nodes are unconnected. On the other hand, there can be several causal orders for a given DAG. For example, a fully-disconnected DAG satisfies any causal order. However, even if the causal order does not identify a unique DAG, it still conveys important information. In particular, each causal order defines a class of DAGs that agree with respect to the non-ancestral relations. Therefore, once a causal order is available, one can estimate the complete DAG by using regularised regression methods. This idea has been exploited, e.g., by Shimizu et al. (2011) and Bühlmann et al. (2014). In addition, Bühlmann et al. (2014) and Peters and Bühlmann (2015) argue that knowledge of a causal order is useful *per se*. In fact, given a correct causal order, one can construct a fully-connected DAG that describes interventional distribution across the variables.

For any heavy-tailed linear SCM and induced DAG $G = (V, E)$, we define the matrix $\Gamma \in \mathbb{R}^{p \times p}$ with entries Γ_{ij}, the causal tail coefficients between all pairs of variables X_i and X_j, $i, j \in V$; see Definition 1.1. Theorem 1.3 tells us how the entries of Γ encode the causal relationships between the random variables of the SCM. To recover the causal order of the DAG G, we propose Algorithm 1, named *extremal ancestral search* (EASE).

Algorithm 1 is a greedy algorithm that identifies root nodes of the current subgraph at each step. In the first step, the algorithm finds a root node $i_1 \in V$ as the one that minimises the score $M_i^{(1)} = \max_{j \neq i} \Gamma_{ji}$, $i \in V$. In fact, by Theorem 1.3, $M_i^{(1)} < 1$ if and only if i is a source node. Once the first node is selected, the algorithm searches for a second root node in the subgraph where i_1 is removed. The procedure continues until all nodes have been selected. In Appendix A.3, one example illustrates how EASE finds a causal order for a given DAG.

The next result states that, in the population case, the EASE algorithm yields a correct causal order of the underlying DAG.

Proposition 1.5. *Consider a heavy-tailed linear SCM over p variables, as defined in Section 1.2.1, and let $G = (V, E)$ be the induced DAG. If the input Γ is the matrix of causal tail coefficients associated with the SCM, then EASE returns a permutation π that is a causal order of G.*

For a proof see Appendix A.2.4.

Algorithm 1 Extremal ancestral search (EASE)

INPUT: A matrix $\Gamma \in \mathbb{R}^{p\times p}$ of causal tail coefficients related to a DAG $G = (V, E)$ with $V = \{1, \dots, p\}$.
RETURNS: Permutation of the nodes $\pi : V \to \{1, \dots, p\}$.

0. Set $V_1 = V$.
1. FOR $s \in \{1, \dots, p\}$
   (a) Let $M_i^{(s)} = \max_{j\in V_s\setminus\{i\}} \Gamma_{ji}$, for all $i \in V_s$.
   (b) Let $i_s \in \arg\min_{i\in V_s} M_i^{(s)}$.
   (c) Set $\pi(i_s) = s$.
   (d) Set $V_{s+1} = V_s \setminus \{i_s\}$.
2. RETURN the permutation π.

COMPLEXITY: $O(p^2)$.

1.3.2 Sample properties for the EASE algorithm

For finite samples, the EASE algorithm will take an estimate of the causal coefficient matrix Γ as input. Based on the empirical non-parametric estimator $\widehat{\Gamma}$ and its asymptotic properties, we assess the performance of the algorithm. Let $\widehat{\Gamma} \in \mathbb{R}^{p\times p}$ denote the matrix where each entry $\widehat{\Gamma}_{ij}$ is defined as in (1.2.6) in Section 1.2.3, for $i, j \in V$. We say that a procedure makes a mistake when it returns a permutation $\pi \notin \Pi_G$. We derive an upper bound for the probability that EASE makes a mistake when the matrix Γ is estimated by $\widehat{\Gamma}$.

Proposition 1.6. *Consider a heavy-tailed linear SCM over p variables $X = (X_1, \dots, X_p)$, as defined in Section 1.2.1, with induced DAG G. Let $\widehat{\Gamma}$ be the estimated causal coefficient matrix related to G. Let $\widehat{\pi}$ denote the permutation returned by EASE based on $\widehat{\Gamma}$. Then,*

$$\mathbb{P}\left(\widehat{\pi} \notin \Pi_G\right) \leq p^2 \max_{i,j\in V: i\neq j} \mathbb{P}\left(\left|\widehat{\Gamma}_{ij} - \Gamma_{ij}\right| > \frac{1-\eta}{2}\right),$$

where $\eta = \max_{u\notin \mathrm{An}(v,G)} \Gamma_{uv} < 1$.

For a proof see Appendix A.2.5. The bound for the probability of making a mistake in the estimated causal order is expressed in terms of the distance between the true Γ_{ij} and the estimated $\widehat{\Gamma}_{ij}$. This bound in combination with the consistency result of Theorem 1.4 yields the consistency of the EASE algorithm in the sample case.

Corollary 1.7. *Let $\widehat{\pi}$ be the permutation computed by EASE under the assumptions of Proposition 1.6. Let $k_n \in \mathbb{N}$ be an intermediate sequence with*

$$k_n \to \infty \quad \text{and} \quad k_n/n \to 0, \quad n \to \infty.$$

If the von Mises' condition (1.2.7) *holds, then the EASE algorithm is consistent, i.e.,*

$$\mathbb{P}\left(\widehat{\pi} \notin \Pi_G\right) \to 0, \qquad \text{as } n \to \infty.$$

The result above is for fixed dimension p. To prove consistency in a regime where p scales with the sample size n we would need to establish concentration inequalities for $\widehat{\Gamma}$ or asymptotic normality in Theorem 1.4. Both would require stronger assumptions on the tails of the noise variables and a second-order analysis in line with the proof of Theorem 1.4.

1.3.3 Computational complexity

The EASE algorithm is based on pairwise quantities and is therefore computationally efficient. To estimate the matrix $\widehat{\Gamma}$ of causal tail coefficients, which is the input for EASE, first, we need to rank the n observations for each of the variables, with a computational complexity of $O(pn \log n)$. Then we compute the coefficients $\widehat{\Gamma}_{ij}$ for each pair $i, j \in V$, with a computational complexity of $O(k_n p^2)$. The computational complexity of EASE grows with the square of the number of variables, i.e., $O(p^2)$. The overall computational complexity of estimating the matrix $\widehat{\Gamma}$ and running the EASE algorithm is therefore $O\left(\max(pn \log n, k_n p^2)\right)$.

1.4 Extensions

1.4.1 Real-valued coefficients

Until now, we have worked with a heavy-tailed linear SCM with positive coefficients (see Section 1.2.1 for a detailed explanation of the model). In the current section, we relax this assumption and let the coefficients of the SCM be real-valued, i.e., $\beta_{jk} \in \mathbb{R}$, $j, k \in V$. Additionally, we assume that $\beta_{j \to k}$ is non-zero, if j is an ancestor of k. Given that the coefficients are real-valued, we want to consider both the upper and the lower tails of the variables. We assume that the noise variables $\varepsilon_1, \ldots, \varepsilon_p$, of the SCM have comparable upper and lower tails, that is, as $x \to \infty$

$$\mathbb{P}(\varepsilon_j > x) \sim c_j^+ \ell(x) x^{-\alpha}, \quad \mathbb{P}(\varepsilon_j < -x) \sim c_j^- \ell(x) x^{-\alpha},$$

where $c_j^+, c_j^- > 0$, $j \in V$ and $\ell \in \mathrm{RV}_0$. Furthermore, we define a causal tail coefficient that is sensitive to both tails as

$$\Psi_{jk} = \lim_{u \to 1^-} \mathbb{E}\left[\sigma(F_k(X_k)) \mid \sigma(F_j(X_j)) > u\right], \quad j, k \in V, \tag{1.4.1}$$

if the limit exists, where $\sigma : x \mapsto |2x - 1|$. The nonlinear transformation $x \mapsto \sigma(x)$, $x \in [0, 1]$, makes the Ψ_{jk} coefficient in (1.4.1) sensitive to both large positive and negative values of X_k, conditional on the event that X_j takes large positive or negative values. Since $F_j(X_j) \sim \mathrm{Unif}[0, 1]$, $j \in V$, we can rewrite (1.4.1) as

$$\begin{aligned} \Psi_{jk} = & \lim_{u \to 1^-} \frac{1}{2} \mathbb{E}\left[\sigma(F_k(X_k)) \mid F_j(X_j) > u\right] \\ & + \lim_{u \to 0^+} \frac{1}{2} \mathbb{E}\left[\sigma(F_k(X_k)) \mid F_j(X_j) < u\right] \\ = & \Psi_{jk}^+ + \Psi_{jk}^-, \end{aligned} \tag{1.4.2}$$

where the first and second terms correspond to the cases where X_j is extremely large and extremely small, respectively.

In the current setting, the coefficient defined in (1.4.1) always exists, and it has a closed form expression that encodes causal relationships between the variables.

Lemma 1.8. *Consider a heavy-tailed SCM over p variables, where the coefficient $\beta_{jk} \in \mathbb{R}$, $j, k \in V$. Assume that $\beta_{j\to k} \neq 0$ if j is an ancestor of k. Then, for $j, k \in V$ and $j \neq k$,*

$$\Psi_{jk} = \frac{1}{2} + \frac{1}{4}\frac{\sum_{h\in A_{jk}} c_{hj}^{+}|\beta_{h\to j}|^{\alpha}}{\sum_{h\in \mathrm{An}(j,G)} c_{hj}^{+}|\beta_{h\to j}|^{\alpha}} + \frac{1}{4}\frac{\sum_{h\in A_{jk}} c_{hj}^{-}|\beta_{h\to j}|^{\alpha}}{\sum_{h\in \mathrm{An}(j,G)} c_{hj}^{-}|\beta_{h\to j}|^{\alpha}},$$

where $A_{jk} = \mathrm{An}(j, G) \cap \mathrm{An}(k, G)$, and

$$c_{hj}^{+} = \begin{cases} c_h^{+}, & \beta_{h\to j} > 0, \\ c_h^{-}, & \beta_{h\to j} < 0, \end{cases} \qquad c_{hj}^{-} = \begin{cases} c_h^{-}, & \beta_{h\to j} > 0, \\ c_h^{+}, & \beta_{h\to j} < 0. \end{cases} \tag{1.4.3}$$

A proof is provided in Appendix A.2.6. The interpretation of the result is as follows. The baseline of the coefficient is 1/2, which can be checked to be the value of Ψ_{jk} when two variables are independent. The other two terms account for the equally weighted contribution from the lower and upper tail, respectively. The result stated in Lemma 1.8 allows us to extend Theorem 1.3 to the more general setting where the heavy-tailed SCM has coefficients $\beta_{jk} \in \mathbb{R}$, $j, k \in V$.

Theorem 1.9. *Consider a heavy-tailed linear SCM over p variables including X_1 and X_2, and assume that $\beta_{jk} \in \mathbb{R}$, $j, k \in V$. In addition, assume that $\beta_{j\to k} \neq 0$ if j is an ancestor of k. Then, knowledge of Ψ_{12} and Ψ_{21} allows us to distinguish the following cases: (a) X_1 causes X_2, (b) X_2 causes X_1, (c) there is no causal link between X_1 and X_2, (d) there is a node $j \notin \{1, 2\}$, such that X_j is a common cause of X_1 and X_2 and neither X_1 causes X_2 nor X_2 causes X_1. The corresponding values for Ψ_{12} and Ψ_{21} are shown in Table 1.4.1.*

Table 1.4.1: Summary of the possible values of Ψ_{12} and Ψ_{21} and the implications for causality.

	$\Psi_{21} = 1$	$\Psi_{21} \in (1/2, 1)$	$\Psi_{21} = 1/2$
$\Psi_{12} = 1$		(a) X_1 causes X_2	
$\Psi_{12} \in (1/2, 1)$	(b) X_2 causes X_1	(d) common cause	
$\Psi_{12} = 1/2$			(c) no causal link

The proof is identical to the proof of Theorem 1.3, replacing Γ_{ij} with Ψ_{ij} and by referring to Lemma 1.8 instead of Lemma 1.2. Moreover, as in Theorem 1.3, condition (a) and (b) can also include the presence of a common cause X_j. Theorem 1.9 implies that if we run the EASE algorithm based on the matrix $\Psi \in \mathbb{R}^{p\times p}$, containing the pairwise Ψ_{ij}, $i, j \in V$, then we retrieve a causal order of the underlying DAG. This is the analogue to Proposition 1.5 for heavy-tailed linear SCM with real-valued coefficients.

We define an empirical estimator $\widehat{\Psi}_{ij}$ of Ψ_{ij} in a similar fashion as the estimator $\widehat{\Gamma}_{ij}$ in (1.2.6). The proof of Lemma 1.8 shows that the coefficient Ψ_{ij} can be decomposed in the same way as Γ_{ij} in the proof of Lemma 1.2. Therefore, following the lines of the proof of Theorem 1.4 with some minor modifications, we obtain the consistency $\widehat{\Psi}_{ij} \xrightarrow{P} \Psi_{ij}$ as $n \to \infty$ for any intermediate sequence $k_n \to \infty$ and $k_n/n \to 0$, and under the assumption of the von Mises' condition for both the upper and the lower tail of X_j, $j \in V$.

We can then estimate a permutation $\widehat{\pi}$ by the EASE algorithm based on the matrix $\widehat{\Psi} \in \mathbb{R}^{p\times p}$ that contains the estimators $\widehat{\Psi}_{ij}$, $i, j \in V$, as entries. For this permutation

we obtain the same bound for the probability that EASE makes a mistake as shown in Proposition 1.6 by replacing $\widehat{\Gamma}_{ij}$ and Γ_{ij} by $\widehat{\Psi}_{ij}$ and Ψ_{ij}, respectively. This together with the consistency of $\widehat{\Psi}_{ij}$ yields the following result.

Corollary 1.10. *Assume the general setup of the heavy-tailed linear SCM with real-valued coefficients of this section. Let $\widehat{\pi}$ be the permutation computed by EASE based on the matrix $\widehat{\Psi}$. Assume the von Mises' condition for the upper and the lower tail of X_j and let $k_n \in \mathbb{N}$ be an intermediate sequence with*

$$k_n \to \infty \quad \textit{and} \quad k_n/n \to 0, \quad n \to \infty.$$

Then, the EASE algorithm is consistent, i.e.,

$$\mathbb{P}\left(\widehat{\pi} \notin \Pi_G\right) \to 0, \qquad \textit{as } n \to \infty.$$

1.4.2 Presence of hidden confounders

A frequent assumption in causality is that one can observe all the relevant variables. However, in many real-world situations, it is hard, if not impossible, to do so. When some of the hidden variables are confounders (i.e., common causes), the causal inference process might be compromised. Therefore, an attractive property of a causal inference algorithm involves its robustness to hidden confounders. In this section, we show that EASE is capable of dealing with hidden common causes and, under certain assumptions, it recovers the causal order of the observed graph both in the population and in the asymptotic case.

Consider a heavy-tailed linear SCM with *real-valued* coefficients, as defined in Section 1.4.1, consisting of both observed and hidden variables. This SCM induces a DAG $G = (V, E)$, with $V = V_O \cup V_H$, $V_O \cap V_H = \varnothing$, where V_O (V_H) denotes the set of nodes corresponding to the observed (hidden) variables. Our goal is to recover a causal order for the subset of the observed variables X_j, $j \in V_O$. In particular, we say that the EASE algorithm recovers a causal order π over the observed variables if

$$\pi(i) < \pi(j) \implies j \notin \mathrm{an}(i, G), \quad \text{for all } i, j \in V_O. \tag{1.4.4}$$

In fact, the results of the previous sections hold even in the presence of hidden confounders.

Regarding the population properties, Theorem 1.3 and 1.9 still apply: they state that the causal tail coefficients Γ and Ψ reflect the causal relationships between pairs of variables without taking into account other variables, e.g., by conditioning. In addition, the result of Proposition 1.5, and the corresponding extension in Section 1.4.1, are also valid. The proof of Proposition 1.5 depends only on the assumption that the input matrix contains the pairwise causal effects between the variables. Therefore, if we use matrix Γ (or Ψ) as input for the EASE algorithm, we recover a causal order π that satisfies (1.4.4).

Regarding the asymptotic properties, $\hat{\Gamma}$ and $\hat{\Psi}$ are consistent even in the presence of hidden common causes: in the proof of Theorem 1.4, the other variables do not appear. In addition, we can still find an upper bound for the probability that the EASE algorithm makes a mistake. To do so, one needs to adjust the proof of Proposition 1.6 by replacing the full DAG G with the subgraph $G_O = (V_O, E_O)$ containing only the observed variables, where $E_O = E \cap (V_O \times V_O)$. Combining the two previous arguments, it follows that Corollary 1.7 and 1.10 hold even in the presence of hidden confounders.

The ability to deal with hidden confounders is a property that, in general, is not shared by all methods in causality. For example, the PC algorithm (Spirtes et al., 2000, Sec. 5.4.2) might retrieve a Markov equivalence class that contains DAGs with a wrong causal order if some of the variables are not included in the analysis. Similarly, the standard version of the LiNGAM algorithm (Shimizu et al., 2006) might produce a wrong DAG in the presence of hidden common causes. Hoyer et al. (2008), Entner and Hoyer (2010), and Tashiro et al. (2014) proposed extensions of LiNGAM that deal with hidden variables. While Entner and Hoyer (2010), and Tashiro et al. (2014) show good performance in practice, all three methods suffer from some drawbacks. For example, the LiNGAM version of Hoyer et al. (2008) requires *a priori* the number of hidden variables in the SCM (or needs to estimate it from data). The main limitation of Entner and Hoyer (2010) is that it recovers causal information only for subsets of variables that are not affected by hidden confounders. For some non-ancestral graphs, the method by Tashiro et al. (2014) does not identify all ancestral relationships (Wang and Drton, 2020). In addition, both the work of Hoyer et al. (2008) and Tashiro et al. (2014) are computationally intensive, with the latter showing a computational time that grows exponentially with the sample size and the number of observed variables. Among the constraint-based methods, Spirtes et al. (2000, Sec. 6.7) proposed the FCI method, which is an extension to the PC algorithm that deals with arbitrarily many hidden confounders and produces a partial ancestral graph (see Zhang, 2008). Due to the high number of independence tests, the FCI algorithm can be slow when the number of variables is large. For this reason, Claassen et al. (2013) proposed the FCI+ algorithm, a faster version of FCI that is consistent in sparse high-dimensional settings with arbitrarily many hidden variables. In general, FCI type algorithms produce an equivalence class of graphs. They are not guaranteed to recover the causal order of the variables.

Compared to the methods mentioned above, our algorithm has the advantage of being computationally fast, and being able to produce a causal order without assumptions on the number of hidden variables and the sparsity of the true underlying DAG.

1.4.3 Noise variables with different tails

We have so far considered the case where the noise variables of a given SCM share the same tail coefficient $\alpha > 0$ and the same slowly varying function ℓ. Consider now a heavy-tailed SCM over p variables, as defined in Section 1.2.1, with the difference that the noise variables have possibly different tail indices $\alpha_1, \dots, \alpha_p > 0$ and slowly varying functions $\ell_1, \dots, \ell_p \in \mathrm{RV}_0$, i.e., for $j = 1, \dots, p$,

$$\mathbb{P}(\varepsilon_j > x) \sim \ell_j(x) x^{-\alpha_j}, \quad x \to \infty.$$

We say that ε_j has heavier (upper) tail than ε_k if either $0 < \alpha_j < \alpha_k$, or $\alpha_j = \alpha_k$ and $\ell_j(x)/\ell_k(x) \to \infty$ as $x \to \infty$. Denote by $G = (V, E)$ the DAG induced by the SCM. With similar arguments to the proof of Lemma 1.2, the causal tail coefficient for $j, k \in V$ can then be expressed as

$$\Gamma_{jk} = \frac{1}{2} + \frac{1}{2} \lim_{x \to \infty} \frac{\sum_{h \in A_{jk}} \beta_{h \to j}^{\alpha_h} \mathbb{P}(\varepsilon_h > x)}{\sum_{h \in \mathrm{An}(j,G)} \beta_{h \to j}^{\alpha_h} \mathbb{P}(\varepsilon_h > x)}, \tag{1.4.5}$$

where $A_{jk} = \mathrm{An}(j, G) \cap \mathrm{An}(k, G)$, and the sum over an empty index set equals zero. From (1.4.5), we can study the different constellations of X_j and X_k and the corresponding values of Γ_{jk}. We obtain the following three statements.

1. If X_j and X_k are independent, the causal tail coefficient satisfies, as before, $\Gamma_{jk} = 1/2$.

2. If X_j is an ancestor of X_k then, as before, $\Gamma_{jk} = 1$.

3. In all other scenarios, it holds that $\Gamma_{jk} < 1$ as long as the noise variables ε_h, $h \in A_{jk}$, of the common ancestors of X_j and X_k have tails that are lighter than (or as light as) the one of ε_j. On the other hand, if there is some common ancestor of X_j and X_k for which the noise variable's tail is heavier than the one of ε_j, then $\Gamma_{jk} = 1$.

These statements help to understand in which cases the values of Γ_{jk} indicate a correct causal relation.

Example 1.2. Suppose that X_j is an ancestor of X_k, and there is possibly a common ancestor X_0 (which can also be a hidden confounder). Since $\Gamma_{jk} = 1$, we will never mistakenly detect the existence of a causal effect from X_k to X_j. If either X_j or X_0 has a heavier tail than X_k, then $\Gamma_{kj} = 1$ and we cannot detect the causal effect from X_j to X_k. ◁

Example 1.3. Suppose neither X_j causes X_k nor X_k causes X_j, and X_0 is a common ancestor of X_j and X_k. If X_0 has a tail that is lighter than (or as light as) the one of X_k, but heavier than the one of X_j, then $\Gamma_{jk} = 1 > \Gamma_{kj}$. Therefore, the causal tail coefficient indicates a wrong causal effect from X_j to X_k. ◁

Whenever there exists a causal effect between two variables, we can, at worst, fail to detect it (that is, the causal tail coefficient does not indicate a causal effect in the wrong direction). When there is no causal connection between two variables, the causal tail coefficient might indicate a wrong causal effect. However, this does not affect the correctness of the EASE algorithm, on the population level. Indeed, if $\Gamma_{jk} = 1 > \Gamma_{kj}$, for $j, k \in V$, there are two possibilities. If X_j is an ancestor of X_k, then the algorithm correctly chooses j before k. If X_j and X_k share a common ancestor, but none of them is causing the other (see Example 1.3), then any permutation of j and k yields a valid causal order. Example 1.2 shows that EASE could make mistakes when X_j is an ancestor of X_k and $\Gamma_{jk} = \Gamma_{kj} = 1$, since the causal tail coefficient does not indicate any causal effect. In this case, one could remove one variable at a time to obtain a subset $\tilde{V} \subset V$ that satisfies $\Gamma_{hm} < 1$ or $\Gamma_{mh} < 1$ for each $h, m \in \tilde{V}$. By applying the EASE algorithm to the subset of the remaining variables, one would recover a correct causal order on such subset.

For simplicity we only considered the causal tail coefficient Γ, but similar conclusions hold for Ψ.

1.5 Numerical results

1.5.1 Simulation study

We assess the performance of EASE in estimating a causal order of a graph induced by a heavy-tailed SCM. We simulate the SCMs with real-valued coefficients and different numbers of variables p and samples n. The noise variables have Student's t distributions with different degrees of freedom α and we consider four different settings, including unobserved confounders and model misspecification; see Appendix A.4 for details. Since the coefficients in the SCM are real-valued, we use the causal tail coefficient Ψ defined

in Section 1.4.1 for our EASE algorithm. Our code is available as an `R` package at https://github.com/nicolagnecco/causalXtreme. Scripts generating all our figures and results can be found at the same url.

Competing methods and evaluation metric We compare our algorithm to three well-established methods in causality, the Rank PC algorithm (Harris and Drton, 2013), ICA-LiNGAM (Shimizu et al., 2006), and Pairwise LiNGAM (Hyvärinen and Smith, 2013).

The classic PC algorithm (Spirtes et al., 2000, Sec. 5.4.2) belongs to the class of constraint-based methods for causal discovery. It estimates the Markov equivalence class of a DAG, encoded as a completed partially directed acyclic graph (CPDAG). The PC algorithm retrieves a CPDAG by performing conditional independence tests on the variables. The Rank PC algorithm, proposed by Harris and Drton (2013), is an extension of the PC algorithm and uses the rank-based Spearman correlation to perform the independence tests. This modification ensures that the method is more robust to non-Gaussian data.

The algorithms that fit our problem best are ICA-LiNGAM and Pairwise LiNGAM. ICA-LiNGAM, proposed by Shimizu et al. (2006), leverages the results of independent component analysis (ICA) (Comon, 1994) to estimate the DAG of a linear SCM under the only assumption that the noise is non-Gaussian. Pairwise LiNGAM, proposed by Hyvärinen and Smith (2013), is a likelihood-ratio-based method to identify the exogenous variables within the DirectLiNGAM framework. DirectLiNGAM, introduced by Shimizu et al. (2011), is an algorithm based on two iterative steps, namely, finding an exogenous variable (i.e., a node in the DAG with no parents), and regressing this variable out of all the others. In this simulation study, we let ICA-LiNGAM and Pairwise LiNGAM return only a causal order (and not a complete DAG structure), in order to make a fair comparison with EASE.

The algorithms return different types of causal information. On the one hand, EASE, ICA-LiNGAM, and Pairwise LiNGAM estimate a causal order. On the other hand, the Rank PC algorithm computes a CPDAG that represents a Markov equivalence class of DAGs. Therefore, when it comes to evaluating the performance of the algorithms, it becomes crucial to use a measure that is meaningful for all of them. We choose the structural intervention distance (SID) proposed by Peters and Bühlmann (2015). The SID takes as input either a pair of DAGs or a DAG and a CPDAG and returns the number of falsely inferred interventional distributions (Peters and Bühlmann, 2015, Definition 3). We standardise the SID to lie between zero and one. For each method, we compute the distance between the simulated DAG, i.e., the ground truth, and the estimated DAG or CPDAG. An estimated causal order $\widehat{\pi}$ corresponds to a fully connected DAG $G = (V, E)$, where $(i, j) \in E$ if $\widehat{\pi}(i) < \widehat{\pi}(j)$. As a caveat, we slightly adapt the SID in the case of hidden confounders (see Setting 2 of our simulations), since it is not designed to work in such a situation.

Results In this simulation experiment we use the implementation of the Rank PC, and ICA-LiNGAM algorithm developed by Kalisch et al. (2012). We implemented Pairwise LiNGAM in `C++` and included it in our software package.

Regarding the hyperparameter settings, for the Rank PC algorithm, we perform a conditional independence test based on Spearman's correlation coefficient, as proposed by Harris and Drton (2013), and we set the level of the independence tests to 0.0005.

Concerning the choice of the number of exceedances k_n in the EASE algorithm, we perform a small preliminary simulation. Figure A.3 shows the SID of EASE for $k_n = \lfloor n^\nu \rfloor$ and different fractional exponents $\nu > 0$. The best fractional exponent in Figure A.3 seems to depend on the tail heaviness of the noise variables, and in particular it appears to be smaller for larger values of the degree of freedom α of the Student's t distribution. Our estimators $\widehat{\Gamma}_{ij}$ and $\widehat{\Psi}_{ij}$ are similar in construction to Hill's estimator (Hill, 1975). For the latter, the optimal number k_n^* of exceedances depends on the tail index and an index related to a second-order condition; see Section 3.2 in de Haan and Ferreira (2006) for details. For the Student's t distribution with α degrees of freedom it can be shown that $k_n^* \sim C_\alpha n^{1/(\alpha+1)}$, where $C_\alpha > 0$ is a constant. This intuitive explanation coincides well with the optimal fractional exponents in Figure A.3. In the sequel, we choose $k_n = \lfloor n^{0.4} \rfloor$ because it lies within the best range for the fractional exponent. This result also agrees with the assumptions of Theorem 1.4, where $k_n \to \infty$ and $k_n/n \to 0$, as $n \to \infty$.

Regarding the simulation settings, we let n denote the number of observations, p the number of variables, and $\alpha > 0$ the tail index of the simulated distribution. For each combination of $n \in \{500, 1000, 10000\}$, $p \in \{4, 7, 10, 15, 20, 30, 50\}$ and $\alpha \in \{1.5, 2.5, 3.5\}$ we simulate 50 random SCMs under four different settings. The simulated data is independent of the data used to choose the best fractional exponent of k_n (see Figure A.3). The first setting corresponds to linear SCMs with real-valued coefficients described in Section 1.4.1. In the second setting, we introduce hidden confounders. The third setting corresponds to nonlinear SCMs. In the fourth setting, we first generate linear SCMs and then transform each variable to uniform margins. Further details on the generation of the SCMs are in Appendix A.4. For each simulation and setting we evaluate the performance of EASE, ICA-LiNGAM, Pairwise LiNGAM, and Rank PC algorithm with the SID. As a baseline, in each simulation, we also compute the SID of a randomly generated DAG, where we randomly choose the causal order, the sparsity and the edges of the graph.

Figure 1.3 displays the results of the simulations when the tail index $\alpha = 1.5$. We can observe that EASE is quite robust across the four different settings. We explain this finding as follows. In the presence of hidden confounders (Setting 2), EASE can retrieve a correct causal order, asymptotically. Furthermore, the nonlinear setting used in this simulation (Setting 3) is such that the relationships between the variables are kept linear in the tails. Therefore, our algorithm is only moderately affected by this model misspecification. Finally, EASE is not affected by the transformation to uniform margins (Setting 4) because the causal tail coefficient Ψ is invariant under any strictly monotone increasing transformation.

Compared to the other methods, we observe that EASE performs better than Rank PC across all settings, and better than ICA-LiNGAM in Setting 2 and 4. Pairwise LiNGAM is overall the best performing method, except in Setting 4. Also, Pairwise LiNGAM is less affected by misspecifications in the bulk of the data distribution (Setting 3), compared to ICA-LiNGAM. One reason is that Pairwise LiNGAM relies on ordinary least square regression that is sensitive to high-leverage points. In this particular setting, ICA-LiNGAM and Pairwise LiNGAM gain in robustness if we discard the data in the bulk of the distribution; see Table A.5.1 in Appendix A.5. Furthermore, both ICA-LiNGAM and Pairwise LiNGAM are the algorithms with the best convergence for high dimension p, as n increases. This result is not surprising because EASE uses only the $k_n < n$ upper order statistics to recover the causal structure. In addition, we notice that Pairwise LiNGAM outperforms ICA-LiNGAM, in agreement to the findings of Hyvärinen and Smith (2013). Regarding the Rank PC algorithm, we can see that it is quite stable under different set-

tings, but it performs only marginally better than the random method. Moreover, in Figure A.10 in Appendix A.5, we observe that the performance of the Rank PC algorithm is almost constant for significance levels of the independence test between $5 \cdot 10^{-4}$ and 0.5. The results do not change when we consider the standard PC algorithm, which is based on partial correlation as a conditional independence test. The results for tail indices $\alpha = 2.5, 3.5$ are in Figures A.6 and A.7 in Appendix A.5. Increasing values of α correspond to lighter tails, and we observe that it becomes more challenging for EASE to recover the correct causal order. In the extreme case where $alpha \to \infty$, the variables are asymptotically independent and the causal tail coefficient does not identify the causal direction anymore; this leads the EASE algorithm to fail in recovering a valid causal order.

In addition to the competitive performance in the simulations, a further advantage of EASE is its computational efficiency. The algorithm performs computations only on the tails of the dataset and relies on simple non-parametric estimators of the causal tail coefficient; see Section 1.3.3. Figure A.9 in Appendix A.5 shows that EASE can be up to two orders of magnitude faster than the other methods.

1.5.2 Financial application

In general, one cannot easily reason about causality in financial markets. Several factors influence financial returns, and most of them are unobserved. In addition, the effect of these factors varies in time. However, under particular circumstances, it is possible to conjecture the existence of a specific causal relationship, with a reasonable degree of confidence. For example, in the Swiss financial market, one can argue that very large (both positive and negative) changes to the Euro Swiss franc exchange rate (EURCHF) induce changes in the Swiss Market Index (SMI), the main stock index in Switzerland. This is due to multiple reasons such as the multinational nature and the high export dependency of the Swiss economy. Consider, for instance, the decision of the Swiss National Bank (SNB) to discontinue the minimum exchange rate between Swiss franc and Euro, on January 15, 2015. This event can be deemed as a *large intervention* with a plausible causal interpretation (in the spirit of Cox and Wermuth (1996, Sec. 8.7)). Following the SNB decision, the EURCHF plummeted more than 30 standard deviations, and all the stocks included in the SMI dropped in value on the same day.

For this reason, we consider the returns of the Euro Swiss franc exchange rate (EURCHF) and the three largest Swiss stocks in terms of market capitalisation, namely, Nestlé (NESN), Novartis (NOVN) and Roche (ROG). We choose to analyse three individual stocks instead of the SMI for three reasons. First, we deem it more appropriate to test our assumptions on more than two variables. Moreover, the three stocks make up 50% of the SMI composition, and thus they are representative of the index itself. Furthermore, all three companies are multinational corporations with a homogeneous exposure to foreign markets and a negligible fraction of revenues coming from the Swiss market (see Nestlé, 2019; Novartis, 2019; Roche, 2019). The last point suggests that the effect (if any) of EURCHF on these stocks does not depend on the idiosyncrasies of each firm.

The dataset consists of daily returns spanning from January 2005 to September 2019 and includes $n = 3832$ observations. The goal is to assess whether EASE can retrieve a correct causal order for the set of four variables. As ground truth, we conjecture that large changes in EURCHF (both positive and negative) will cause large changes in the stock returns, but not *vice versa*. Figure 1.4 shows the causal structure corresponding to our hypothesis.

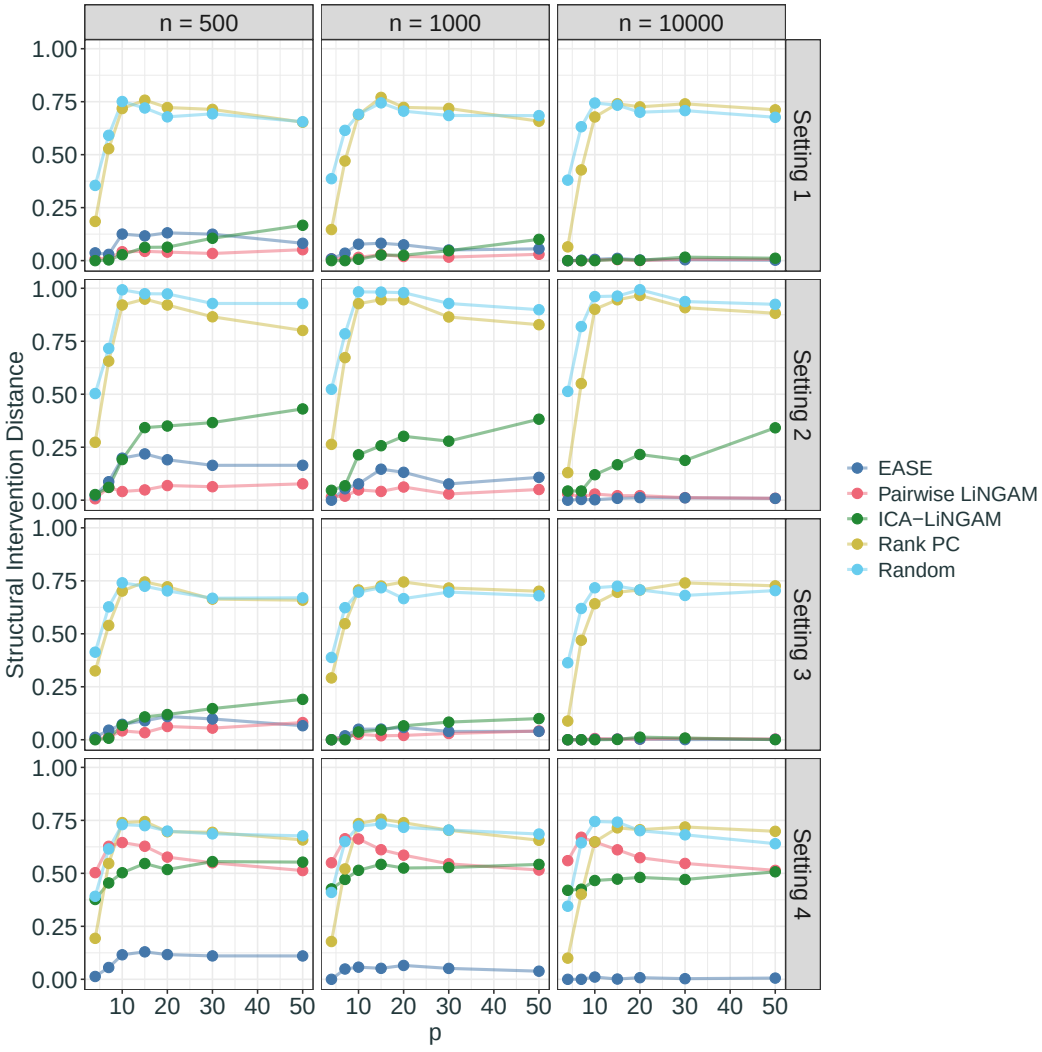


Figure 1.3: The figure refers to Section 1.5.1. It shows the SID averaged over 50 simulations, for each method, setting, sample size n and dimension p, when the tail index is $\alpha = 1.5$. Each row of the figure corresponds to one setting. In order, the settings are: (1) Linear SCM; (2) Linear SCM with hidden confounders; (3) Nonlinear SCM; (4) Linear SCM where each variable is transformed to a uniform margin.

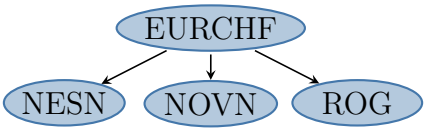


Figure 1.4: A DAG representing a plausible causal structure among the daily returns of Euro Swiss franc exchange rate (EURCHF), Nestlé (NESN), Novartis (NOVN) and Roche (ROG).

Before running the EASE algorithm, we assess the tail behaviour of each variable by estimating the shape parameter ξ of a generalised Pareto distribution (see Embrechts et al., 1997, Sec. 3.4) on the threshold data. Recall that ξ is the reciprocal of the tail index $\alpha = 1/\xi$, if $\xi > 0$. For each variable and each tail (upper and lower) we estimate the ξ parameter using 200 observations, corresponding to the 95%-quantile, approximately. The estimated parameters and their standard errors for the upper tails are 0.31 (0.08)

for EURCHF, 0.25 (0.08) for NESN, 0.16 (0.07) for NOVN, and 0.25 (0.09) for ROG. Regarding the lower tail, the estimated parameters and their standard errors are 0.27 (0.08) for EURCHF, 0.12 (0.08) for NESN, 0.17 (0.08) for NOVN, and 0.23 (0.10) for ROG. By adding and subtracting two standard errors from each point estimate of ξ, we observe that the lower shape parameter of Nestlé and Novartis is not significantly different from zero. Moreover, the shape parameters for the Euro Swiss franc exchange rate and for Roche are significantly different from zero in both tails. In addition, since the confidence intervals of all estimates are overlapping, the assumption of a common shape parameter seems reasonable. It seems however that the returns of Nestlé and Novartis have slightly lighter tails compared to the other two variables.

With the goal of recovering a causal order with EASE, first we estimate the $\widehat{\Psi}$ matrix from the full dataset, by setting the number of exceedances to $k = 10$ (this corresponds to $\lfloor n^{0.3} \rfloor$, approximately). We run the EASE algorithm on the matrix $\widehat{\Psi}$, and we obtain the causal order $\widehat{\pi}^{-1}$ = (EURCHF, NOVN, ROG, NESN); this agrees with the proposed ground truth of Figure 1.4. As a comparison, also ICA-LiNGAM and Pairwise LiNGAM recover a causal order that agrees with our hypothesis.

Since our results are based on the $k = 10$ upper order statistic, we assess the variability of the estimates $\widehat{\Psi}$ for different values of $k_n = \lfloor n^{\nu} \rfloor$, with $\nu \in [0.2, 0.7]$. Figure A.5 shows the estimated coefficients $\widehat{\Psi}$ for the pairs (EURCHF, NESN), (EURCHF, NOVN), and (EURCHF, ROGN), with the corresponding 90% bootstrap confidence intervals. In the three plots, the black (blue) line corresponds to the estimated coefficient $\widehat{\Psi}_{\text{EURCHF},i}$ ($\widehat{\Psi}_{i,\text{EURCHF}}$), with i = NESN, NOVN, ROG. We can interpret the difference between the black and blue lines as a causal signal, since $\Psi_{ij} - \Psi_{ji} > 0$ if variable i causes variable j, for $i, j \in V$ (see Section 1.4.1). For the pairs (EURCHF, NESN) and (EURCHF, NOVN) the blue and the black lines overlap for all values of the upper order statistic k, and therefore any possible causal effect is not identified by the estimated coefficient $\widehat{\Psi}$. This result agrees with Example 1.2 of Section 1.4.3 which shows that the causal tail coefficients do not identify a causal signal when the ancestor has a heavier tail than its descendant — as is the case for the pairs (EURCHF, NESN) and (EURCHF, NOVN). In contrast, if we consider the pair (EURCHF, ROGN) we notice that the difference $\widehat{\Psi}_{\text{EURCHF,ROGN}} - \widehat{\Psi}_{\text{ROGN,EURCHF}}$ is positive for all fractional exponents $\nu \leq 0.4$. This can be explained by the fact that EURCHF and ROGN have comparable tail indices, and therefore it is easier for the coefficient Ψ to detect a possible causal effect between the variables. In Appendix A.6, we show the dynamic evolution of the $\widehat{\Psi}$ coefficient across time.

Given the highly complex nature of financial markets, we do not take the conclusion of this experiment as a definite answer but rather consider it as an indication for a possible causal relationship in this data.

1.5.3 River data

We apply the EASE algorithm to the average daily discharges of the rivers located in the upper Danube basin. This dataset has been studied in Asadi et al. (2015), Engelke and Hitz (2020) and Mhalla et al. (2020), and it is made available by the Bavarian Environmental Agency (http://www.gkd.bayern.de). We consider average daily discharges for 12 stations along the basin, representing the different tributaries and different sections of the Danube, and 11 of them are a subset of the 31 stations selected by Asadi et al. (2015). We exclude some of the 31 stations that are spatially very close since those are highly

dependent and almost indistinguishable. For convenience, we name the stations with the same numbers used in Asadi et al. (2015). While Asadi et al. (2015) decluster the data prior to their analysis in order to obtain independent samples, we use all observations despite the possible temporal dependence. In fact, for extreme value copulas, Zou et al. (2021) show that the use of a larger but possibly dependent dataset can decrease the asymptotic estimation error. Moreover, Fawcett and Walshaw (2007) argue that considering all exceedances over a high threshold reduces the bias of the maximum likelihood estimators compared to a declustered analysis. To account for the time dependence of the exceedances, they adjust the standard errors using methods presented by Smith (1990). In this experiment, we compute the standard errors according to the adjustment proposed by Fawcett and Walshaw (2007).

The dataset spans from 1960 to 2009, where we consider only the summer months, i.e., June, July, and August. The rationale is that most of the extreme observations occur in summer due to heavy rainfall. The final dataset contains $n = 4600$ observations. The rivers have an average volume that ranges between 20 m^3/s (for the upstream rivers) and 1400 m^3/s (for the downstream rivers). A map of the basin can be seen in Figure A.8 in Appendix A.5. In order to implement our method, we first assess whether the equal tail index assumption is satisfied. To do so, we consider a regional model similar to the one presented by Asadi et al. (2015). We split the stations into four separate regions. Region 1 contains three stations in the southwest of the upper Danube basin and the catchment areas are located at mid-altitude; region 2 includes three stations in the Inn-Salzach basin whose tributaries are located in high-altitude alpine regions; region 3 contains four stations along the main Danube with large average water volume; region 4 comprises two stations in the north of the Danube. For each region, we fit a Poisson point process likelihood (Coles, 2001, Chap. 7) by considering exceedances over the 90% quantile and by constraining the shape parameter ξ to be equal across the stations within the same region. To address the presence of temporal dependence in the exceedances, we adjust the standard errors as shown by Fawcett and Walshaw (2007) and, based on these, we compute approximate confidence intervals. For each region, the estimated shape parameter and the corresponding confidence intervals are 0.167 (0.062, 0.273), 0.145 (0.047, 0.242), 0.133 (0.027, 0.239) and 0.229 (0.099, 0.358), respectively. The fact that the confidence intervals overlap suggests that our assumption of equal tail index across the variables is satisfied. Moreover, all confidence intervals do not include the zero value, and therefore the data can be deemed to be heavy-tailed.

We perform two separate analyses to identify causal structures both in space and time. Regarding the spatial structure, the goal is to recover the causal order of the network flow of the 12 stations on the rivers. We consider observations that occur on the same day because the 12 stations are at most 200 km apart from each other and the water flows at ten kilometer per hour, on average. Recovering the causal order of the spatial network is a non-trivial task for two reasons. First, the water discharges at the stations can be confounded by rainfall that spreads out across the region of interest. Second, the large difference in water volume between the stations can further mask the causal structure of the water flow. The true DAG of the spatial disposition of the stations is shown in Figure 1.5. We run the EASE algorithm based on the Γ coefficient, considering the upper tails, and setting the number of exceedances to $k = 29 = \lfloor n^{0.4} \rfloor$. The estimated causal order $\widehat{\pi}^{-1} = (23, 32, 26, 28, 19, 21, 11, 9, 7, 14, 13, 1)$ is correct, and the corresponding fully connected DAG has an SID equal to 0. We also run ICA-LiNGAM and Pairwise LiNGAM on the same dataset, and we obtain an SID of 0 and 0.053, respectively. Clearly,

in this example, the causal structure in the bulk of the distribution is the same as in the extremes. To assess the variability of our results, we compute the average SID of EASE

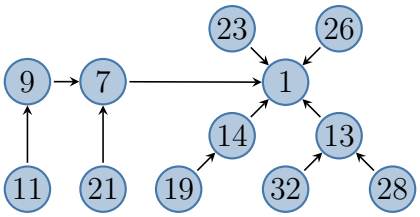


Figure 1.5: DAG representing the spatial configuration of the stations across the upper Danube basin.

over 50 bootstrap samples, for different values of the threshold parameter $k = \lfloor n^\nu \rfloor$, $\nu \in [0.2, 0.7]$ — see Figure A.4. From this figure, we observe that the fractional exponent $\nu \approx 0.4$ yields a good performance both in terms of SID and variability. The value for the optimal fractional exponent agrees with the empirical findings of Section 1.5.1.

Concerning the time-series analysis, we consider each station individually and try to recover the direction of time from the lagged data. Consider an autoregressive (AR) process of order $p \geq 1$,

$$X_t = \sum_{j=1}^{p} \beta_j X_{t-j} + \varepsilon_t, \qquad t \geq 0,$$

where the ε_t are regularly varying with comparable tails, and β_j satisfies the classical stationarity condition for AR processes, $j = 1, \ldots, p$ (see Brockwell and Davis, 2002, Chapter 3.1, equation (3.1.4)). For a detailed discussion of such time series models we refer to Basrak and Segers (2009) and Embrechts et al. (1997, Chapter 7). Peters et al. (2009) prove that an AR(p) process is time reversible, i.e., can be represented by an AR(p) process in the reversed time direction, if and only if the noise is Gaussian. This means that for heavy-tailed random variables, one can in principle detect the direction of time from the data. For each station, we construct a dataset D where the rows correspond to different days and the columns to different lags. We denote by $X_0, X_1, \ldots, X_6$ the columns containing the current and lagged values of the station discharge. We then run EASE on the dataset D and recover a causal order $\hat{\pi}$ over the seven variables $X_0, \ldots, X_6$. We say that the direction of time is correctly inferred if the estimated causal order places the lags in the correct position, i.e., $\hat{\pi}(j) < \hat{\pi}(i)$ if $j < i$, for $i, j = 0, \ldots, 6$. EASE successfully recovers the direction of time for 11 out of the 12 stations. As a comparison, ICA-LiNGAM, and Pairwise LiNGAM find the correct order for 9 and 11 stations, respectively.

So far, we have not considered a multivariate time series analysis of the dataset. On the one hand, in this particular application of the river data, the effects among the stations are almost instantaneous — due to the closeness of the water catchment areas, the fact that we have daily values and the river speed, which is about ten kilometres per hour. On the other hand, time information usually helps in estimating causal relations. For this reason, we apply multivariate Granger causality (Granger, 1969) to the river data, considering one day lag. For each pair of stations (i, j), we say that i Granger-causes j if the corresponding p-value is significant at a 0.05 level, after the Bonferroni correction. We sort the significant p-values in ascending order, and we sequentially add the directed edge (i, j) if nodes i and j are not connected. We continue until the skeleton of the resulting graph is connected, or all the p-values have been selected. The resulting directed tree achieves an SID of 0.083. Alternatively, if we sequentially add directed edges (i, j) that

do not create cycles (until all p-values have been selected) we obtain a DAG with an SID of 0.196.

1.6 Discussion and future work

In several real-world phenomena, the causal mechanisms between the variables appear more clearly during extreme events. Moreover, there are situations where the causal relationship in the bulk of the distribution differs from the structure in the tails. We have introduced an algorithm, named extremal ancestral search (EASE), that is shown to consistently recover the causal order of a DAG from extreme observations only. EASE has the advantage of relying on the pairwise causal tail coefficient, and therefore it is computationally efficient. In addition, our algorithm can deal with the presence of hidden confounders and performs well for small sample sizes and high dimensions. The EASE algorithm is robust to model misspecifications, such as nonlinear relationships in the bulk of the distribution, and strictly monotone increasing transformations applied marginally to each variable.

This work sheds light on a connection between causality and extremes, and thereby opens new directions of research. In particular, it might be interesting to study the properties of the causal tail coefficient under more general conditions. This includes high-dimensional settings where the dimension grows with the sample size, more general SCMs where the functional relations between the variables can be nonlinear, and settings where the noise variables have lighter tails. For example, in future research, one could combine EASE with regression techniques to obtain the complete DAG structure, compute the residuals and then test them for the assumption of common tail indices.

Another possible extension may consider multivariate time series data, where the temporal order of cause and effect could help to estimate causal relationships among variables, see, e.g., Granger (1969). Future work might study how to combine our approach with the Granger causality framework. For instance, one could first perform a Granger causality analysis, and then, apply EASE to the residuals of the vector autoregression model. For a careful study in this direction, it would also be necessary to investigate the statistical properties of the residuals.

Acknowledgements

We thank Cesare Miglioli, Stanislav Volgushev and Linbo Wang for helpful discussions. We are grateful to the editorial team and two anonymous referees for constructive comments that improved the paper. JP was supported by research grants from VILLUM FONDEN and the Carlsberg Foundation, and SE was supported by the Swiss National Science Foundation.

Chapter 2

Extremal Random Forests

Joint work with

Edossa Merga Terefe, and Sebastian Engelke

Abstract

Classical methods for quantile regression fail in cases where the quantile of interest is extreme and only few or no training data points exceed it. Asymptotic results from extreme value theory can be used to extrapolate beyond the range of the data, and several approaches exist that use linear regression, kernel methods or generalized additive models. Most of these methods break down if the predictor space has more than a few dimensions or if the regression function of extreme quantiles is complex. We propose a method for extreme quantile regression that combines the flexibility of random forests with the theory of extrapolation. Our extremal random forest (ERF) estimates the parameters of a generalized Pareto distribution, conditional on the predictor vector, by maximizing a local likelihood with weights extracted from a quantile random forest. Under certain assumptions, we show consistency of the estimated parameters. Furthermore, we penalize the shape parameter in this likelihood to regularize its variability in the predictor space. Simulation studies show that our ERF outperforms both classical quantile regression methods and existing regression approaches from extreme value theory. We apply our methodology to extreme quantile prediction for U.S. wage data.

Keywords: extreme quantiles, local likelihood estimation, quantile regression, random forests, threshold exceedances.

2.1 Introduction

Quantile regression is a well-established technique to model statistical quantities that go beyond the conditional expectation that is used for standard regression analysis (Koenker and Bassett, 1978). This is particularly valuable in applications such as economics, survival analysis, medicine, and finance (Angrist et al., 2006; Yang, 1999; Heagerty and Pepe, 1999; Taylor, 1999; Yu et al., 2003), where one needs to model the heteroschedasticity of the response or conditional quantiles such as the median.

In this paper, we consider the problem of estimating high conditional quantiles of a response variable $Y \in \mathbb{R}$ given a set of predictors $X \in \mathbb{R}^p$ in large, but fixed, dimensions,

an important task in risk assessment for rare events (Chernozhukov, 2005). For a fixed predictor value x, define $Q_x(\tau)$ as the quantile at level $\tau \in (0,1)$ of the conditional distribution of $Y \mid X = x$. We are interested in the estimation of extreme quantiles where $\tau \approx 1$ is close to one. This estimation problem exhibits two fundamental challenges that are illustrated in Figure 2.1, which shows a simulation similar to Athey et al. (2019, Figure 2). The predictor space has $p = 40$ dimensions and only the first variable X_1 has a signal corresponding to a scale shift in Y; see Example 2.1 in Section 2.3.1 for details.

The first challenge in estimating $Q_x(\tau)$ relates to the fact that for an extreme probability level, say $\tau = 0.9995$ as in Figure 2.1, there are typically only few or no observations in the sample that exceed the corresponding conditional τ-quantiles. Indeed, for a sample of size n, the expected number of exceedances above the conditional τ-quantile is $n(1-\tau)$, which becomes smaller than one if $\tau > 1 - 1/n$. Therefore, using an empirical estimator based on quantile loss leads to a large bias. A second challenge stems from the possibly large, while fixed, dimension of the predictor space $\mathbb{R}^p$, where there might be no training observations close to x; note that the Figure 2.1 only shows the first of the 40 dimensions of X. Too simple regression models may then introduce additional bias.

The first challenge can be addressed by relying on tail approximations motivated by extreme value theory (e.g., de Haan and Ferreira, 2006), which allow the extrapolation to quantile levels beyond the range of the data. Such methods typically consider (transformations of) linear (Chernozhukov, 2005; Wang and Tsai, 2009; Wang et al., 2012; Wang and Li, 2013) functions, additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), non-parametric regression (Beirlant et al., 2004; Martins-Filho et al., 2015) and local smoothing methods (Daouia et al., 2011; Gardes and Stupfler, 2019; Velthoen et al., 2019). However, these existing approaches are either not flexible enough to model complex response surfaces or do not scale well in higher dimensions p of the predictor space.

Regarding the second challenge, several quantile regression methods have been proposed in the statistical and machine learning literature that can cope with large, while fixed, dimensions of the predictor space and complex regression surfaces (Taylor, 2000; Friedman, 2001b). In particular, there exist several forest-based approaches for quantile regression (Meinshausen, 2006; Athey et al., 2019). These methods are based on (extensions of) the random forest originally developed by Breiman (2001) and can estimate flexible quantile regression functions. Compared to methods such as gradient boosting and neural networks, the main advantage of forest-based approaches is that they require little tuning and that their statistical properties are relatively well understood (Athey et al., 2019). Moreover, they scale well with the dimension of the predictor space as opposed to approaches based on generalized additive models (Koenker, 2011) and kernel-based methods (Yu and Jones, 1998). While these methods work well for estimation of quantiles inside the data range, such as $\tau_0 = 0.8$ in Figure 2.1, their performance deteriorates for quantile estimation at extreme levels $\tau \approx 1$ close to the upper endpoint of the response distribution.

In this paper, we bring together ideas from extreme value theory and forest-based regression methods to tackle the challenges of extreme quantile regression in predictor spaces with possibly large, but fixed, dimensions p. To extrapolate beyond the data range, we rely on the approximation by the generalized Pareto distribution (GPD) of the exceedances over an intermediate threshold; see the triangles in Figure 2.1. Under mild assumptions, the conditional quantile of Y, given $X = x$, at level $\tau \approx 1$ can the be

approximated by (Balkema and de Haan, 1974; Pickands, 1975)

$$Q_x(\tau) \approx Q_x(\tau_0) + \frac{\sigma(x)}{\xi(x)} \left[\left(\frac{1-\tau}{1-\tau_0} \right)^{-\xi(x)} - 1 \right], \tag{2.1.1}$$

where $Q_x(\tau_0)$ is an intermediate quantile at level $\tau_0 < \tau$ and the second term on the right-hand side is quantile function of the GPD indexed by the conditional scale $\sigma(x) > 0$ and shape parameter $\xi(x) \in \mathbb{R}$. This includes responses with heavy tails ($\xi(x) > 0$), light tails ($\xi(x) = 0$) and with finite upper end points ($\xi(x) < 0$). The intermediate quantile level τ_0 is chosen small enough such that the conditional quantiles $Q_x(\tau_0)$ can be estimated by classical regression methods. At the same time, it should be large enough so that the approximation in (2.1.1) by the GPD is accurate.

In order to cope with complex response surfaces and large-dimensional predictor spaces, we rely on ideas from the random forest literature (Meinshausen, 2006; Athey et al., 2019). Our new extremal random forest (ERF) localizes the estimation of the GPD parameter vector $\theta(x) = (\sigma(x), \xi(x))$ around the predictor value x using forest-based weights. Since only few extreme observations are typically available for training, the simple tuning of random forests turns out to be of great advantage. Under certain conditions, we show consistency of the ERF estimator $\hat{\theta}(x)$ for the true conditional parameter vector $\theta(x)$. Since our loss function, namely the GPD log-likelihood, is non-convex, the proof strategy of Athey et al. (2019) cannot be used, and we rely on the theory of Newey (1991).

Our ERF algorithm combines the advantages of accurate tail extrapolation at levels $\tau \approx 1$ with a flexible regression method that scales well with predictor dimension. In simulations, we show that ERF outperforms extreme value theory and quantile regression methods to estimate extreme quantiles. Moreover, it is competitive with the recent gradient boosting by Velthoen et al. (2021) and has the advantage of significantly easier tuning and the theoretical guarantee of our consistency result. Finally, we apply our methodology to extreme quantile prediction for U.S. wage data (Angrist et al., 2009). The ERF algorithm is available as an `R` package on https://github.com/nicolagnecco/erf.

2.2 Background

2.2.1 Extreme Value Theory

The first challenge of extreme quantile regression is that only a few or even no data points exceed the quantiles of interest. This section considers the classical case of unconditional extremes without predictors. Let $Y_1, \ldots, Y_n$ be n independent copies of a real-valued random variable Y. The notion of an extreme quantile $\tau = \tau_n$ is typically expressed relative to the sample size n. The expected number of observations in the sample that exceed the τ_n-quantile is then $n(1-\tau_n)$. A quantile with level $\tau_n \to 1$ such that $n(1-\tau_n) \to \infty$ is called an intermediate quantile. Empirical estimation in this case still works well since the effective sample size, that is, the number of exceedances, grows to infinity (de Haan and Ferreira, 2006). For risk assessment, the most critical case is if the quantile of interest is eventually beyond the range of the data, that is, $(1-\tau_n)n \to 0$ as $n \to \infty$. Then, we can no longer rely on empirical estimators but must resort to asymptotically motivated approximations from extreme value theory.

Let $u^* \in (0, \infty]$ be the upper endpoint of the distribution of Y. Under mild regularity assumptions on the tail of Y, the Pickands–Balkema–De Haan theorem (Balkema and

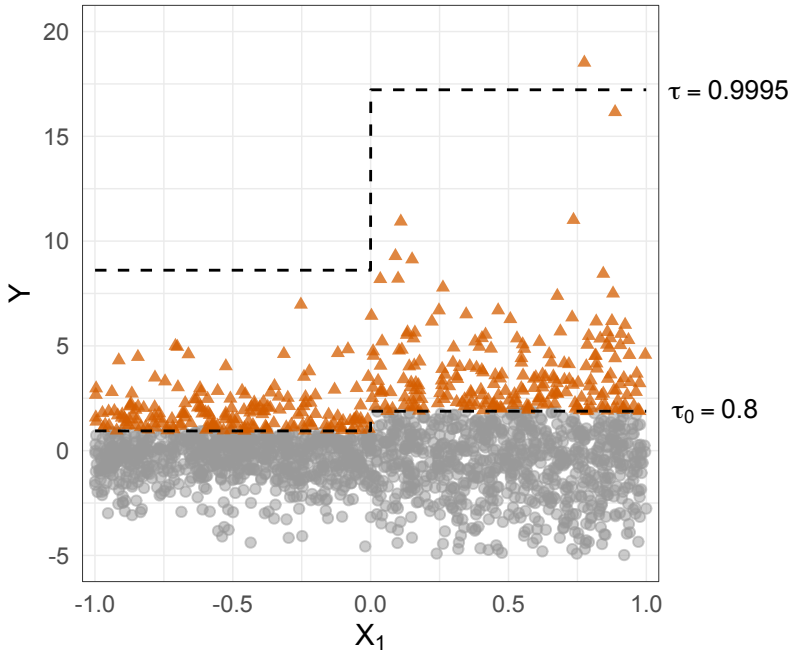


Figure 2.1: Realization of $n = 2000$ samples from the generative model in Example 2.1 in Section 2.3.1. Response Y is plotted against the first predictor X_1. Dashed lines represent the quantile functions associated to the intermediate $\tau_0 = 0.8$ and high $\tau = 1 - 1/n = 0.9995$ quantile levels. Triangles are observations above the intermediate threshold.

de Haan, 1974; Pickands, 1975) states that there exists a normalizing function $\sigma(u) > 0$ such that

$$\lim_{u \to u^*} \mathbb{P}\left(\frac{Y - u}{\sigma(u)} \leq z \mid Y > u \right) = G(z; (1, \xi)), \tag{2.2.1}$$

where the limit on the right-hand side is the distribution function of the generalized Pareto distribution (GPD) (Pickands, 1975) given by

$$G(z; \theta) = 1 - \left(1 + \frac{\xi}{\sigma} z \right)_+^{-1/\xi}, \quad z > 0, \tag{2.2.2}$$

and $\theta = (\sigma, \xi) \in (0, \infty) \times \mathbb{R}$ is the parameter vector consisting of scale and shape, respectively. The shape parameter $\xi \in \mathbb{R}$, also known as the extreme value index (Beirlant et al., 2005), characterizes the decay of the tail of Y. If $\xi > 0$, then Y is heavy-tailed; if $\xi = 0$, then Y is light-tailed; if $\xi < 0$ then Y has a finite upper endpoint. Moreover, the GPD is a natural model for the distribution tails since it is the only possible limit of threshold exceedances as in (2.2.1).

The GPD approximation can be directly translated into an approximation for the small probability of Y exceeding a high threshold y. By Bayes' theorem and (2.2.1) we obtain

$$\mathbb{P}(Y > y) = \mathbb{P}(Y > u)\, \mathbb{P}(Y > y \mid Y > u) \approx \mathbb{P}(Y > u) \left\{ 1 - G(y - u; \sigma, \xi) \right\},$$

where $u < y$ denotes an intermediate threshold. Combining this approximation with (2.2.2) and letting $\mathbb{P}(Y > y) = 1 - \tau$ and $\mathbb{P}(Y > u) = 1 - \tau_0$, we obtain an approximation for the

τ-quantile of Y as

$$Q(\tau) \approx Q(\tau_0) + \frac{\sigma}{\xi}\left[\left(\frac{1-\tau}{1-\tau_0}\right)^{-\xi} - 1\right], \tag{2.2.3}$$

where $Q(\tau_0) := F_Y^{-1}(\tau_0)$ denotes the intermediate quantile at level $\tau_0 < \tau$.

In applications, the scale and shape parameters of the GPD have to be estimated from independent observations $Y_1, \ldots, Y_n$ of Y. We fix an intermediate quantile level τ_0 and define the exceedances $Z_i = (Y_i - \hat{Q}(\tau_0))_+$, $i = 1, \ldots, n$, where $\hat{Q}(\tau_0)$ denotes the empirical τ_0 quantile. We can estimate the GPD parameter vector θ by maximum-likelihood, where the negative log-likelihood (or deviance) contribution of the ith exceedance Z_i is

$$\ell_\theta(Z_i) = \log\sigma + \left(1 + \frac{1}{\xi}\right)\log\left(1 + \frac{\xi}{\sigma}Z_i\right), \quad \theta \in (0,\infty) \times \mathbb{R}, \tag{2.2.4}$$

if $Z_i > 0$, and zero otherwise.

2.2.2 Quantile Regression and Generalized Random Forests

Given a pair (X, Y) of predictor vector $X \in \mathbb{R}^p$ and response variable $Y \in \mathbb{R}$, quantile regression deals with modeling the conditional τ-quantile $Q_x(\tau)$ of the conditional distribution of Y given that $X = x$ for a particular predictor value $x \in \mathbb{R}^p$. The main challenge is that the dimension p of the predictor space may be large, while fixed, and that the quantile surface $Q_x(\tau)$ as a function x may be a complex, highly non-linear function.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n independent copies of the random vector (X, Y). In contrast to the setting in Section 2.2.1, classical methods for quantile regression consider a fixed quantile level $\tau \equiv \tau_n$ that does not change with the sample size. On population level, these methods exploit the fact that the conditional quantile function is the minimizer of the expectation of the quantile loss $\rho_\tau(c) = c(\tau - \mathbb{1}\{c < 0\})$, $c \in \mathbb{R}$, (Koenker and Bassett, 1978), that is,

$$Q_x(\tau) = \arg\min_{q \in \mathbb{R}} \mathbb{E}[\rho_\tau(Y - q) \mid X = x]. \tag{2.2.5}$$

The above expectation cannot be estimated directly on the sample level since the set of observed predictor values does not typically include the value x. A natural estimator is

$$\hat{Q}_x(\tau) = \arg\min_{q \in \mathbb{R}} \sum_{i=1}^{n} w_n(x, X_i)\rho_\tau(Y_i - q), \tag{2.2.6}$$

where $x' \mapsto w_n(x, x')$ is a set of localizing similarity weights around the predictor value of interest. The weights can for instance be obtained by a kernel approach (Yu and Jones, 1998), but this is limited to moderately large dimensions (Stone, 1980, 1982).

In order to model more complex quantile surfaces in larger dimensions, Meinshausen (2006) and Athey et al. (2019) propose to use the estimator (2.2.6) with similarity weights $w_n(\cdot, \cdot)$ obtained from a random forest. Random forests (Breiman, 2001) are an ensemble method used for both regression and classification tasks and consist of fitting B decision trees to the training data. In regression settings, each decision tree predicts a test point $x \in \mathbb{R}^p$ by

$$\mu_b(x) := \sum_{i=1}^{n} \frac{\mathbb{1}\{X_i \in L_b(x)\}Y_i}{|\{i : X_i \in L_b(x)\}|}, \quad b = 1, \ldots, B,$$

where $L_b(x) \subset \mathbb{R}^p$ denotes the rectangular region that x belongs to in bth tree. By defining the similarity weights $w_{n,b}(x, X_i) := \mathbb{1}\{X_i \in L_b(x)\}/|\{i : X_i \in L_b(x)\}|$, the random forest predictions can be written as

$$\mu(x) := \frac{1}{B}\sum_{b=1}^{B} \mu_b(x) = \sum_{i=1}^{n} w_n(x, X_i) Y_i,$$

where $w_n(x, X_i) = \sum_{b=1}^{B} w_{n,b}(x, X_i)/B$ is the average weight across B trees.

The original idea of Meinshausen (2006) is to use the weights estimated by this standard regression random forest for quantile regression in (2.2.6). A drawback of this approach is that the similarity weights arise from decision trees that are grown by minimizing the mean squared error loss. This leads to the fact that, as stated in Meinshausen (2006), $w_n(x, X_i)$ takes large values for those observations i such that $\mathbb{E}[Y \mid X = X_i] \approx \mathbb{E}[Y \mid X = x]$. In many situations the conditional expectation is not representative of the whole conditional distribution of $Y \mid X = x$, and it may happen that $w_n(x, X_i)$ is large but $Q_{X_i}(\tau) \not\approx Q_x(\tau)$; see Athey et al. (2019, Figure 2) or our Figure 2.1 where the conditional expectation is constant over the predictor space. In these cases, the similarity weights estimated with standard random forest do not capture the heterogeneity of the quantile function and are thus not well-suited for quantile regression tasks. Athey et al. (2019) introduced generalized random forests (GRF), a method designed to fit random forests with custom loss functions. The GRF retains all the appealing features of classical random forests, i.e., it is simple to fit and requires little tuning of hyperparameters. One of the main applications of GRF is quantile regression, where the trees of the forest are grown to minimize the quantile loss function. In this work, we rely on GRF with quantile loss to estimate similarity weights $w_n(\cdot, \cdot)$ that capture the variation of the entire conditional distribution of $Y \mid X = x$ in the predictor space. In practice, the GRF algorithm estimates simultaneously conditional quantiles at levels $\tau = 0.1, 0.5, 0.9$ as a proxy for the conditional distribution of $Y \mid X = x$. For simplicity, in the sequel, we refer to GRF with quantile loss as GRF.

2.3 Extremal Random Forest

2.3.1 The Algorithm

In this work we study a method for flexible extreme quantile regression where both challenges described in Sections 2.2.1 and 2.2.2 occur simulateneously. Consider the random vector (X, Y) of predictors $X \in \mathcal{X} \subset \mathbb{R}^p$ and response $Y \in \mathbb{R}$, with $\mathcal{X}$ compact. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). In many applications in risk assessment, the goal is to estimate the quantile function $x \mapsto Q_x(\tau) = F^{-1}_{Y|X=x}(\tau)$, at an extreme level $\tau = \tau_n$, where the expected number of observations in the sample that exceed their conditional quantiles is small and possibly tends to 0 as $n \to \infty$; see Section 2.2.1. To illustrate the challenges of this estimation problem, we consider an example where the scale of the response variable Y is modeled as a step function of the covariates X. This corresponds to Athey et al. (2019, Figure 2), except that we assume that the noise of the response variable is heavy-tailed instead of Gaussian.

Example 2.1. Let $X \sim U_p$ be a uniform distribution on the cube $[-1, 1]^p$ in dimension p and $Y \mid X = x \sim s(x)\, T_4$, where T_ν denotes a Student's t-distribution with $\nu > 0$ degrees of freedom. The shape parameter of the conditional distribution $Y \mid X = x$ is

then constant $\xi(x) = 1/\nu(x) \equiv 0.25$ and we choose the $s(x) = 1 + \mathbb{1}\{x_1 > 0\}$ for $x \in \mathbb{R}^p$. The GPD scale parameter $\sigma(x)$ of $Y \mid X = x$ and therefore also the quantile function $Q_x(\tau)$ only depend on X_1. The other predictors are noise variables. ◁

Figure 2.1 in the introduction shows $n = 2000$ observations sampled from the model of Example 2.1 in dimension $p = 40$. The goal is to predict the conditional quantile $Q_x(\tau)$ for a high level of τ, e.g., $\tau = 0.9995$. We observe that the difficulty of the task is twofold. First, because of a possibly large-dimensional, while fixed, predictor space, there might be no training observations close to x; note that we only show the first of the 40 dimensions of X in the figure. Second, the τ-quantile might be out of the range of the data if τ is very close to one. Indeed, for a sample of size n, the expected number of exceedances above the conditional τ-quantile is $n(1-\tau)$, which becomes smaller than one if $\tau > 1 - 1/n$.

Our methodology accurately addresses both of these challenges. For effective localizing in the predictor space, even in large dimensions, we use the weights emerging from GRF (Athey et al., 2019). For correct extrapolation in the tail of the conditional response variable, we rely on the asymptotic theory of extremes and fit a localized generalized Pareto distribution; see Section 2.2.1. More precisely, for an intermediate quantile level τ_0, we assume that the distribution function of $Y - Q_x(\tau_0)$, conditional on $Y > Q_x(\tau_0)$, is approximately generalized Pareto (Balkema and de Haan, 1974) with scale and shape parameters depending on the predictor value x, that is, for any $z > 0$,

$$\mathbb{P}\left(Y - Q_x(\tau_0) \le z \mid Y > Q_x(\tau_0), X = x\right) \approx G\left(z; \theta(x)\right), \tag{2.3.1}$$

where $\theta(x) = (\sigma(x), \xi(x))$, and the scale and shape are continuous functions $\sigma : \mathcal{X} \to (0, \infty)$ and $\xi : \mathcal{X} \to \mathbb{R}$, respectively. This assumption is a conditional version of (2.2.1) and means that the GPD approximation holds for the distribution of $Y \mid X = x$ for any $x \in \mathcal{X}$. It is satisfied by most data generating processes as for instance in Example 2.1.

In order to formulate our estimators of the conditional GPD parameters $\theta(x)$ and the extreme quantile $Q_x(\tau)$, we define the exceedances in the training data as

$$Z_i := (Y_i - \hat{Q}_{X_i}(\tau_0))_+, \quad i = 1, \dots, n; \tag{2.3.2}$$

see the triangles in Figure 2.1. Here, $\tau_0 \in (0, 1)$ is an intermediate probability level that is chosen such that the estimator $\hat{Q}_x(\tau_0)$ of the conditional quantile function can be obtained by classical quantile regression techniques; see Section 2.2.2. In principle, any quantile regression method can be used to fit $Q_x(\tau_0)$. Here, we choose GRF with quantile loss (Athey et al., 2019) since it is a flexible method well-suited for large, but fixed, dimensional quantile regression problems and it requires little tuning.

For the estimation of the GPD parameter vector $\theta(x) = (\sigma(x), \xi(x))$ we rely on those exceedances that carry most information on the tail of the distribution of $Y \mid X = x$. To do so, we use the localizing weight functions $w_n(x, X_i)$ estimated from a GRF (Athey et al., 2019) that may be *different* from the one used to estimate the intermediate quantile $\hat{Q}_x(\tau_0)$. We would like to define the estimator of the conditional GPD parameter $\hat{\theta}(x)$ as the minimizer of the weighted (negative) log-likelihood

$$L_n(\theta; x) = \sum_{i=1}^{n} w_n(x, X_i) \ell_\theta(Z_i) 1\{Z_i > 0\}, \quad x \in \mathcal{X}, \tag{2.3.3}$$

where ℓ_θ is defined in (2.2.4). In practice, the parameter space $\theta(\mathcal{X}) = \{\vartheta \in (0, \infty) \times \mathbb{R} : \vartheta = \theta(x) \text{ for some } x \in \mathcal{X}\}$ is unknown. As explained by Dombry (2015), it is not

guaranteed that the log-likelihood of the generalized extreme value distribution has a global optimum over the parameter space $(0, \infty) \times \mathbb{R}$. In fact, Smith (1985) shows that there exists no maximum likelihood estimator when $\xi \leq -1$. Analogous results apply to the GPD log-likelihood $L_n(\theta; x)$ (Drees et al., 2004). We therefore follow Bücher and Segers (2017) and define $\hat{\theta}(x)$ as the optimizer of $L_n(\theta; x)$ over an arbitrarily large compact set $\Theta \subset (0, \infty) \times (-1, \infty)$ such that $\theta(\mathcal{X}) \subset \operatorname{Int} \Theta$, that is,

$$\hat{\theta}(x) = \operatorname*{arg\,min}_{\theta \in \Theta} L_n(\theta; x). \tag{2.3.4}$$

If there is more than one minimizer, let $\hat{\theta}(x)$ be the smallest with respect to lexicographic order (see Dombry, 2015). The estimated pair $(\hat{Q}_x(\tau_0), \hat{\theta}(x))$ of intermediate quantile and conditional GPD parameters can be plugged into the extrapolation formula (2.1.1) to obtain an estimate $\hat{Q}_x(\tau)$ of the extreme conditional quantile at level $\tau > \tau_0$.

In Algorithm 2, we describe our prediction method, which we call the extremal random forest (ERF). The algorithm consists of two subroutines, namely ERF-Fit and ERF-Predict. The first one estimates a similarity weight function $(x, y) \mapsto w_n(x, y)$ and an intermediate quantile function $x \mapsto \hat{Q}_x(\tau_0)$ from the training data, for $x, y \in \mathcal{X}$. The second procedure predicts the extreme τ-quantile $\hat{Q}_x(\tau_0)$ at point $x \in \mathcal{X}$ by estimating the GPD parameter vector $\theta(x)$ as in (2.3.4). Appendix B.3 shows the estimated GRF

Algorithm 2 Extremal random forest (ERF)

Denote by $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ the training data. Let $x \in \mathbb{R}^p$ be a test predictor value. Specify intermediate and extreme quantile levels τ_0 and τ, respectively, with $\tau_0 < \tau$. Let α be a vector of hyperparameters supplied to GRF.

procedure ERF-Fit($\mathcal{D}, \tau_0, \alpha$)
  $w_n(\cdot, \cdot) \leftarrow \text{GRF}(\mathcal{D}, \alpha)$ ▷ Fit similarity weight function.
  $\hat{Q}_\cdot(\tau_0) \leftarrow$ QuantileRegression($\mathcal{D}$) ▷ Fit intermediate quantile function.
  output `erf` $\leftarrow [\mathcal{D}, w_n(\cdot, \cdot), \hat{Q}_\cdot(\tau_0)]$ ▷ Return an `erf` object.

procedure ERF-Predict(`erf`, x, τ)
  $Z_i \leftarrow (Y_i - \hat{Q}_{X_i}(\tau_0))_+$, with $i = 1, \ldots, n$ ▷ Compute exceedances.
  $\hat{\theta}(x) \leftarrow \arg\min_\theta L_n(\theta; x)$ as in (2.3.3) ▷ Estimate GPD parameters.
  output $\hat{Q}_x(\tau)$ as in (2.1.1) ▷ Return extreme quantile.

The subroutine GRF estimates the similarity weight function $w_n(\cdot, \cdot)$ using the generalized random forest of Athey et al. (2019). The subroutine QuantileRegression fits the intermediate conditional quantile function $\hat{Q}_\cdot(\tau_0)$ using a classical quantile regression technique. The object `erf` is a list containing the training data $\mathcal{D}$, the fitted intermediate quantile $\hat{Q}_\cdot(\tau_0)$, and the estimated similarity weight function $w_n(\cdot, \cdot)$.

weights $w_n(x, X_i)$ used in the likelihood in (2.3.3) for Example 2.1 and specific values of x. It can be seen that the weights are large for training observations X_i where the distribution of $Y \mid X = X_i$ is equal to the one of $Y \mid X = x$.

2.3.2 Consistency

Our ERF provides an estimate $\hat{\theta}(x) = (\hat{\sigma}(x), \hat{\xi}(x))$ of the conditional GPD parameter $\theta(x)$ that describes the tail of the distribution of $Y \mid X = x$. The method is at the interface of

random forests and extreme value theory, and both fields have their challenges related to the analysis of asymptotic properties.

Consistency and asymptotic normality of classical (Meinshausen, 2006; Biau, 2012; Scornet et al., 2015; Wager and Athey, 2018) and generalized random forests (Athey et al., 2019) have only recently been established. The results by Athey et al. (2019) require regularity conditions (see Assumptions 1–6 of their paper) that are not satisfied in our setting. In particular, the negative GPD log-likelihood $\theta \mapsto \ell_\theta(z)$ that we consider is not a convex function and, therefore, it does not satisfy Assumption 6 in Athey et al. (2019). On the other hand, the asymptotic analysis of extreme value estimators is notoriously difficult due to the pre-limit approximation in (2.3.1) and changing distributional support (Smith, 1985; Drees et al., 2004). Recent papers have worked out the asymptotics for the unconditional i.i.d. case (Dombry, 2015; Bücher and Segers, 2017; Dombry and Ferreira, 2019).

We will not show consistency of the ERF under the most general conditions on the distributional tail of $Y \mid X = x$ since the required technicalities would be beyond the scope of this paper. We list all assumptions needed for our theorem and discuss possible relaxations after the statement. The first assumption deals with the data generating process.

Assumption 2.1. Let $X \in \mathcal{X}$ have a density that is bounded away from 0 and ∞ and support $\mathcal{X} := [0,1]^p$. For large enough τ_0, suppose the conditional intermediate quantile function $Q_X(\tau_0)$ is known. Furthermore, assume that the distribution function of $Y - Q_X(\tau_0)$, conditional on $Y > Q_X(\tau_0)$, is *exactly* generalized Pareto with parameter vector $\theta(X)$.

The next assumption addresses how the parameter vector $\theta(x)$ depends on the predictor $X = x$. We consider only the most relevant case of positive shape parameter $\xi(x) > 0$, that is, where $Y \mid X = x$ is heavy-tailed.

Assumption 2.2. Let $\theta(x) = (\sigma(x), \xi(x))$ denote the bivariate regression function for the GPD parameters, for $x \in \mathcal{X}$. Assume $\sigma : \mathcal{X} \to (0,\infty)$ and $\xi : \mathcal{X} \to (0,\infty)$ are continuous functions on $\mathcal{X}$. Furthermore, assume their first order partial derivatives are continuous in the interior and exist on the boundary of $\mathcal{X}$; we refer to Appendix B.2 for a definition of partial derivative on the boundary. Notice that the parameter space $\theta(\mathcal{X}) \subset (0,\infty) \times (0,\infty)$ is compact and bounded away from the origin.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). The first step of our algorithm consists of fitting a generalized random forest on the training data to obtain similarity weights. To show consistency, we make the following standard assumptions on how this forest is built.

Assumption 2.3. Let $w_n(x, y)$ denote the similarity weights for $x, y \in \mathcal{X}$ estimated by a GRF. We assume the forest satisfies Specification 1 of Athey et al. (2019). In particular, we assume that each tree in the forest is symmetric, places balanced splits, and is randomized (see Athey et al., 2019). We require that each tree is fitted on a subsample of the training data with size $s < n$, such that $s \to \infty$ and $s/n \to 0$ as $n \to \infty$, and that the forest consists of $\binom{n}{s}$ trees fitted on all possible subsamples of size s.

In practice, one builds a forest by estimating B trees. Our theoretical results hold for forests made of $\binom{n}{s}$ trees fitted on all possible subsamples of size s. For this reason,

similarly to Wager and Athey (2018), we assume that B is large enough so that the Monte Carlo effect is negligible. Furthermore, Assumption 2.3 does not require that the trees in the forest are honest in the sense of Athey et al. (2019). The reason is that, as opposed to Athey et al. (2019), our conditional response distribution belongs to the parametric GPD family. In practice, we find that honesty helps our algorithm perform better, and the result below remains true under this additional, stronger assumption.

Theorem 2.4. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y) as specified in Assumptions 2.1 and 2.2. Let $x \in \operatorname{Int} \mathcal{X}$ be a fixed test predictor value, and denote by $w_n(x, X_i)$ the similarity weights estimated with a forest satisfying Assumption 2.3. Let $\Theta \subset (0, \infty) \times (0, \infty)$ be an arbitrary compact set such that $\theta(\mathcal{X}) \subset \operatorname{Int}(\Theta)$, and let $\hat{\theta}(x)$ denote a sequence of estimators minimizing* (2.3.3)*. Then, $\hat{\theta}(x) \to \theta(x)$ in probability as $n \to \infty$.*

The proof relies on the theory of Newey (1991) and is in Appendix B.1. To the best of our knowledge, this is the first consistency proof for a tree-based extreme quantile regression method. Wang and Tsai (2009) show asymptotic normality for the model parameters for the heavy-tailed case, but only in the situation where the covariate dependence is linear. There are no asymptotic results for models for generalized Pareto distributions with parameters depending in a more complex way on the covariates such as through generalized additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), trees (Farkas et al., 2020) or gradient boosting (Velthoen et al., 2021).

Similarly to Wang and Tsai (2009), we focus on the heavy-tailed case where $\xi(x) > 0$ for all $x \in \mathcal{X}$. Relaxing this assumption to $\xi(x) \in \mathbb{R}$ would make the support of the generalized Pareto distribution depend on the model parameters. This would require a different proof strategy and additional care in terms of Lipschitz conditions, but some ideas from the i.i.d. case in Bücher and Segers (2017, Lemma E.2) might be helpful.

A further simplification in our setup is that we assume that the approximation in (2.3.1) is an equality. Dropping this assumption would require additional conditions to control the approximation error and would add a further level of technicality to the proofs. Similar assumptions are often made in the literature as for instance in Bücher and Segers (2017) for the i.i.d. case for generalized extreme value distributions.

2.3.3 Hyperparameter Tuning

Generalized random forests have several tuning parameters, such as the number of predictors selected at each split and the minimum node size. This section presents a cross-validation scheme to tune such hyperparameters within our algorithm. For large values of $\tau \approx 1$, the quantile loss is not a reliable scoring function since there might be few or no test observations above this level. In our case, we can instead rely on the tail approximation in (2.3.1) and use the deviance of the GPD as a reasonable metric for cross-validation. Let $\mathcal{N}_1, \ldots, \mathcal{N}_M$ be a random partitioning of $\{1, \ldots, n\}$ into M equally sized folds of the training data. For a sequence $\alpha_1, \ldots, \alpha_J$ of tuning parameters, we fit an `erf` object on the training set (X_i, Y_i), $i \notin \mathcal{N}_m$, for each α_j and each fold m as described in the ERF-Fit function in Algorithm 2. Given the fitted `erf` object, we estimate the GPD parameter vector $\hat{\theta}(X_i; \alpha_j)$ on the validation set (X_i, Y_i), $i \in \mathcal{N}_m$ as in the ERF-Predict function in Algorithm 2, and evaluate the cross-validation error by

$$CV(\alpha_j) = \sum_{m=1}^{M} \sum_{i \in \mathcal{N}_m} \ell_{\hat{\theta}(X_i; \alpha_j)}(Z_i) 1\{Z_i > 0\},$$

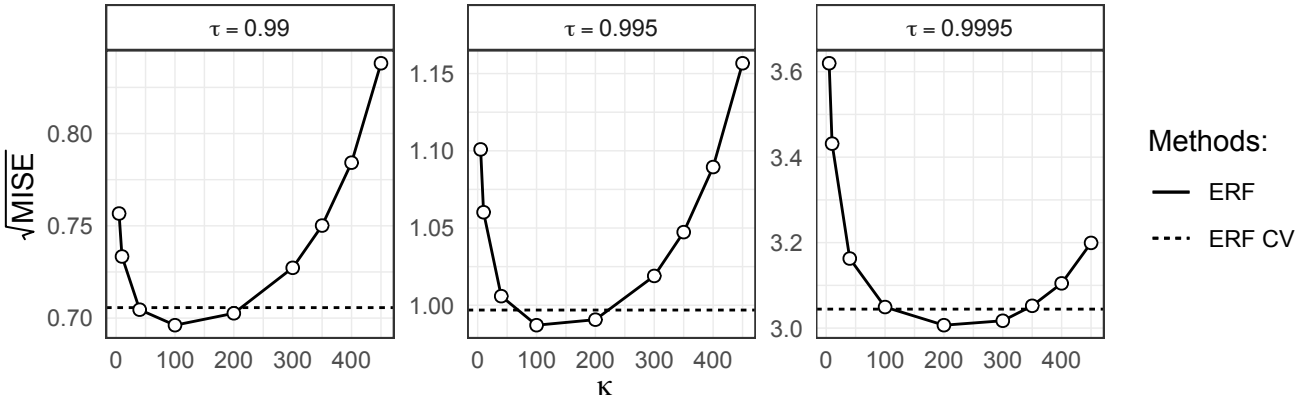


Figure 2.2: Solid line shows the square root of the MISE of ERF for different minimum node sizes κ over 50 simulations. The dashed line shows the square root MISE of the cross-validated ERF. The data is generated according to Example 2.1.

where $\theta \mapsto \ell_\theta(z)$ is the deviance of the GPD and $Z_i := (Y_i - \hat{Q}_{X_i}(\tau_0))_+$ are the exceedances. Finally, we select the optimal tuning parameter α^* as the minimizer of $CV(\alpha_j)$, $j = 1, \ldots, J$. To make the problem computationally tractable, we first fit the intermediate quantile function $x \mapsto \hat{Q}_x(\tau_0)$ on the entire data set. Then, on each fold, we estimate the similarity weight function $(x, y) \mapsto w_n(x, y)$ with "small" forests made of 50 trees. We repeat the cross-validation scheme several times to reduce the variability of the results.

Even though, in principle, one could perform cross-validation on several tuning parameters, we find that the minimum node size $\kappa \in \mathbb{N}$ plays the most critical role for ERF. The reason is that κ controls the model complexity of the individual trees in the forest and consequently of the similarity weights $w_n(\cdot, \cdot)$. Small (large) values of κ correspond to trees with few (many) observations in each leaf and produce strongly (weakly) localized weight functions $w_n(\cdot, \cdot)$. The estimates of the shape parameter $\hat{\xi}(x)$ in (2.3.4) may be sensitive to small changes of the localizing weights in the covariate space, leading to unstable quantile predictions through (2.1.1). To reduce the variance of $\hat{\xi}(x)$, it is useful to stabilize the log-likelihood $x \mapsto L_n(\theta; x)$ by estimating the similarity weights $w_n(\cdot, \cdot)$ with a forest made of trees with relatively large leaves. Notice that $w_n(x, y)$ influences the effective number of observations used in the weighted (negative) log-likelihood $L_n(\theta; x)$ (2.3.3).

Figure 2.2 shows numerical results of cross-validating the minimum node size κ for the model described in Example 2.1. Here, we perform 5-fold cross-validation repeated three times by growing forests of 50 trees on each fold. We measure the performance as the square root of the mean integrated squared error (MISE) between the estimated and the true quantile function over 50 simulations; see Section 2.4 for the definition of the MISE. We observe that the cross-validated performance of ERF (dashed line) is close to the minimum square root MISE, suggesting that the proposed cross-validation scheme works well.

2.3.4 Penalized Log-Likelihood

The shape ξ of the GPD is the most crucial parameter since it determines the tail behavior of Y at extreme quantile levels; the extrapolation formula (2.2.3) shows the highly non-linear influence of the shape parameter on large quantiles.

Estimation of the shape parameter is notoriously challenging, and the maximization of the GPD likelihood may exhibit convergence problems for small sample sizes (Coles

and Dixon, 1999). In general, penalization can help to reduce the variance of an estimator at the cost of higher bias (Hastie et al., 2009). Coles and Dixon (1999) propose a penalty function that restricts the shape parameter values to $\xi < 1$ and favors smaller values of ξ. Several penalization schemes can be interpreted in a Bayesian sense by considering a prior distribution on the regularized parameter. For example, de Zea Bermudez and Turkman (2003) introduce a Bayesian approach to estimate the ξ by using different priors for the cases $\xi > 0$ and $\xi < 0$, respectively. In the context of the generalized extreme value distribution, other penalization methods have been proposed by Smith and Naylor (1987).

While the above regularization methods are tailored to i.i.d. data, in our setting we want to penalize the variation of the shape function $x \mapsto \xi(x)$ across the predictor space $\mathcal{X}$. In spatial applications, for instance, it is common to assume a constant shape parameter at different locations (e.g., Ferreira et al., 2012; Engelke et al., 2019). Similarly, in ERF we shrink the estimates $\hat{\xi}(x)$ to a constant shape parameter ξ_0. In general, ξ_0 can be given by expert knowledge, but often a good choice is the unconditional fit $\xi_0 = \hat{\xi}$ obtained by minimizing the GPD deviance in (2.3.3) with constant weights $w_n(x, y) = 1$ for all $x, y \in \mathcal{X}$.

We propose to penalize the weighted GPD deviance (2.3.3) with the squared distance between the estimates of $\xi(x)$ and the constant shape parameter ξ_0, that is,

$$\hat{\theta}(x) = \underset{(\sigma,\xi)=\theta\in\Theta}{\arg\min} \frac{1}{(1-\tau_0)} L_n(\theta; x) + \lambda(\xi - \xi_0)^2, \tag{2.3.5}$$

where $\lambda \geq 0$ is a tuning parameter, and τ_0 is the intermediate quantile level. The parameter λ allows interpolating between a simpler model with a constant shape when $\lambda \to \infty$, and a more complex model with a varying shape over the predictor space when λ is small. This penalized negative log-likelihood can be interpreted in a Bayesian sense: it is equivalent to the maximum *a posteriori* GPD estimator when putting Gaussian prior $N(\xi_0, 1/(2\lambda))$ on the shape parameter ξ. Bücher et al. (2020) propose the same penalization as in (2.3.5) to estimate the generalized extreme value distribution parameters, where the prior distribution is centered around an expert belief ξ_0 and $\lambda \geq 0$ reflects the confidence in such belief.

In practice, when we penalize the shape parameter we modify Algorithm 2 by replacing Line 3 of the ERF-Predict subroutine with (2.3.5). Similarly, we cross-validate λ using the scheme presented in Section 2.3.3 on the modified Algorithm 2. Figure 2.3 shows the square root MISE over 50 simulations for different values of λ and different quantile levels. Here, we set ξ_0 as the estimated unconditional shape parameter.

2.4 Simulation Study

2.4.1 Setup

We compare ERF to other quantile regression methods on simulated data sets, assessing the properties of the different approaches. In the three experiments, we simulate n training observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ as independent copies of a random vector (X, Y). We always generate the predictor $X \in \mathbb{R}^p$ from a uniform distribution on the cube $[-1, 1]^p$ for different dimensions p. We let the conditional response variable $Y \mid X = x$ follow distributions such as Gaussian or Student's t, with tail heaviness depending on the

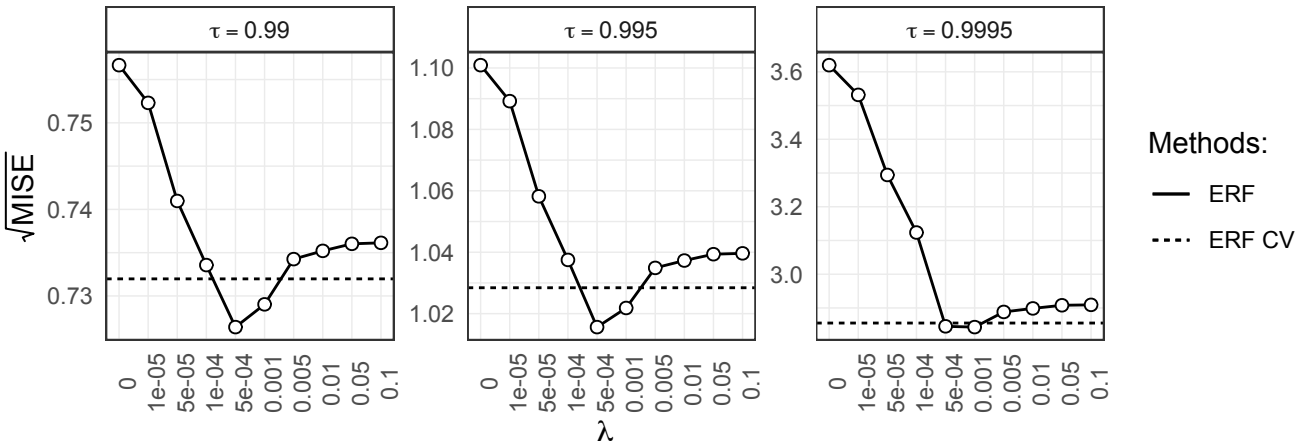


Figure 2.3: Square root MISE of ERF for different penalty values λ and quantile levels τ over 50 simulations. The data is generated according to Example 2.1.

simulation study. The parameters of these distributions, and therefore also the parameters of the GPD corresponding to their tails, vary as functions of the predictor value x. Different response surfaces are considered. The goal is to predict the quantiles $Q_x(\tau)$ of the conditional response $Y \mid X = x$ for moderately to very extreme quantile levels $\tau > 0$.

We evaluate the performance of the method on a test data set $\{x_i\}_{i=1}^{n'}$ of $n' = 1000$ observations generated with a Halton sequence (Halton, 1964) on the cube $[-1, 1]^p$. For a fitted quantile regression function $x \mapsto \hat{Q}_x(\tau)$, $\tau \in (0, 1)$, we then compute the integrated squared error (ISE) on the test data set as

$$\text{ISE} = \frac{1}{n'} \sum_{i=1}^{n'} \left(\hat{Q}_{x_i}(\tau) - Q_{x_i}(\tau) \right)^2,$$

where $x \mapsto Q_x(\tau)$ is the true quantile function of the model. Repeating the simulation, fitting and evaluation $m = 50$ times, we obtain the mean integrated squared error (MISE) as the average of the different ISEs.

In the first experiment, we study how ERF performs on the two challenges of high quantile levels and large, but fixed, dimensions of the predictor spaces illustrated in Figure 2.1. The data sets follow the model of Example 2.1 where the response has a Student's t-distribution with scale shift according to a step function. We consider the methods' performances for different dimensions p of the predictor space and different quantile levels τ.

The second experiment illustrates the performance of ERF and other methods under different tail heaviness of the noise distribution. The data generating function is the same as in the first experiment, except that the tail of the noise ranges from the light-tailed Gaussian case with $\xi = 0$ to the relatively heavy tails of Student's t distributions with large $\xi > 0$.

In the last experiment (see Appendix B.4.2), we consider more complex regression functions for the conditional response variables to assess the performance of the quantile regression methods on complex data. The underlying models depend on more than one predictor value, and both the scale and the shape parameters vary simultaneously.

2.4.2 Competing Methods

Among the forest-based algorithms, we consider the quantile regression forest by Meinshausen (2006), denoted by QRF, and the generalized random forest by Athey et al.

(2019), denoted by GRF. Since these methods do not rely on the GPD likelihood, it is not possible to cross-validate their tuning parameters as in Section 2.3.3 for prediction error of extreme quantiles. However, in independent simulations, we notice that their tuning parameters do not have big influence on the results. We set their tuning parameters to the default values and and fit the quantile functions $\hat{Q}_x^{QRF}(\tau)$, $\hat{Q}_x^{GRF}(\tau)$ on the training data for some $\tau \in (0,1)$. More details on forest-based approaches can be found in Section 2.2.2.

As a hybrid method that uses forest-based weights, we consider the method EGP Tail proposed by Taillardat et al. (2019) who assume that the entire conditional distribution $Y \mid X = x$ follows a parametric family called extended generalized Pareto (EGP) distribution. They estimate the covariate dependent parameters of the EGP through a probability-weighted method of moments using the estimated quantiles $\hat{Q}_x^{GRF}(\tau)$ of the GRF. We follow the authors' implementation and use the default parameter values.

Our ERF method is part of the class of extrapolation approaches that model the exceedances Z_i in (2.3.2) by conditional GPD distributions. Among the numerous methods that follow this strategy we present only those from Youngman (2019) and Velthoen et al. (2021) as they turn out to be most competitive. Other existing extrapolation based methods are not flexible enough in our setting (Wang and Tsai, 2009; Wang et al., 2012) or do not perform well with larger noise dimensions (Daouia et al., 2011; Gardes and Stupfler, 2019). For the sake of comparability, for all extrapolation methods we use the same exceedances $Z_i = (Y_i - \hat{Q}_x^{GRF}(\tau_0))_+$, which are computed from a GRF with intermediate quantile level $\tau_0 = 0.8 \leq \tau$. To assess the sensitivity of our method to the intermediate threshold τ_0, we perform a simulation study for a data set generated according to Example 2.1; see Figure B.2 in Appendix B.4.1. In this setup, the intermediate threshold does not strongly influence the results. In general, the optimal choice will depend on the properties of the data (see de Haan and Ferreira, 2006, Section 3.2) and numerous data-driven methods for choosing the threshold exist (e.g., Embrechts et al., 2012, Section 6.2.2).

The method from Youngman (2019) uses generalized additive models to estimate the parameters of a GPD distribution. Here, we model the scale and shape parameters as smooth additive functions of the covariates without interaction effects. In the sequel, we abbreviate this method by EGAM. Velthoen et al. (2021) propose the GBEX method to estimate the GPD parameters using gradient boosting (Friedman, 2001b, 2002). In particular, they grow two sequences of gradient trees to model the conditional scale and shape parameter, respectively. To fit GBEX, we use 5-fold cross validation with a maximum number of trees per fold set to $B_{\max} = 500$. We set the depth of each gradient tree $D = 2$, and we set the learning rate for the scale parameter to $\lambda^\sigma = 0.1$. We set the other tuning parameters to their default values. We also consider the unconditional model as a baseline, where we fit constant GPD parameters (σ, ξ) to the conditional exceedances Z_i.

Concerning our ERF method, we fit the parameters as described in Algorithm 2 using the repeated cross-validation scheme described in Section 2.3.3. In particular we repeat three times 5-fold cross-validation to tune the minimum node size $\kappa \in \{10, 40, 100\}$ and the penalty $\lambda \in \{0, 0.01, 0.001\}$ for the shape parameter. We leave the other tuning parameters of the random forests at their default values; see the documentation for `quantile_forest` in Tibshirani et al. (2021). All simulation results can be reproduced following the description and code on https://github.com/nicolagnecco/erf-numerical-results.

2.4.3 Experiment 1

In this simulation study, the data follows the model of Example 2.1 where the response variable $Y \mid X = x$ follows a Student's t-distribution with $\nu(x) \equiv 1/\xi(x) = 4$ degrees of freedom and scale $s(x) = 1 + \mathbb{1}\{x_1 > 0\}$. This is the same setup as in the simulation in Athey et al. (2019, Section 5), except that here we use Student's t-distribution instead of Gaussian for the noise. There is only one signal variable X_1 and $p-1$ noise variables. We generate $n = 2000$ training data and consider different dimensions p and quantile levels τ.

We first fix the dimension $p = 10$ and investigate the effect of different target quantile levels τ on the prediction performances of the competing methods. The top panel of Figure 2.4 shows the MISE, as defined in Section 2.4.1, for varying values of τ close to 1, and its decomposition in terms of bias and variance. At the intermediate quantile level $\tau_0 = 0.8$ all methods show a similar performance; in fact, the extrapolation methods coincide at this level since they use the same GRF based estimator for the intermediate quantile. When the quantile level τ increases, or equivalently, the expected number of exceedances $n(1-\tau)$ in the training sample decreases, we observe that the performance curves diverge. The forest-based quantile regression methods that do not explicitly use extreme value theory for tail approximations cannot extrapolate well to extreme quantile levels. This includes the EGP Tail method that does not focus on modeling the tail. Their degradation in performance is mainly driven by high variance. Among the extrapolation methods, the unconditional baseline does not perform well since it cannot capture the shift in the scale function and therefore it presents a high bias. While the EGAM does better, it shows a pretty large bias already in this setup with ten predictors, a fact that we discuss in detail below. By far, the best methods are our ERF and the GBEX. Both combine the flexibility in the predictor space with correct extrapolation originating from the GPD approximation.

We next compare the performances for varying dimensions p of the predictor space. The bottom panel of Figure 2.4 shows the MISE as a function of p for fixed quantile levels $\tau = 0.9995$, and its bias and variance decomposition. QRF and GRF look relatively robust against growing dimensions and additional noise variables, but the performance is not competitive for higher quantiles levels. As before, their performance is mainly explained by large variance. For smaller dimensions, the methods deteriorate because of the overfitting; the trees can only place split on the signal variable X_1, increasing the variance. The performance of EGAM clearly illustrates that this method suffers from high bias in large dimensions. The method cannot filter the signal from the many noise variables even though, in principle, it is flexible enough to model the response function; the latter is indicated by the good performance for very small noise dimension. Moreover, as mentioned by Youngman (2019), the method becomes computationally demanding as p grows. The unconditional model, while biased, has constant performance across different dimensions since it does not use the predictor values. Both ERF and GBEX combine the advantages of the two types of approaches. They are both robust against additional noise variables and perform well even for large dimensional predictor spaces.

2.4.4 Experiment 2

In the second experiment, we illustrate the performance of the quantile regression methods under several tail heaviness of the noise distribution in a large dimension. The simulation setup is similar to the previous section and data follows the model of Example 2.1, where

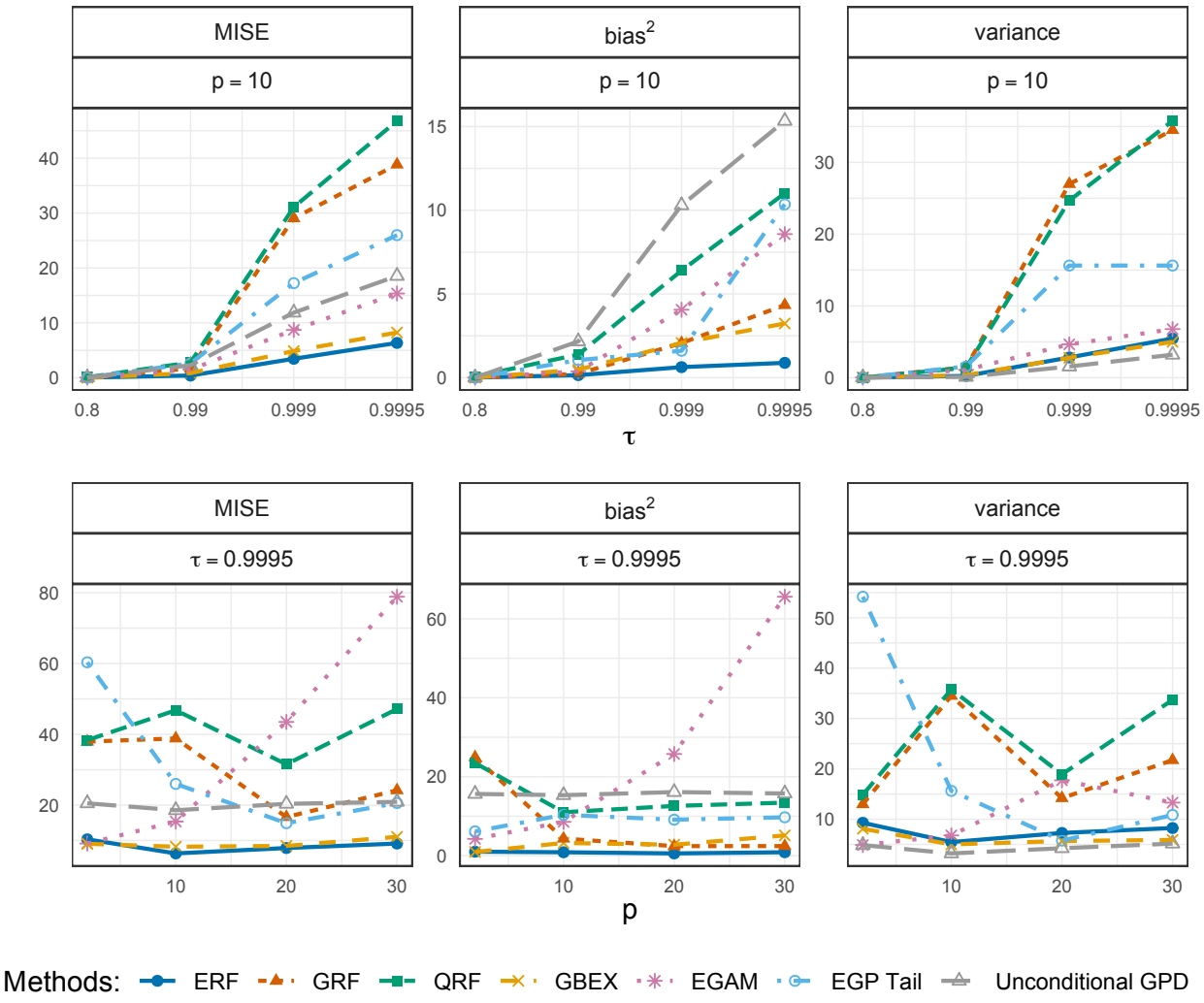


Figure 2.4: Square root MISE for different methods against the quantile level τ in dimension $p = 10$ (left), and against the model dimension p for quantile levels $\tau = 0.9995$ (right).

we set $p = 40$. We simulate data for noise distributions with shape parameters $\xi = 0, 1/4, 1/3$, where for the light-tailed case $\xi = 0$ we choose a Gaussian distribution and otherwise a Student's t distribution with $\xi = 1/4, 1/3$ corresponding $v = 4, 3$ degrees of freedom, respectively. We exclude EGAM in this experiment since its performance decreases for large p and it becomes computationally prohibitive (see Figure 2.4).

Figure 2.5 shows boxplots of the $\sqrt{\text{ISE}}$ for the extreme quantile level $\tau = 0.9995$ for the different methods and different shape parameters. The triangles correspond to the average values. To make the plot easier to visualize, we remove large outliers of GRF and QRF. The picture is similar for the three noise distributions. We observe that ERF performs very well also in the Gaussian case. Since our method relies on the GPD, estimation is not restricted to positive shape parameters, as opposed to approaches based on the Hill estimator (e.g., Wang et al., 2012; Wang and Li, 2013). Unsurprisingly, as the noise becomes very heavy-tailed (right-hand side of Figure 2.5) the performances of all methods become closer since the problem becomes increasingly difficult. We further note that the performance of both QRF and GRF degrades for large values of ξ. They exhibit increasingly large outliers that result in an average exceeding the upper quartile.

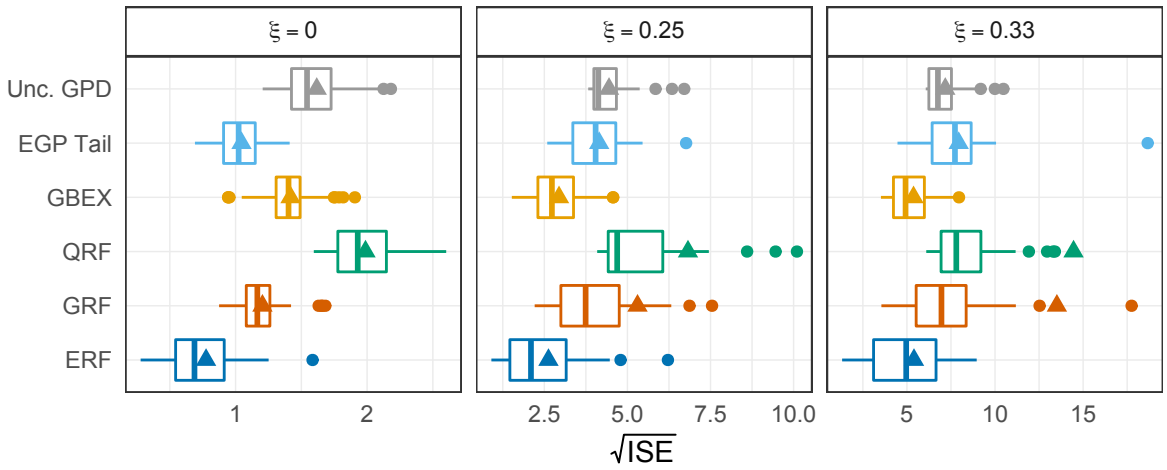


Figure 2.5: Boxplots of $\sqrt{\text{ISE}}$ over $m = 50$ simulations, for different tail indices in the noise distribution at the quantile level $\tau = 0.9995$. The predictor space dimension is $p = 40$. Triangles represent the average values.

This underlines that classical methods without proper extrapolation are insufficient for extreme quantile regression.

2.5 Analysis of the U.S. Wage Structure

We compare the performance of ERF, GBEX, GRF, and the unconditional GPD on the U.S. census microdata for the year 1980 (Angrist et al., 2009). As described therein, the data set consists of 65,023 U.S.-born black and white men of age between 40–49, with five to twenty years of education, and with positive annual earnings and hours worked in the year before the census. The large number of observations makes this dataset suitable to assess the performance of the different methods at very high quantile levels. The response Y describes the weekly wage, expressed in 1989 U.S. dollars computed as the annual income divided by the number of weeks worked. The predictor vector consists of the numerical variables age and years of education and the categorical predictor whether the person is black or white. To make the data set higher dimensional, we add ten random predictors sampled independently from uniform distributions on the interval $[-1, 1]$, resulting in a predictor space's dimension $p = 13$.

Throughout this analysis, we fit ERF repeating three times 5-fold cross-validation to tune the minimum node size $\kappa \in \{5, 40, 100\}$. To stabilize the variance of the shape parameter, we set the penalty $\lambda = 0.01$. Regarding the other methods, we use the same tuning parameter setup as in 2.4.2. In particular, we use GRF to predict the intermediate conditional quantiles at level $\tau_0 = 0.8$ for all extrapolation-based methods. We split the original data into two halves, i.e., 32,511 and 32,512 samples, respectively. We use the first portion to perform an exploratory data analysis and the second one to fit and evaluate the different methods.

For the exploratory data analysis, we fit ERF on a random subset made of 10% of the data (i.e., 3,251 observations), and predict the GPD parameters $\hat{\theta}(x) = (\hat{\sigma}(x), \hat{\xi}(x))$ on the left-out observations (i.e., 29,260 observations). Figure 2.6 shows the estimated GPD parameters $\hat{\theta}(x)$ as a function of years of education. We observe that the scale parameter depends positively on years of education, whereas it is quite homogeneous between the

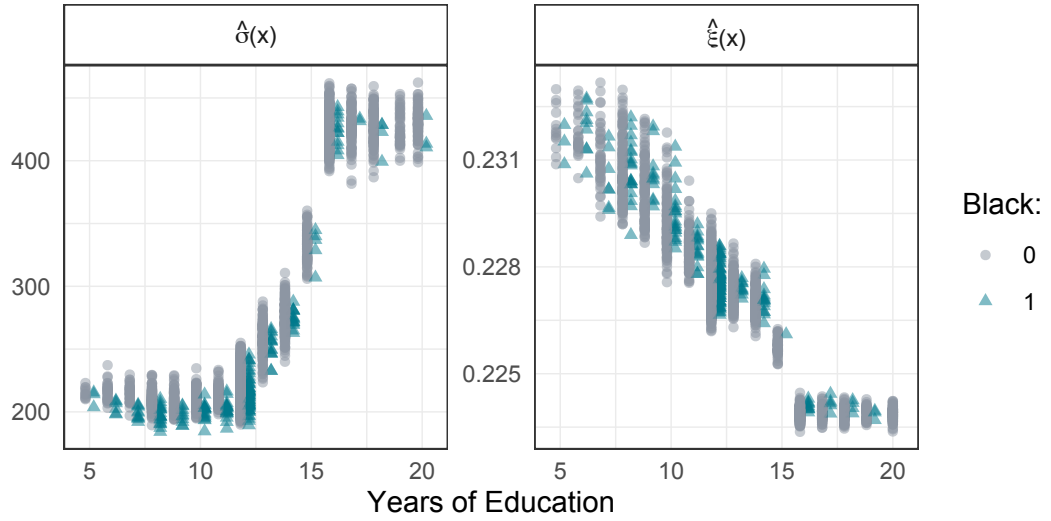


Figure 2.6: Estimated GPD parameters $\hat{\theta}(x)$ as a function of the years of education for the black (triangles) and white (circles) subgroups.

black and white groups. In particular, it has a clear jump around 15-16 years of education, which corresponds to the end of the undergraduate studies. The shape parameter is relatively homogeneous for the black and white group and looks stable for education. It ranges between 0.22 and 0.24, indicating heavy-tails throughout the predictor space. Moreover, Figure B.4 in Appendix B.5.1 shows that the scale and shape parameters do not seem to depend on the predictor age.

In Figure 2.7 we compare the ERF quantile predictions to the ones obtained by the other methods at levels $\tau = 0.9, 0.995$. To help with the visualization, we removed all the quantiles above 6,000 predicted by GRF. We observe that the extrapolation methods retain a good shape of the quantile function even for high levels. This does not hold for GRF, whose profile worsens as τ increases, and the discrete structure of the largest training observations becomes visible. The unconditional method seems to capture the variability of the conditional quantiles for $\tau = 0.9$, but we observe that it loses flexibility for larger values of τ. The reason for this is that the unconditional method cannot produce different scale parameters of the GPD, while Figure 2.6 indicates that this is necessary for this data set. ERF and GBEX model well the variability of the conditional quantiles for all values of τ, and they agree on the magnitude of the estimates.

After the exploratory analysis, we assess the quantitative performance of ERF compared to the other methods. We consider the prediction metric proposed by Wang and Li (2013),

$$\mathcal{R}_n\left(\hat{Q}_\cdot(\tau)\right) := \frac{\sum_{i=1}^n \mathbb{1}\{Y_i < \hat{Q}_{X_i}(\tau)\} - n\tau}{\sqrt{n\tau(1-\tau)}}, \tag{2.5.1}$$

where n is the number of test observations, and $\hat{Q}_\cdot(\tau)$ is the τ-th conditional quantile estimated on the training data set. This metric compares the normalized estimated proportion of observations with $Y_i < \hat{Q}_{X_i}(\tau)$ with the theoretical level τ. Using the true quantile function $Q_\cdot(\tau)$, the random variable $\mathbb{1}\{Y_i < Q_{X_i}(\tau)\}$ follows a Bernoulli distribution with expectation τ and variance $\tau(1-\tau)$, and by the central limit theorem the metric with oracle quantile function $\mathcal{R}_n(Q_\cdot(\tau))$ is asymptotically standard normal. We partition the 32,512 observations not used in the exploratory analysis into ten random folds. On

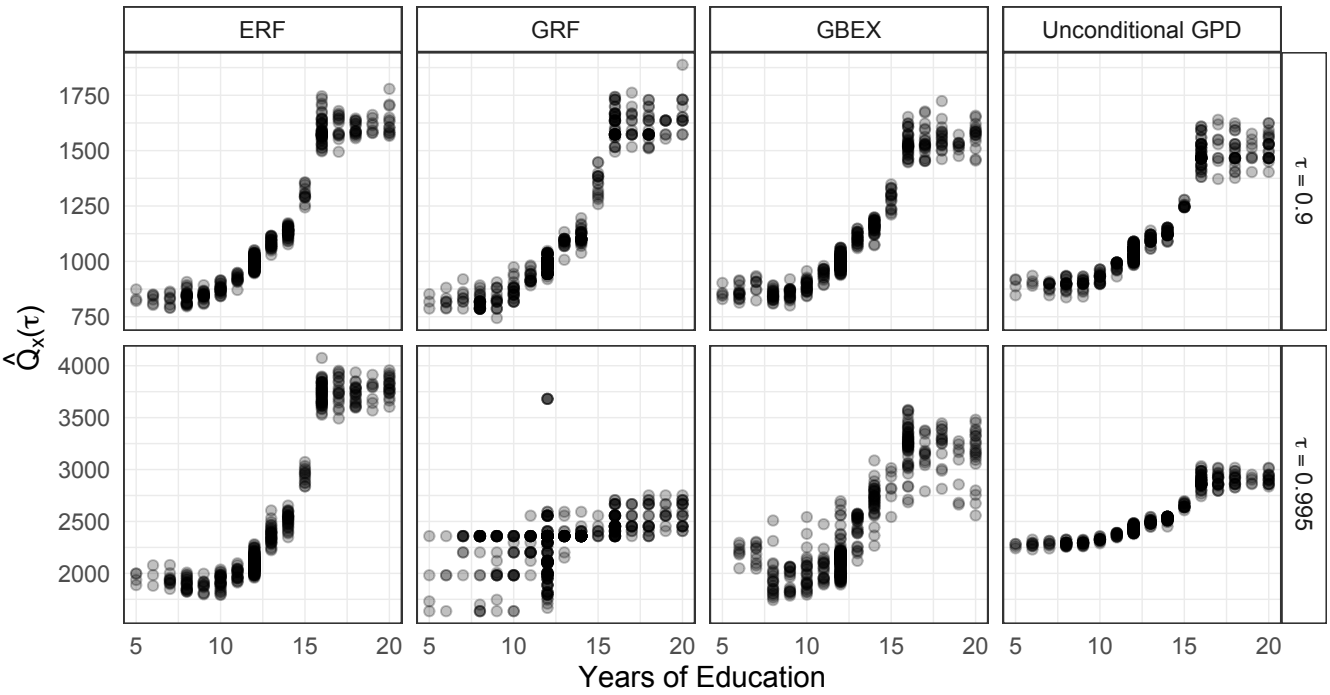


Figure 2.7: Predicted quantiles at levels $\tau = 0.9, 0.995$ for ERF, GRF, GBEX, and the unconditional method.

each fold, we fit the different methods and evaluate them on the left-out observations, using the absolute value of (2.5.1). Unlike classical cross-validation, we fit the methods using a single fold and validate them on the remaining ones; this allows us to have enough observations to gauge their performance for high quantile levels τ. Figure 2.8 shows the performance of ERF, GRF, GBEX, and the unconditional method over the ten repetitions for different quantile levels. The shaded area represents the 95% interval of the absolute value of a standard normal distribution, corresponding to the 95% confidence level of the oracle method with true quantile function. We observe that both ERF and GBEX have very good performance compared to the oracle for increasing quantile levels, and they outperform the unconditional method for large values of τ. This is because they are flexible to model the scale and shape as a function of the predictors, unlike the unconditional method. While GRF performs well for the quantile level $\tau = 0.9$, it worsens quite quickly for larger values of τ. This is expected since GRF does not rely on extrapolation results from extreme value theory and cannot accurately predict very high quantiles.

For the same data set, Angrist et al. (2006) consider the natural logarithm of the wage as a response variable for quantile regression with fixed, non-extreme quantile levels. In Appendix B.5.1 we perform our analysis above for extreme quantiles again with this log-transformed response since it highlights several interesting properties of the ERF algorithm. In particular, Figure B.6 in Appendix B.5.2 shows that the flexible methods ERF and GBEX have the desirable property that the predictions do not change much under marginal transformations. The unconditional method, on the other hand, seems to be sensitive to marginal transformations; for an explanation and details, see Appendix B.5.1. In general, therefore, it is advised to use a flexible extrapolation method, such as ERF or GBEX, that performs well on any marginal distributions.

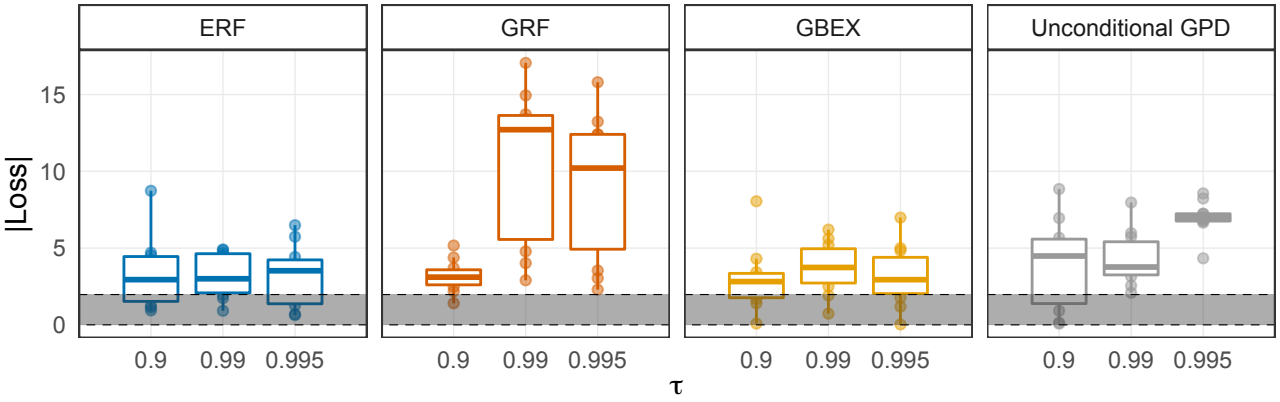


Figure 2.8: Absolute value of the loss (2.5.1) for the different methods fitted on the original response of the U.S. wage data. The shaded area represents the 95% interval of the absolute value of a standard normal distribution.

Chapter 3

Distribution generalization in semi-parametric models: A control function approach

Joint work with

Sebastian Engelke, Niklas Pfister, and Jonas Peters

Abstract

Distribution generalization aims at learning a function with predictive guarantees when the test distribution differs from the training. In this work, we adopt a causal approach to address the problem by modelling distributional shifts with causal interventions (Pearl, 2009a; Peters et al., 2017). We consider the problem of predicting a real-valued response when the data comes from different environments that shift the mean of the predictors. We assume the presence of hidden confounders and a possibly large dimensional predictor space. Our goal is to learn a nonparametric function that minimizes the worst-case mean squared error over unseen environments. Existing literature provides minimax guarantees when the function class is linear. However, in the case of nonlinear function classes, existing methods do not provide such guarantees or do not scale to large dimensions. Here, we propose a method to learn a function that has invariant predictions across environments and is as predictive as possible. We define such function as the invariant most predictive (IMP), and we show identification using control variables (Ng and Pinkse, 1995; Newey et al., 1999). Furthermore, we provide minimax guarantees over unseen environments over the class of square-integrable functions. Lastly, we propose an adaptation of the regression tree algorithm (Breiman et al., 1984) to learn the IMP function nonparametrically in large dimensions.

Keywords: distribution generalization; nonparametric regression; structural causal models; control functions; regression trees.

3.1 Introduction

Let $Y \in \mathbb{R}$ be a real valued response and $X \in \mathbb{R}^p$ a vector of mean-zero predictors. Suppose the data (X, Y) is generated from different environments $\mathcal{E}$ that induce a mean shift in the

predictors. That is, for all $e \in \mathcal{E}$ there exists some mean vector $\mu_e \in \mathbb{R}^p$ and independent noise term $V_e \sim \mathbb{P}_V$ such that $X_e = \mu_e + V_e$. For all $e \in \mathcal{E}$, denote by $\text{supp}(X_e)$ the support of X_e and define the support over the environments $\mathcal{E}$ as $\text{supp}(\mathcal{E}) := \cup\{\text{supp}(X_e) : e \in \mathcal{E}\}$. The span generated by the environments $\mathcal{E}$ is defined as $\text{span}(\mathcal{E}) := \text{span}(\{\mu_e : e \in \mathcal{E}\})$. At training time, we collect data (X_e, Y_e) with $e \in \mathcal{E}_{\text{tr}}$ for a subset of environments $\mathcal{E}_{\text{tr}} := \{1, \dots, r\} \subseteq \mathcal{E}$ that is "rich enough", i.e., $\text{span}(\mathcal{E}_{\text{tr}}) = \text{span}(\mathcal{E})$ and $\text{supp}(\mathcal{E}_{\text{tr}}) = \text{supp}(\mathcal{E})$. The goal of this work is to learn a nonparametric function $f^\diamond : \mathbb{R}^p \to \mathbb{R}$ that predicts well on all environments $\mathcal{E}$, i.e.,

$$f^\diamond := \arg\min_{f \in \mathcal{F}} \sup_{e \in \mathcal{E}} \mathbb{E}\left[(Y_e - f(X_e))^2\right], \tag{3.1.1}$$

where $\mathcal{F}$ is a given function class. One can think of $\mathcal{E}$ as the set of all possible environments that might occur at test time. Rothenhäusler et al. (2021) show how to learn $f^\diamond$ when $\mathcal{F}$ is the class of linear functions. Their idea is to exploit the heterogeneity of the observed environments $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}$ to learn a function $\tilde{f} : \mathbb{R}^p \to \mathbb{R}$ that (i) has invariant performance across the observed environments, and (ii) is as predictive as possible. Property (i) means that the function of interest $\tilde{f} \in \mathcal{F}$ satisfies

$$Y_e = \tilde{f}(X_e) + \tilde{U}_e, \text{ for all } e \in \mathcal{E}_{\text{tr}}, \tag{3.1.2}$$

where $\tilde{U}_e \sim \mathbb{P}_U$ is possibly correlated with V_e. Property (ii) means that the function of interest $\tilde{f} \in \mathcal{F}$ minimizes the mean squared prediction error (MSPE). When $\mathcal{F}$ consists of linear functions, Rothenhäusler et al. (2021) show that the proposed function $\tilde{f}$ minimizes the worst-case MSPE defined in (3.1.1). We extend the work of Rothenhäusler et al. (2021) to the case where $\mathcal{F}$ is the class of arbitrary functions with bounded second moment. We define the *invariant most predictive* (IMP) function (see Definition (3.2)), and characterize its identification with the control function approach of Ng and Pinkse (1995) and Newey et al. (1999). Also, we show that the IMP function is minimax optimal in the sense of (3.1.1). To learn the IMP function in large dimensions, we propose an adaptation of the regression tree algorithm (Breiman et al., 1984).

To motivate the analysis, we set up a simple example.

Example 3.1. Consider the problem of predicting a patient's blood pressure $Y \in \mathbb{R}$ based on the normalized age and weight predictors $X \in \mathbb{R}^2$. Suppose we collect an equal amount of data from two hospitals $\mathcal{E}_{\text{tr}} := \{-1, 1\}$, such that $X_e = \mu_e + V_e$, for all $e \in \mathcal{E}_{\text{tr}}$. Figure 3.1 shows the distribution of the predictors across the two environments. The hospitals means are $\mu_1 = (-1, -1)$ and $\mu_2 = (1, 1)$, respectively. Moreover, the support of the training environments, $\text{supp}(\mathcal{E}_{\text{tr}})$, is bounded. The heterogeneity in the two environments creates a mean shift in the $\text{span}(\mathcal{E}_{\text{tr}}) := \text{span}(\{\mu_1, \mu_2\}) \subseteq \mathbb{R}^2$. By exploiting the heterogeneity in the observed data, one can learn the IMP function $f_{\text{IMP}} : \mathbb{R}^2 \to \mathbb{R}$ that (i) has invariant performance across the two hospitals, and (ii) is as predictive as possible. In fact, f_{IMP} minimizes the worst-case MSPE (3.1.1) over the test environments $\mathcal{E}$, as long as $\text{span}(\mathcal{E}) = \text{span}(\mathcal{E}_{\text{tr}})$, and $\text{supp}(\mathcal{E}) = \text{supp}(\mathcal{E}_{\text{tr}})$. ◁

Why is $\mathbb{E}[Y \mid X = x]$ not enough? The regression function $x \mapsto \mathbb{E}[Y \mid X = x]$ minimizes the MSPE on the training observations, and therefore is the most predictive function on the observed data. However, when the predictors are endogenous (Wooldridge, 2010), e.g., they are correlated with the errors in the response, then $x \mapsto \mathbb{E}[Y \mid X = x]$ is not invariant. This means that the distribution of the residuals $Y - \mathbb{E}[Y \mid X]$ depends

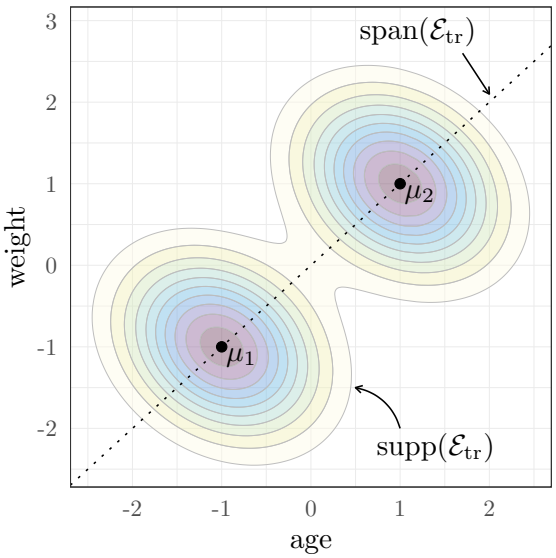


Figure 3.1: Distribution of the predictors across two environments. The support of the training environments $\text{supp}(\mathcal{E}_{\text{tr}})$ is bounded because the noise terms V_e have bounded support for all $e \in \mathcal{E}_{\text{tr}}$, and because $\mathcal{E}_{\text{tr}}$ is finite.

on the environments. In turn, this might lead to poor performance of the regression function under certain unseen environments in $\mathcal{E}$. On the other hand, the IMP function is the most predictive function among the invariant ones, and therefore achieves a good performance across all environments $\mathcal{E}$. Figure 3.2 compares the IMP function to the regression function $\mathbb{E}[Y \mid X = x]$ when the possible future environments are $\mathcal{E} = \{1, 2, 3\}$ and the training environments are $\mathcal{E}_{\text{tr}} = \{1, 2\}$, so that the unseen environment is $\mathcal{E} \setminus \mathcal{E}_{\text{tr}} = \{3\}$. The IMP function achieves a low and constant MSPE on all environments. In contrast, the regression function performs well on the training environments, but not on unseen environments.

Do invariant functions exist? Existence of an invariant function as in (3.1.2) is a modeling assumption that we make precise in Section 3.2. Uniqueness, instead, depends on the dimension of the predictor space and the heterogeneity induced by the training environments $\mathcal{E}_{\text{tr}}$. When the $X \in \mathbb{R}^p$ and $\text{span}(\mathcal{E}_{\text{tr}}) = \mathbb{R}^p$, there exists a unique invariant function (see Section 3.3); this relates to the identification condition in the instrumental variable setting (Angrist et al., 1996; Imbens, 2014). On the other hand, when the number of predictors is large, it is unlikely to observe enough environments such that $\text{span}(\mathcal{E}_{\text{tr}}) = \mathbb{R}^p$. In such a case, there exists a class of invariant functions $\mathcal{I}$, where each $f \in \mathcal{I}$ satisfies (3.1.2). Among these invariant functions, our goal is to identify and estimate the most predictive one (see Definition 3.2).

3.1.1 Related work

The problem of predicting in new unseen environments is of great importance in several applications, and it has been named out-of-distribution (OOD) generalization in the machine learning and statistics community. The goal of OOD generalization is to learn a predictive function when the training and test distributions are different (Quiñonero-Candela et al., 2009). In the machine learning literature, one well-established approach is to assume that the unseen environments are sampled from a distribution that is 'close' to

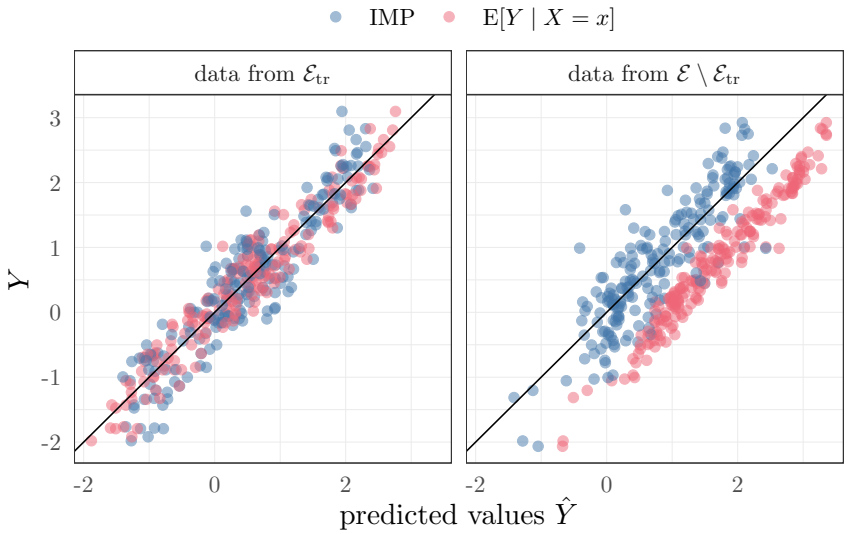


Figure 3.2: Illustrative example comparing the IMP function to the regression function $x \mapsto \mathbb{E}[Y \mid X = x]$. The IMP function achieves a low and constant MSPE on all environments. In contrast, the the regression function performs well on the training environments, but not on unseen environments.

the observed distribution with respect to the Wasserstein distance (Abadeh et al., 2015; Sinha et al., 2018). Recently, Meinshausen (2018); Bühlmann (2020); Rothenhäusler et al. (2021); Christiansen et al. (2021) cast the problem of OOD generalization under a causal perspective, where the shifts in distributions are generated by causal interventions (Pearl, 2009a; Peters et al., 2017). Rothenhäusler et al. (2021) consider a linear instrumental variable (IV) setup, where the causal function is possibly not identified and the instruments can act directly on the response variable. They introduce the anchor regression approach, which interpolates between the ordinary least squares (OLS) and the IV solution, and they show minimax guarantees as in (3.1.1) when $\mathcal{F}$ is the class of linear functions. Bühlmann (2020) extends anchor regression to a nonlinear setting, where $\mathcal{F}$ consists of nonlinear functions and the environments have a nonlinear effect on the predictors. However, it is not clear whether the proposed approach has the minimax guarantees as in (3.1.1). Christiansen et al. (2021) introduce the NILE algorithm, which extends anchor regression to the class of smooth functions that extrapolate linearly outside the observed support. While their algorithm is competitive in experiments, it has no minimax guarantees and, in practice, it works only when the predictor space has small dimensions (e.g., $p = 1, 2$).

3.2 Setup

Unless specified otherwise, we consider the following structural causal model (SCM)

$$\begin{aligned} E &:= \epsilon_E, \; V := \epsilon_V, \; U := \gamma^T V + \epsilon_U \\ X &:= ME + V \\ Y &:= f(X) + U, \end{aligned} \tag{3.2.1}$$

where $(\epsilon_E, \epsilon_U, \epsilon_V) \sim Q$ are jointly independent noise variables, $E \in \mathbb{R}^r$ is a vector that encodes the environments, $X \in \mathbb{R}^p$ are predictors, and $Y \in \mathbb{R}$ is a response. We assume that $\mathbb{E}[\epsilon_E] = 0$, $\mathbb{E}[\epsilon_E \epsilon_E^T] \succ 0$, and ϵ_U, ϵ_V are standard Gaussian. We denote this model by

$\mathcal{C} = (f, M, \gamma, Q)$, where $f \in \mathcal{F} := \{f : \mathbb{R}^p \to \mathbb{R} : \int f(X)^2 \, d\mathbb{P}_X < \infty\}$, $M : \mathbb{R}^{p\times r} \to \mathbb{R}^p$ is full column rank, and $\gamma \in \mathbb{R}^p$. Any model $\mathcal{C}$ as defined in (3.2.1), induces a distribution $\mathbb{P}_{\mathcal{C}}$ over the observed variables (E, X, Y). In addition to the observational distribution, we only consider the distributions arising from hard interventions on the environment variable, i.e., $\text{do}(E := e)$. We denote the data generating model by $\mathcal{C}_0 = (f_0, M_0, \gamma_0, Q_0)$, and we refer to the function f_0 as the *target function*. A comment is in order: model $\mathcal{C}$ in (3.2.1) does not describe the distributions induced by interventions on X or Y, and so, the function f can differ from the causal function between X and Y (see Remark 3.3).

To better understand model $\mathcal{C}$, we consider three examples. First, we illustrate how we can rewrite Example 3.1 within the framework of the SCM $\mathcal{C}_0$.

Remark 3.1. Consider again the problem of predicting a patient's blood pressure $Y \in \mathbb{R}$ based on the age and weight predictors $X \in \mathbb{R}^2$. We collect an equal amount of data from two hospitals, with $\mu_1 = (-1, -1)$ and $\mu_2 = -\mu_1$, so that $X_e = \mu_e + V_e$ and $V_e \sim \mathbb{P}_V$, $e \in \mathcal{E}_{\text{tr}} = \{-1, 1\}$. Let $E \in \{1, -1\} \subseteq \mathbb{R}$ be a random variable encoding the two environments, with $\mathbb{P}(E = 1) = 1/2$. Furthermore, denote by $M_0 = \mu_1 \in \mathbb{R}^{2\times 1}$ the matrix whose image $\text{im}(M_0) = \text{span}(\{\mu_1, \mu_2\})$. Then, the predictor vector can be written as $X = M_0 E + V$, with $V \sim \mathbb{P}_V$. ◁

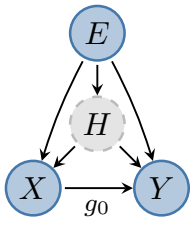
Our SCM $\mathcal{C}_0$ is also related to the instrumental variable setup (Angrist et al., 1996; Imbens, 2014).

Remark 3.2. Let $E \in \mathbb{R}^r$ denote a vector of r valid instruments, and let $f_0 : \mathbb{R}^p \to \mathbb{R}$ denote the causal function from X to Y. Then our SCM $\mathcal{C}_0$ is similar to a semi-parametric instrumental variable model, where the relationship is linear between E and X, and nonparametric between X and Y. Unlike the instrumental variable setup, however, here we do not require identification of the causal function, since we allow the number of instruments to be less than or equal to the number of predictors, i.e., $r \leq p$. Moreover, we do not require the instruments to be valid, in the sense that E can also directly affect the response variable. If this is the case, $f_0 : \mathbb{R}^p \to \mathbb{R}$ is not anymore the causal function from X to Y (see Remark 3.3). ◁

Finally, we provide an example of an SCM $\tilde{\mathcal{C}}$ that is equivalent to $\mathcal{C}_0$ under the observational distribution and under interventions $\text{do}(E := e)$.

Remark 3.3. Consider the following SCM $\tilde{\mathcal{C}}$ and related directed acyclic graph (DAG),

$$\begin{aligned} E &:= \epsilon_E \in \mathbb{R}^r, \\ H &:= B_{HE}E + \epsilon_H \in \mathbb{R}^q, \\ X &:= B_{XE}E + B_{XH}H + \epsilon_X \in \mathbb{R}^p, \\ Y &:= g_0(X) + B_{YE}E + B_{YH}H + \epsilon_Y \in \mathbb{R}, \end{aligned} \tag{3.2.2}$$



where $(\epsilon_E, \epsilon_H, \epsilon_X, \epsilon_Y) \sim N(0, \Sigma)$ such that Σ is positive definite and $\epsilon_E, \epsilon_H, \epsilon_X, \epsilon_Y$ are jointly independent. Here, $g_0 : \mathbb{R}^p \to \mathbb{R}$ denotes the causal function from X to Y, $B_{\cdot\cdot}$ are linear maps of suitable sizes, and $H \in \mathbb{R}^q$ is a vector of hidden confounders. Model (3.2.2) corresponds to the nonlinear anchor regression setup with a nonparametric causal function (Rothenhäusler et al., 2021; Bühlmann, 2020). If we only allow for interventions on E, then (3.2.2) is observationally and interventionally equivalent (when considering

interventions on E, X, and Y) to

$$\begin{aligned} X &:= M_0 E + V, \\ Y &:= g_0(X) + \eta_0^T E + B_{YH}\epsilon_H + \epsilon_Y, \end{aligned}$$

where $V := B_{XH}\epsilon_H + \epsilon_X \perp\!\!\!\perp E$, $M_0 := B_{XE} + B_{XH}B_{HE}$, and $\eta_0^T := B_{YE} + B_{YH}B_{HE}$. Since $M_0 \in \mathbb{R}^{p\times r}$ has full column rank, we can express $E = M_0^+(X - V)$, where M_0^+ denotes the Moore–Penrose inverse of M_0. Moreover, define $f_0(X) := g_0(X) + \eta_0^T M_0^+ X$, and $U := B_{YH}\epsilon_H + \epsilon_Y - \eta_0^T M_0^+ V$. Then, we can write

$$\begin{aligned} X &:= M_0 E + V, \\ Y &:= f_0(X) + U, \end{aligned}$$

with $E \perp\!\!\!\perp (U, V)$. In this example, the target function f_0 differs from the causal function g_0 because the environment variable E directly affects the response Y. ◁

3.3 Invariant most predictive function

Our aim is to learn a function $f : \mathbb{R}^p \to \mathbb{R}$ that makes good predictions on unseen environments. That is, by using the causal framework of Section 3.2, we rewrite our target of inference in (3.1.1) as

$$f^\diamond := \arg\min_{f\in\mathcal{F}} \sup_{e\in\mathbb{R}^r} \mathbb{E}[(Y - f(X))^2 \mid \mathrm{do}(E := e)], \tag{3.3.1}$$

where the expectation is taken with respect to the interventional distribution induced by $\mathrm{do}(E := e)$. In words, the function $f^\diamond$ minimizes the worst-case risk over any hard intervention on the environment vector $E \in \mathbb{R}^r$. The right-hand side of (3.3.1) involves evaluating the MSPE for all possible interventional distributions, and therefore it cannot be evaluated directly. Therefore, we propose to identify from the observational distribution $\mathbb{P}_{\mathcal{C}_0}$ the invariant most predictive model (IMP) and show that is a solution for (3.3.1).

First, we introduce the class of invariant functions. These functions are natural candidates to solve (3.3.1) since their residuals do not depend on the environment vector E.

Definition 3.1. *Consider an SCM $\mathcal{C}_0$ as defined in* (3.2.1)*. We define the class of invariant functions related to $\mathcal{C}_0$ as*

$$\mathcal{I} = \{f \in \mathcal{F} : Y - f(X) \perp\!\!\!\perp E\}. \tag{3.3.2}$$

While it is clear that $f_0 \in \mathcal{I}$, it is interesting to study under which conditions $\{f_0\} \subsetneq \mathcal{I}$. In fact, the cardinality $\mathcal{I}$ depends on the relationship between the dimension of predictors $X \in \mathbb{R}^p$ and the environment vector $E \in \mathbb{R}^r$. Consider the following example.

Example 3.2. Consider again the setup of Remark 3.1. The two environments span a one-dimensional space $\mathrm{im}(M_0)$; see Figure 3.3. On the other hand, the predictor space is two-dimensional, and therefore, the orthogonal complement of $\mathrm{im}(M_0)$, denoted by $\ker(M_0^T)$, is one-dimensional. Furthermore,

$$\delta \in \ker(M_0^T) \implies \delta^T X = \delta^T M_0 E + \delta^T V = \delta^T V.$$

Therefore, any function $f(x) = f_0(x) + \delta^T x$, with $\delta \in \ker(M_0^T)$, is such that $Y - f(X) \perp\!\!\!\perp E$, and so $f \in \mathcal{I}$. ◁

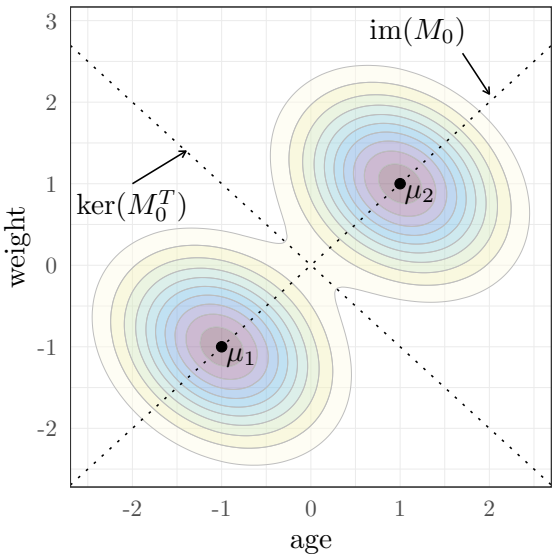


Figure 3.3: Distribution of the predictors across two different environments. The two environments shift the mean of the predictors along the one-dimensional $\text{im}(M_0)$. The orthogonal complement $\ker(M_0^T)$, by construction, is invariant to any shift in the direction of $\text{im}(M_0)$.

Having introduced the invariant set $\mathcal{I}$, we now define the invariant most predictive (IMP) function, i.e., the invariant function minimizing the MSPE.

Definition 3.2. *Consider an SCM $\mathcal{C}_0$ as defined in* (3.2.1)*. Let $\mathcal{I}$ denote the class of invariant functions related to $\mathcal{C}_0$. The invariant most predictive (IMP) function is defined as*

$$f_{\text{IMP}} := \arg\min_{f \in \mathcal{I}} \mathbb{E}\left[(Y - f(X))^2\right], \tag{3.3.3}$$

where the expectation is with respect to the observational distribution $\mathbb{P}_{\mathcal{C}_0}$.

Unlike (3.3.1), the IMP function in (3.3.3) can be identified and estimated from $\mathbb{P}_{\mathcal{C}_0}$. However, direct optimization over $\mathcal{I}$ is challenging. In principle, one could identify (3.3.3) by using econometric methods based on the generalized methods of moments (GMM) (Poirier, 2017) or by solving a non-linear ill-posed inverse problem (Dunker et al., 2014; Dunker, 2021). One limitation of these approaches, however, is that they estimate the target function "globally" over the predictor space $\mathbb{R}^p$. In other words, they minimize the mean squared error while enforcing at the same time invariance of the residuals. This "global" approach is not well suited for nonparametric regression techniques which, instead, learn the optimal function *locally*, e.g., at each point in the predictor space $x_0 \in \mathbb{R}^p$, they estimate $f(x_0) = \mathbb{E}[Y \mid X = x_0]$.

Remark 3.4. For an arbitrary function class $\mathcal{F}$, the zero-covariance condition (a) $\mathbb{E}[(Y - f(X))E] = 0$ used in linear instrumental variables and anchor regression (Greene, 2003; Rothenhäusler et al., 2021; Bühlmann, 2020) is not sufficient to enforce an invariant function $f \in \mathcal{I}$. Even the slightly stronger condition (b) $\mathbb{E}[Y - f(X) \mid E] = 0$ used in nonparametric instrumental variables (Newey and Powell, 2003) does not imply $f \in \mathcal{I}$. Next, we show an example where (a) and (b) hold but $f \notin \mathcal{I}$. ◁

Example 3.3. Consider an SCM $\mathcal{C}_0$, as defined in (3.2.1), with two-dimensional predictor vector $X \in \mathbb{R}^2$. Let $V \sim N(0, I_2)$, $U \sim N(0,1)$, and $E \in \{-1, 1\}$ follows a Rademacher distribution, with $E \perp\!\!\!\perp (U, V)$. Moreover, let $M_0 = (1,0)^T$. We can write $\mathcal{C}_0$ as

$$X_1 = E + V_1, \; X_2 = V_2, \; Y = f_0(X) + U.$$

Notice that $X_1 \perp\!\!\!\perp X_2$ and $X_1 \perp\!\!\!\perp X_2 \mid E$. Consider the function $f(x_1, x_2) = f_0(x_1, x_2) + x_1 x_2$ for all $(x_1, x_2) \in \mathbb{R}^2$. We have that $Y - f(X_1, X_2) = U - X_1 X_2$. Therefore,

$$\mathbb{E}[(Y - f(X))E] = \mathbb{E}[UE] - \mathbb{E}[X_1 X_2 E] = \mathbb{E}[U]\ \mathbb{E}[E] - \mathbb{E}[X_1 E]\ \mathbb{E}[X_2] = 0.$$

Also,

$$\mathbb{E}[Y - f(X) \mid E] = \mathbb{E}[U \mid E] - \mathbb{E}[X_1 X_2 \mid E] = \mathbb{E}[U] - \mathbb{E}[X_1 \mid E]\ \mathbb{E}[X_2] = 0.$$

However, even invariance of the first moment of $Y - f(X)$ is not enough. In fact,

$$\begin{aligned}\mathbb{E}[(Y - f(X))^2 \mid E] &= \mathbb{E}[U^2 + X_1^2 X_2^2 - 2U X_1 X_2 \mid E] \\ &= \mathbb{E}[U^2] + \mathbb{E}[X_1^2 \mid E]\ \mathbb{E}[X_2^2] - 2\mathbb{E}[U]\ \mathbb{E}[X_1 X_2 \mid E] \\ &= 1 + \mathbb{E}[E^2 + V_1^2 + 2EV_1 \mid E] = 2 + E^2,\end{aligned}$$

and thus $f \notin \mathcal{I}$. ◁

Given the difficulty of directly optimizing (3.3.3), we consider a subset $\mathcal{J} \subseteq \mathcal{I}$ that is simpler to deal with, and can be estimated with nonparametric regression techniques.

Definition 3.3. *Consider an SCM $\mathcal{C}_0$ as defined in* (3.2.1)*. We define the class of* linear *invariant functions related to $\mathcal{C}_0$ as*

$$\mathcal{J} := \left\{ f \in \mathcal{F} : \text{there exists } \delta \in \ker(M_0^T) \text{ s.t. } f(x) = f_0(x) + \delta^T x \text{ for all } x \in \mathbb{R}^p \right\}. \tag{3.3.4}$$

The set $\mathcal{J}$ is a subset of $\mathcal{I}$ and contains all functions that differ from f_0 by an element $\delta \in \ker(M^T)$. The following proposition shows that $\mathcal{J}$ is a proper subset of $\mathcal{I}$.

Proposition 3.4. *Consider an SCM $\mathcal{C}_0$ as defined in* (3.2.1)*. Let $\mathcal{I}$ and $\mathcal{J}$ denote the classes of invariant and linear invariant functions related to $\mathcal{C}_0$, respectively. Then, it holds that $\mathcal{J} \subsetneq \mathcal{I}$.*

A proof can be found in Appendix C.1.1. By Proposition 3.4, it is clear that,

$$\min_{f \in \mathcal{I}} \mathbb{E}\left[(Y - f(X))^2\right] \leq \min_{f \in \mathcal{J}} \mathbb{E}\left[(Y - f(X))^2\right]. \tag{3.3.5}$$

We will later show that the two quantities in (3.3.5) are equal (see Corollary 3.9).

3.3.1 Identification

We show that the set $\mathcal{J}$ is identified from the observed data by a conditional expectation, and therefore, any function $f \in \mathcal{J}$ can be estimated using nonparametric regression techniques. Identification of $\mathcal{J}$ relies on the control function approach developed by Ng and Pinkse (1995); Newey et al. (1999).

Proposition 3.5. *Consider an SCM $\mathcal{C}_0$ as defined in* (3.2.1)*. Let $\mathcal{J}$ denote the class of linear invariant functions related to $\mathcal{C}_0$. Then, for any $f \in \mathcal{F}$, if holds that $f \in \mathcal{J}$ if and only if there exists $\gamma \in \mathbb{R}^p$ such that for almost every (a.e.) $x, v \in \mathbb{R}^p$ it holds that*

$$\mathbb{E}[Y \mid X = x, V = v] = f(x) + \gamma^T v, \tag{3.3.6}$$

where the expectation is with respect to the observational distribution $\mathbb{P}_{\mathcal{C}_0}$.

A proof can be found in Appendix C.1.2. Proposition 3.5 states that any function $f \in \mathcal{J}$ can be identified by performing an additive regression of the form $f(x)+\gamma^T v$. This allows us to adapt existing nonparametric techniques to compute $f \in \mathcal{J}$; see Section 3.3.3.

If $\ker(M_0^T)$ contains non-zero vectors, the set $\mathcal{J}$ contains infinitely many functions. To obtain the most predictive function in $\mathcal{J}$, we perform two steps. First, we identify an element $\tilde{f} \in \mathcal{J}$ by fitting a nonparametric regression model as defined in (3.3.6). Given $\tilde{f} \in \mathcal{J}$, we then optimize over the invariant space $\ker(M_0^T)$ to find an optimal $\tilde{\delta} \in \ker(M_0^T)$, i.e.,

$$\min_{f \in \mathcal{J}} \mathbb{E}\left[(Y - f(X))^2\right] = \min_{\delta \in \ker(M_0^T)} \mathbb{E}\left[(Y - \tilde{f}(X) - \delta^T X)^2\right],$$

so that the optimal function writes $f^*(x) = \tilde{f}(x) + \tilde{\delta}^T x$, for all $x \in \mathbb{R}^p$. The next proposition shows that the optimal function f^* has an explicit expression in terms of the data generating model $\mathcal{C}_0 = (f_0, M_0, \gamma_0, Q_0)$. Moreover, it shows that the optimal function f^* is well-defined, i.e., it does not depend on the choice of representative function $\tilde{f} \in \mathcal{J}$.

Proposition 3.6. *Consider an SCM $\mathcal{C}_0$ as defined in* (3.2.1)*. Let $V \sim N(0, S)$ with $S \succ 0$, and let $R \in \mathbb{R}^{p \times (p-r)}$ denote an orthonormal basis for $\ker(M_0^T)$. Define*

$$\delta_0 := R(R^T S R)^{-1} R^T S \gamma_0 \in \mathbb{R}^p. \tag{3.3.7}$$

Then, it holds

$$\delta_0 = \operatorname*{arg\,min}_{\delta \in \ker(M_0^T)} \mathbb{E}\left[(Y - f_0(X) - \delta^T X)^2\right],$$

and so, the optimal function writes $f^*(x) = f_0(x) + \delta_0^T x$, for all $x \in \mathbb{R}^p$. Moreover, the optimal function f^* is well-defined.

A proof can be found in Appendix C.1.3.

3.3.2 Distribution generalization

Given the set $\mathcal{J}$, we can compute

$$f^* := \operatorname*{arg\,min}_{f \in \mathcal{J}} \mathbb{E}[(Y - f(X))^2], \tag{3.3.8}$$

where the expectation is taken with respect to $\mathbb{P}_{\mathcal{C}_0}$. The function f^* is invariant, and is the most predictive among all members of $\mathcal{J}$. In fact, it turns out that f^* is minimax optimal under all hard interventions on the environment vector $E \in \mathbb{R}^r$, that is

$$\min_{f \in \mathcal{J}} \mathbb{E}[(Y - f(X))^2] = \min_{f \in \mathcal{F}} \sup_{e \in \mathbb{R}^r} \mathbb{E}[(Y - f(X))^2 \mid \mathrm{do}(E := e)].$$

Before stating the minimax result, we provide the following lemma.

Lemma 3.7. *Consider an SCM $\mathcal{C}_0$ as defined in (3.2.1). Let $V \sim N(0, S)$ with $S \succ 0$. Define the function $F : \mathcal{F} \times \mathbb{R}^r \to [0, \infty)$ for all $h \in \mathcal{F}$ and $e \in \mathbb{R}^r$ by*

$$F(h, e) := \mathbb{E}[(V^T\gamma_0 - h(M_0 e + V))^2].$$

Let $\delta_0 \in \ker(M_0^T)$ be the minimizer defined in (3.3.7). Then, it holds that

$$\mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] = \inf_{h\in\mathcal{F}} \sup_{e\in\mathbb{R}} F(h, e).$$

A proof can be found in Appendix C.1.4. Based on Lemma 3.7, we show that the minimizer of the MSPE over $\mathcal{J}$ achieves the minimax loss under all hard interventions on the environment vector.

Theorem 3.8. *Consider an SCM $\mathcal{C}_0$ as defined in (3.2.1). Let $\mathcal{J}$ denote the class of invariant linear functions related to $\mathcal{C}_0$. Then, under any hard intervention $\mathrm{do}(E := e)$, $e \in \mathbb{R}^r$, it holds that*

$$\min_{f\in\mathcal{J}} \mathbb{E}[(Y - f(X))^2] = \min_{f\in\mathcal{F}} \sup_{e\in\mathbb{R}^r} \mathbb{E}[(Y - f(X))^2 \mid \mathrm{do}(E := e)]. \tag{3.3.9}$$

A proof can be found in Appendix C.1.5. Building on this minimax result, we can now show that minimization of the MSPE over the set of invariant functions $\mathcal{I}$ is attained in $\mathcal{J}$.

Corollary 3.9. *Consider an SCM $\mathcal{C}_0$ as defined in (3.2.1). Let $\mathcal{I}$ and $\mathcal{J}$ denote the classes of invariant and linear invariant functions related to $\mathcal{C}_0$, respectively. Then, it holds that*

$$\min_{f\in\mathcal{J}} \mathbb{E}\left[(Y - f(X))^2\right] = \min_{f\in\mathcal{I}} \mathbb{E}\left[(Y - f(X))^2\right]. \tag{3.3.10}$$

A proof can be found in Appendix C.1.6.

3.3.3 Learn the IMP function

In this section, we propose an adaptation of the regression tree algorithm (Breiman et al., 1984) to learn the IMP function in large dimensions. Let $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ be a predictor vector and response variable generated from an SCM $\mathcal{C}_0$ as defined in (3.2.1). A regression tree with K leaves, is a function $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ defined for all $x \in \mathbb{R}^p$ such that

$$\hat{f}(x) = \sum_{k=1}^{K} \theta_k \mathbb{1}\{x \in t_k\},$$

where $t_1, \ldots, t_K \subseteq \mathbb{R}^p$ are rectangular regions, and θ_k is a constant value in the region t_k. We propose to approximate the conditional expectation defined in (3.3.6) with

$$g(x, v) = \hat{f}(x) + \gamma^T v = \sum_{k=1}^{K} \theta_k \mathbb{1}\{x \in t_k\} + \gamma^T v.$$

To grow such a tree, we adapt the greedy algorithm of Breiman et al. (1984) as follows. Let $P \subseteq \mathbb{R}^p$ denote a parent node. Then, for any partition $C_1, C_2 \subseteq \mathbb{R}^p$ of P, we define the criterion

$$\begin{aligned} \mathrm{err}(C_1, C_2, \theta) &= \mathbb{E}[(Y - \theta_{C_1}\mathbb{1}\{X \in C_1\} - \theta_{C_2}\mathbb{1}\{X \in C_2\} - \gamma^T V)^2 \mid X \in P] \\ &= \mathbb{E}[(Y - \theta^T \tilde{X})^2 \mid X \in P], \end{aligned} \tag{3.3.11}$$

where $\theta = (\theta_{C_1}, \theta_{C_2}, \gamma) \in \mathbb{R}^{p+2}$ and $\tilde{X} = (\mathbb{1}\{X \in C_1\}, \mathbb{1}\{X \in C_2\}, V) \in \mathbb{R}^{p+2}$. Then, for each partition C_1, C_2, the optimal parameter writes

$$\theta^* = \mathbb{E}[\tilde{X}\tilde{X}^T \mid X \in P]^{-1}\mathbb{E}[\tilde{X}Y \mid X \in P]. \tag{3.3.12}$$

To draw a parallel, in classical regression trees we consider the criterion

$$\mathrm{err}(C_1, C_2, \theta) = \mathbb{E}[(Y - \theta_{C_1}\mathbb{1}\{X \in C_1\} - \theta_{C_2}\mathbb{1}\{X \in C_2\})^2 \mid X \in P]$$

in place of (3.3.11), and the resulting optimal constant values over the partition C_j, $j = 1, 2$, writes,

$$\theta^*_{C_j} = \mathbb{E}[\mathbb{1}\left\{X \in C_j\right\} \mid X \in P]^{-1}\mathbb{E}[\mathbb{1}\left\{X \in C_j\right\}Y \mid X \in P] = \mathbb{E}[Y \mid X \in C_j].$$

Building upon (3.3.11) and (3.3.12), we can therefore approximate the conditional expectation in (3.3.6) nonparametrically, and learn the IMP function even when the dimension of the predictor space is large.

Acknowledgments

We thank Jeffrey Glenn Adams, Cesare Miglioli, and Sorawit Saengkyongam for helpful discussions. SE and NG were supported by the Swiss National Science Foundation.

Appendices

Appendix A

Causal discovery in heavy-tailed data

A.1 Some facts about regular variation

In the sequel, for any two functions f, $g : \mathbb{R} \to \mathbb{R}$, we write $f \sim g$ if $f(x)/g(x) \to 1$ as $x \to \infty$. Also, we write $S_p := Y_1 + \cdots + Y_p$, and $M_p := \max(Y_1, \ldots, Y_p)$.

Consider independent random variables $Y_1, \ldots, Y_p$ and assume that they have comparable upper tails, i.e., there exist $c_j, \alpha > 0$ and $\ell \in \mathrm{RV}_0$ such that for all $j \in \{1, \ldots, p\}$

$$\mathbb{P}(Y_j > x) \sim c_j \ell(x) x^{-\alpha}, \quad x \to \infty. \tag{A.1.1}$$

Lemma A.1. *Let $Y_1, \ldots, Y_p$ be real-valued independent regularly varying random variables with comparable tails. Then,*

$$\mathbb{P}(S_p > x) \sim \sum_{h=1}^{p} \mathbb{P}(Y_h > x), \quad x \to \infty.$$

The proof for $p = 2$ of Lemma A.1 can be found in Feller (1971, p. 278) and can be extended to a general p using induction.

An important property of regularly varying random variable is the max-sum-equivalence presented in the following lemma (Embrechts et al., 1997, Sec. 1.3.1).

Lemma A.2. *Let $Y_1, \ldots, Y_p$ be real-valued, independent regularly varying random variables with comparable tails. Then, as $x \to \infty$,*

$$\mathbb{P}(M_p > x) \sim \mathbb{P}(S_p > x).$$

Proof. We can write, as $x \to \infty$,

$$\begin{aligned} \mathbb{P}\{M_p > x\} &= 1 - \mathbb{P}\left\{M_p \leq x\right\} = 1 - P\left(Y_1 \leq x, \ldots, Y_p \leq x\right) \\ &= \sum_{h=1}^{p} \mathbb{P}(Y_h > x) - \sum_{1 \leq h < h' \leq p} \mathbb{P}(Y_h > x, Y_{h'} > x) + \Delta(x), \end{aligned}$$

where $\Delta(x)$ contains terms of higher order interactions of the sets $\{Y_j > x\}$, $j = 1, \ldots, p$. Because of independence, the probability

$$\mathbb{P}(Y_j > x, Y_{j'} > x) = o\{\mathbb{P}(S_p > x)\}.$$

Similarly, this holds for the terms in $\Delta(x)$. Recalling, by Lemma A.1, that $\mathbb{P}(Y_1 > x) + \cdots + \mathbb{P}(Y_p > x) \sim \mathbb{P}(S_p > x)$, the result follows. □

Lemma A.3. *Let $Y_1, \dots, Y_p$ be real-valued independent regularly varying random variables with comparable tails. Then, for $j = 1, \dots, p$,*

$$\mathbb{P}(Y_j > x, S_p > x) \sim \mathbb{P}(Y_j > x), \quad x \to \infty.$$

Proof. For any $x > 0$ and $\delta \in (0, 1/(2p-2))$, we have, for $j = 1, \dots, p$,

$$\left\{ Y_j > x + (p-1)\delta x, \bigcap_{h \neq j} \{Y_h > -\delta x\} \right\} \subset \{Y_j > x, S_p > x\} \subset \{Y_j > x\}.$$

Considering the upper bound, it holds, $j = 1, \dots, p$,

$$\mathbb{P}(Y_j > x, S_p > x) \leq \mathbb{P}(Y_j > x).$$

Regarding the lower bound, we get, as $x \to \infty$, $j = 1, \dots, p$,

$$\begin{aligned} \mathbb{P}(Y_j > x, S_p > x) &\geq \mathbb{P}\left(Y_j > x + (p-1)\delta x, \bigcap_{h \neq j} \{Y_h > -\delta x\} \right) \\ &= \mathbb{P}\left(Y_j > x + (p-1)\delta x \right) \prod_{h \neq j} \mathbb{P}\left(Y_h > -\delta x \right). \end{aligned}$$

Dividing everything by $\mathbb{P}(Y_j > x)$, and letting first $x \to \infty$ and then $\delta \downarrow 0$, we get the desired result. □

Lemma A.4. *Let $Y_1, \dots, Y_p$ be real-valued independent regularly varying random variables with comparable tails. Then, as $x \to \infty$,*

$$\mathbb{P}\{S_p > x, M_p \leq x\} = o\{\mathbb{P}(S_p > x)\}.$$

Proof. We first write $\mathbb{P}\{S_p > x, M_p \leq x\} = \mathbb{P}(S_p > x) - \mathbb{P}\{S_p > x, M_p > x\}$. Let $I := \{1, \dots, p\}$. By definition of the maximum function and using the inclusion-exclusion principle, it follows that

$$\begin{aligned} \mathbb{P}\{S_p > x, M_p > x\} &= \mathbb{P}\left(S_p > x, \bigcup_{h \in I} \{Y_h > x\} \right) \\ &= \sum_{h \in I} \mathbb{P}\left(S_p > x, Y_h > x \right) \\ &\quad - \sum_{1 \leq h < h' \leq p} \mathbb{P}\left(S_p > x, Y_h > x, Y_{h'} > x \right) + \Delta(x). \end{aligned}$$

Regarding the summands in first term, it holds by Lemma A.3, $h \in I$,

$$\mathbb{P}\left(S_p > x, Y_h > x \right) \sim \mathbb{P}(Y_h > x), \quad x \to \infty.$$

The summands in the second term can be upper bounded by, $1 \leq h < h' \leq p$,

$$\mathbb{P}(Y_h > x, Y_{h'} > x) = o\{\mathbb{P}(S_p > x)\}, \quad x \to \infty.$$

The same holds for $\Delta(x)$ which contains terms of higher order interactions of the sets $\{Y_j > x\}$, $j \in I$. Putting everything together, we obtain

$$\mathbb{P}\{S_p > x, M_p \leq x\} \sim \mathbb{P}(S_p > x) - \sum_{h \in I} \mathbb{P}(Y_h > x) + o\{\mathbb{P}(S_p > x)\} = o\{\mathbb{P}(S_p > x)\},$$

where in the last equality we used Lemma A.1. □

A.2 Proofs

A.2.1 Proof of Lemma 1.2

Proof. Let $j, k \in V$ and $j \neq k$. Recall that each variable X_h, $h \in V$, can be expressed as a weighted sum of the noise terms $\varepsilon_1, \ldots, \varepsilon_p$ belonging to the ancestors of X_h, as shown in (1.2.3). Therefore, we can write X_j and X_k as

$$X_j = \sum_{h \in A_{jk}} \beta_{h\to j}\varepsilon_h + \sum_{h \in A_{jk^*}} \beta_{h\to j}\varepsilon_h,$$
$$X_k = \sum_{h \in A_{jk}} \beta_{h\to k}\varepsilon_h + \sum_{h \in A_{kj^*}} \beta_{h\to k}\varepsilon_h,$$

where $A_{jk} = \mathrm{An}(j, G) \cap \mathrm{An}(k, G)$, $A_{jk^*} = \mathrm{An}(j, G) \cap \mathrm{An}(k, G)^c$ and similarly for A_{kj^*}. We have

$$\mathbb{E}\left[F_k(X_k)\mathbf{1}\{X_j > x\}\right] = \mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{X_j > x, \bigcup_{h\in \mathrm{An}(j,G)} \left\{\beta_{h\to j}\varepsilon_h > x\right\}\right\}\right]$$
$$+ \mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{X_j > x, \max_{h\in \mathrm{An}(j,G)} \left\{\beta_{h\to j}\varepsilon_h\right\} \leq x\right\}\right].$$

The second summand can be bounded by

$$\mathbb{P}\left[X_j > x, \max_{h\in \mathrm{An}(j,G)} \left\{\beta_{h\to j}\varepsilon_h\right\} \leq x\right] = o\{\mathbb{P}(X_j > x)\},$$

by Lemma A.4. For the first term, we use the inclusion-exclusion principle to write

$$\mathbf{1}\left\{X_j > x, \bigcup_{h\in \mathrm{An}(j,G)} \left\{\beta_{h\to j}\varepsilon_h > x\right\}\right\} = \sum_{h\in \mathrm{An}(j,G)} \mathbf{1}\left\{X_j > x, \beta_{h\to j}\varepsilon_h > x\right\}$$
$$- \sum_{h,h'\in \mathrm{An}(j,G), h<h'} \mathbf{1}\left\{X_j > x, \beta_{h\to j}\varepsilon_h > x, \beta_{h'\to j}\varepsilon_{h'} > x\right\} + \Delta(x),$$

where $\Delta(x)$ contains terms of higher order interactions of the sets $\{\beta_{h\to j}\varepsilon_h > x\}$, $h \in \mathrm{An}(j, G)$. The probability

$$\mathbb{P}\left(X_j > x, \beta_{h\to j}\varepsilon_h > x, \beta_{h'\to j}\varepsilon_{h'} > x\right)$$
$$\leq \mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x, \beta_{h'\to j}\varepsilon_{h'} > x\right) = o\{\mathbb{P}(X_j > x)\}.$$

The same holds for all finitely many terms in $\Delta(x)$. We further note that for all $h \in \mathrm{An}(j, G)$, by Lemma A.3,

$$\mathbb{P}(X_j > x, \beta_{h\to j}\varepsilon_h > x) = \mathbb{P}(\beta_{h\to j}\varepsilon_h > x) + o\{\mathbb{P}(X_j > x)\}.$$

Putting everything together, we can rewrite

$$\mathbb{E}\left[F_k(X_k)\mathbf{1}\{X_j > x\}\right] = \sum_{h\in \mathrm{An}(j,G)} \mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] + o\{\mathbb{P}(X_j > x)\}$$
$$= \sum_{h\in A_{jk}} \mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right]$$
$$+ \sum_{h\in A_{jk^*}} \mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] + o\{\mathbb{P}(X_j > x)\}.$$

For $h \in A_{jk}$, let $c = \beta_{h\to j}/\beta_{h\to k} > 0$, and note that for every $x > 0$,

$$\begin{aligned}\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right) &\geq \mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] \\ &\geq \mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{\beta_{h\to k}\varepsilon_h > cx, X_k > cx\right\}\right] \\ &\geq F_k(cx)\mathbb{P}\left(\beta_{h\to k}\varepsilon_h > cx, X_k > cx\right).\end{aligned}$$

Therefore, using Lemma A.3 and that $F_k(cx) \to 1$ as $x \to \infty$, it follows that

$$\mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] \sim \mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right).$$

On the other hand, for $h \in A_{jk^*}$, we have that X_k and ε_h are independent, and therefore

$$\mathbb{E}\left[F_k(X_k)\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] = \frac{1}{2}\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right), \quad x > 0.$$

Consequently,

$$\begin{aligned}\Gamma_{jk} &= \lim_{x\to\infty} \mathbb{E}\left[F_k(X_k) \mid X_j > x\right] \\ &= \lim_{x\to\infty} \sum_{h\in A_{jk}} \frac{\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right)}{\mathbb{P}(X_j > x)} + \lim_{x\to\infty} \frac{1}{2} \sum_{h\in A_{jk^*}} \frac{\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right)}{\mathbb{P}(X_j > x)} \\ &= \frac{1}{2} + \frac{1}{2} \sum_{h\in A_{jk}} \lim_{x\to\infty} \frac{\mathbb{P}(\beta_{h\to j}\varepsilon_h > x)}{\mathbb{P}(X_j > x)} = \frac{1}{2} + \frac{1}{2} \lim_{x\to\infty} \frac{\sum_{h\in A_{jk}} \beta_{h\to j}^{\alpha}\mathbb{P}(\varepsilon_h > x)}{\sum_{h\in \mathrm{An}(j,G)} \beta_{h\to j}^{\alpha}\mathbb{P}(\varepsilon_h > x)} \\ &= \frac{1}{2} + \frac{1}{2}\frac{\sum_{h\in A_{jk}} \beta_{h\to j}^{\alpha}}{\sum_{h\in \mathrm{An}(j,G)} \beta_{h\to j}^{\alpha}},\end{aligned}$$

where the second last equality follows from the fact that $\beta_{h\to j}\varepsilon_h$, $h \in A_{jk}$, are independent regularly varying random variables, see Lemma A.1, and the last equality holds because we assume that the noise variables ε_j, $j \in V$, have comparable tails; see Section 1.2.1. □

A.2.2 Proof of Theorem 1.3

Proof. Recall that $\mathrm{an}(j, G) = \mathrm{An}(j, G) \setminus \{j\}$ and define $A_{12} = \mathrm{An}(1, G) \cap \mathrm{An}(2, G)$.

(a). Suppose X_1 causes X_2, i.e., $1 \in \mathrm{an}(2, G)$. This implies that $\mathrm{An}(1, G) \subset \mathrm{An}(2, G)$ and thus $A_{12} = \mathrm{An}(1, G) \subset \mathrm{An}(2, G)$. By applying Lemma 1.2 we obtain $\Gamma_{12} = 1$ and $\Gamma_{21} \in (1/2, 1)$.

Conversely, suppose that $\Gamma_{12} = 1$ and $\Gamma_{21} \in (1/2, 1)$. If $\Gamma_{12} = 1$ then the numerator and denominator of the second term in Lemma 1.2 must be equal and strictly positive. This implies that $A_{12} = \mathrm{An}(1, G) \cap \mathrm{An}(2, G) = \mathrm{An}(1, G) \neq \varnothing$. It follows that $\mathrm{An}(1, G) \subseteq \mathrm{An}(2, G)$, i.e., $1 \in \mathrm{An}(2, G)$. At the same time, if $\Gamma_{21} \in (1/2, 1)$, then the numerator of the second term in Lemma 1.2 must be positive and smaller than the denominator. This means that $A_{12} \neq \varnothing$ and $A_{12} = \mathrm{An}(1, G) \cap \mathrm{An}(2, G) \subset \mathrm{An}(2, G)$. Thus, it follows that $\mathrm{An}(1, G) \subset \mathrm{An}(2, G)$. Therefore, $1 \in \mathrm{an}(2, G)$, that is, X_1 causes X_2.

(b). By symmetry, as case (a).

(c). Suppose there is no causal link between X_1 and X_2, i.e., $\mathrm{An}(1,G) \cap \mathrm{An}(2,G) = \varnothing$. Then, $A_{12} = \mathrm{An}(1,G) \cap \mathrm{An}(2,G) = \varnothing$ and by Lemma 1.2, we obtain $\Gamma_{12} = \Gamma_{21} = 1/2$.

Suppose now that $\Gamma_{12} = \Gamma_{21} = 1/2$. This means that the numerator of the second term in Lemma 1.2 must be equal to zero. This implies that $A_{12} = \varnothing$ and therefore $\mathrm{An}(1,G) \cap \mathrm{An}(2,G) = \varnothing$, that is, there is no causal link between X_1 and X_2.

(d). Suppose there is a node $j \notin \{1,2\}$ such that X_j is a common cause of X_1 and X_2, i.e., $j \in \mathrm{an}(1,G)$ and $j \in \mathrm{an}(2,G)$. Then $A_{12} = \mathrm{An}(1,G) \cap \mathrm{An}(2,G)$ is non-empty. Since $\mathrm{An}(1,G) \neq \mathrm{An}(2,G)$, it follows that $A_{12} \subset \mathrm{An}(i,G)$, for $i = 1,2$. Thus, according to Lemma 1.2 we have $\Gamma_{12}, \Gamma_{21} \in (1/2, 1)$.

Conversely, suppose that $\Gamma_{12}, \Gamma_{21} \in (1/2,1)$. If $\Gamma_{12} \in (1/2,1)$, then the numerator of the second term in Lemma 1.2 must be positive and smaller than the denominator. This implies that $A_{12} \neq \varnothing$ and $A_{12} = \mathrm{An}(1,G) \cap \mathrm{An}(2,G) \subset \mathrm{An}(1,G)$. Similarly, if $\Gamma_{21} \in (1/2,1)$, it follows that $A_{21} = A_{12} = \mathrm{An}(1,G) \cap \mathrm{An}(2,G) \subset \mathrm{An}(2,G)$. This implies that $\mathrm{An}(1,G) \neq \mathrm{An}(2,G)$ and they are not disjoint. Therefore, there exists a node $j \notin \{1,2\}$ such that $j \in \mathrm{an}(1,G)$ and $j \in \mathrm{an}(2,G)$, i.e., X_j is a common cause of X_1 and X_2.

□

A.2.3 Proof of Theorem 1.4

Proof. For simplicity we will write $k = k_n$ in the sequel. We only show the result for $\widehat{\Gamma}_{21}$, the proof for $\widehat{\Gamma}_{12}$ follows by symmetry. Recall that each variable X_h, $h \in V$, can be expressed as a weighted sum of the noise terms $\varepsilon_1, \ldots, \varepsilon_p$ belonging to the ancestors of X_h, as shown in (1.2.3). Therefore, we can write X_1 and X_2 as follows,

$$\begin{aligned} X_1 &= \sum_{h \in A} \beta_{h\to 1}\varepsilon_h + \sum_{h \in A_1} \beta_{h\to 1}\varepsilon_h, \\ X_2 &= \sum_{h \in A} \beta_{h\to 2}\varepsilon_h + \sum_{h \in A_2} \beta_{h\to 2}\varepsilon_h, \end{aligned}$$

where $A = A_{12} = \mathrm{An}(1,G) \cap \mathrm{An}(2,G)$ and $A_j = A_{jk^*} = \mathrm{An}(j,G) \setminus A$, for $j,k = 1,2$. Thus, the estimator $\widehat{\Gamma}_{21}$ can be rewritten as

$$\begin{aligned} \widehat{\Gamma}_{21} &= \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\ \mathbf{1}\left\{ X_{i2} > X_{(n-k),2},\ \bigcup_{h \in \mathrm{An}(2,G)} \left\{ \beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2} \right\} \right\} \\ &\quad + \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\ \mathbf{1}\left\{ X_{i2} > X_{(n-k),2},\ \max_{h \in \mathrm{An}(2,G)} \beta_{h\to 2}\varepsilon_{ih} \leq X_{(n-k),2} \right\} \\ &= S_{1,n} + S_{2,n}. \end{aligned} \tag{A.2.1}$$

Define the theoretical quantile function as

$$U(x) = F^{\leftarrow}(1 - 1/x), \quad x > 1. \tag{A.2.2}$$

Recall that $X_{(n-k),2} = \widehat{F}_2^{\leftarrow}(1 - k/n)$ is the $(n-k)$-th order statistic of $X_{12}, \ldots, X_{n2}$, and as such an approximation to the theoretical quantile $U_2(n/k)$. Under the von Mises'

condition (1.2.7) this convergence of the intermediate order statistics can be made rigorous (de Haan and Ferreira, 2006, Theorem 2.2.1), namely

$$\sqrt{k}\left(\frac{X_{(n-k),2}}{U_2\left(\frac{n}{k}\right)}-1\right) \xrightarrow{d} N(0,1/\alpha^2), \quad n\to\infty, \tag{A.2.3}$$

where $\alpha>0$ is the tail index of the variables in the SCM. This implies in particular that $X_{(n-k),2}\to\infty$, $n\to\infty$. For any $\delta_1>0$ define the event

$$B_{n\delta_1}=\left\{\left|X_{(n-k),2}/U_2\left(\frac{n}{k}\right)-1\right|<\delta_1\right\}, \tag{A.2.4}$$

and note that by (A.2.3) it holds that $\mathbb{P}(B_{n\delta_1})\to 1$ as $n\to\infty$.

Since the noise terms ε_h are independent regularly varying random variables with comparable tails, then, by Lemma A.1, we have

$$\mathbb{P}(X_2>x)\sim\left(\sum_{h\in\mathrm{An}(2,G)}\beta_{h\to 2}^{\alpha}\right)\ell(x)x^{-\alpha}=:c_2\ell(x)x^{-\alpha}.$$

Furthermore, from Resnick (1987, Prop. 0.8) it holds, for all $h\in\mathrm{An}(2,G)$ and $x>0$,

$$\mathbb{P}\{\varepsilon_h>xU_2(t)\}\sim x^{-\alpha}(c_2t)^{-1}, \quad t\to\infty, \tag{A.2.5}$$

where U_2 is defined as in (A.2.2) for the distribution function F_2.

We treat the two terms in (A.2.1) separately. We can upper bound the absolute value of the second term, for any $\tau,\delta_1>0$, by

$$\begin{aligned}\mathbb{P}(|S_{2,n}|>\tau)\le\mathbb{P}(B_{n\delta_1}^c)+\mathbb{P}\Bigg(&\frac{1}{k}\sum_{i=1}^{n}\mathbf{1}\Big\{X_{i2}>U_2\left(\frac{n}{k}\right)(1-\delta_1),\\ &\max_{h\in\mathrm{An}(2,G)}\beta_{h\to 2}\varepsilon_{ih}\le U_2\left(\frac{n}{k}\right)(1+\delta_1)\Big\}>\tau\Bigg)\end{aligned}$$

where $\mathbb{P}(B_{n\delta_1}^c)\to 0$ as $n\to\infty$ by (A.2.3). The limit superior of the second term, as $n\to\infty$, can be bounded with Markov's inequality by

$$\begin{aligned}&\lim_{n\to\infty}\frac{n}{\tau k}P\left\{X_2>(1-\delta_1)U_2\left(\frac{n}{k}\right),\max_{h\in\mathrm{An}(2,G)}\beta_{h\to 2}\varepsilon_h\le(1+\delta_1)U_2\left(\frac{n}{k}\right)\right\}\\ &\le\lim_{n\to\infty}\frac{1}{\tau}\left(\mathbb{P}\left\{X_2>U_2\left(\frac{n}{k}\right)\right\}\right)^{-1}\left[\mathbb{P}\left\{X_2>(1-\delta_1)U_2\left(\frac{n}{k}\right)\right\}\right.\\ &\quad\left.-\mathbb{P}\left\{X_2>(1+\delta_1)U_2\left(\frac{n}{k}\right),\max_{h\in\mathrm{An}(2,G)}\beta_{h\to 2}\varepsilon_h>(1+\delta_1)U_2\left(\frac{n}{k}\right)\right\}\right]\\ &=\frac{(1-\delta_1)^{-\alpha}-(1+\delta_1)^{-\alpha}}{\tau}.\end{aligned}$$

The last equality holds because X_2 is regularly varying with index α and because, by Lemma A.4, we have

$$\mathbb{P}\left(X_2>x,\max_{h\in\mathrm{An}(2,G)}\beta_{h\to 2}\varepsilon_h>x\right)\sim\mathbb{P}(X_2>x), \quad x\to\infty.$$

Since $\delta_1, \tau > 0$ are arbitrary, it follows that $S_{2,n} = o_P(1)$.

For the first term, we use the inclusion-exclusion principle to write

$$\begin{aligned} S_{1,n} = & \sum_{h \in \mathrm{An}(2,G)} \frac{1}{k} \sum_{i=1}^{n} \widehat{F}_1(X_{i1}) \mathbf{1}\left\{X_{i2} > X_{(n-k),2}, \beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}\right\} \\ & - \sum_{h,h' \in \mathrm{An}(2,G), h<h'} \frac{1}{k} \sum_{i=1}^{n} \widehat{F}_1(X_{i1}) \mathbf{1}\Big\{X_{i2} > X_{(n-k),2}, \beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}, \\ & \beta_{h'\to 2}\varepsilon_{ih'} > X_{(n-k),2}\Big\} + \frac{1}{k} \sum_{i=1}^{n} \widehat{F}_1(X_{i1}) \Delta_i\left(X_{(n-k),2}\right) \\ = & \, T_{1,n} + T_{2,n} + T_{3,n}, \end{aligned}$$

where $\Delta_i(X_{(n-k),2})$ contains terms of higher order interactions of the sets $\{\beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}\}$, $h \in \mathrm{An}(2,G)$, $i = 1, \ldots, n$. First, we show that the terms $T_{2,n}$ and $T_{3,n}$ are $o_P(1)$. Considering $T_{2,n}$, for each $h, h' \in \mathrm{An}(2,G)$, $h < h'$, define

$$\begin{aligned} T_{2,n}^{(h,h')} = & \frac{1}{k} \sum_{i=1}^{n} \widehat{F}_1(X_{i1}) \mathbf{1}\Big\{X_{i2} > X_{(n-k),2}, \beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}, \\ & \beta_{h'\to 2}\varepsilon_{ih'} > X_{(n-k),2}\Big\}. \end{aligned}$$

We can upper bound its absolute value, for any $\tau, \delta_1 > 0$, by

$$\begin{aligned} & \mathbb{P}(|T_{2,n}^{(h,h')}| > \tau) \le \mathbb{P}(B_{n\delta_1}^c) \\ & \quad + \mathbb{P}\left(\frac{1}{k} \sum_{i=1}^{n} \mathbf{1}\Big\{\beta_{h\to 2}\varepsilon_{ih} > (1-\delta_1) U_2\left(\frac{n}{k}\right), \beta_{h'\to 2}\varepsilon_{ih'} > (1-\delta_1) U_2\left(\frac{n}{k}\right)\Big\}\right). \end{aligned}$$

By Markov's inequality,

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{P}(|T_{2,n}^{(h,h')}| > \tau) \le & \lim_{n\to\infty} \mathbb{P}(B_{n\delta_1}^c) \\ & + \lim_{n\to\infty} \frac{n}{\tau k} \mathbb{P}\left\{\beta_{h\to 2}\varepsilon_h > (1-\delta_1) U_2\left(\frac{n}{k}\right)\right\} \mathbb{P}\left\{\beta_{h'\to 2}\varepsilon_{h'} > (1-\delta_1) U_2\left(\frac{n}{k}\right)\right\} \\ = & \lim_{n\to\infty} \frac{k(1-\delta_1)^{-2\alpha} p_{2h} p_{2h'}}{n\tau} = 0, \end{aligned}$$

where in the last line we used property (A.2.5) and the fact that $k/n \to 0$ as $n \to \infty$, and $p_{h2}, p_{2h'}$ are defined as

$$p_{2h} = \frac{\beta_{h\to 2}^{\alpha}}{c_2}, \quad h \in \mathrm{An}(2,G). \tag{A.2.6}$$

Since δ_1, τ are arbitrary, putting together the finitely many terms $h < h'$ where $h, h' \in \mathrm{An}(2,G)$, it follows that $T_{2,n} \xrightarrow{P} 0$. Using a similar argument as the one for $T_{2,n}$, one can show that $T_{3,n} \xrightarrow{P} 0$.

We want to show that $T_{1,n} \xrightarrow{P} \Gamma_{21}$. Rewrite

$$\begin{aligned} T_{1,n} = & \sum_{h\in A} \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\mathbf{1}\left\{\beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}\right\} \\ & + \sum_{h\in A_2} \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\mathbf{1}\left\{\beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}\right\} \\ & - \sum_{h\in \mathrm{An}(2,G)} \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\mathbf{1}\left\{\beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}, X_{i2} \le X_{(n-k),2}\right\} \\ = & \, U_{1,n} + U_{2,n} + U_{3,n}. \end{aligned}$$

Using an argument similar to the one for $S_{2,n}$ on page 68, one can show that $U_{3,n} \xrightarrow{P} 0$.

Regarding $U_{1,n}$, for each $h \in A$, define

$$U_{1,n}^{(h)} = \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\mathbf{1}\left\{\beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}\right\}.$$

Note that, for $h \in A$, we have that both $\beta_{h\to 2} > 0$ and $\beta_{h\to 1} > 0$, therefore we can bound

$$\widehat{F}_1\left(cX_{(n-k),2}\right) V_n^{(h)} - W_n^{(h)} \le U_{1,n}^{(h)} \le V_n^{(h)}, \tag{A.2.7}$$

where $c = \beta_{h\to 1}/\beta_{h\to 2} > 0$ and

$$\begin{aligned} V_n = V_n^{(h)} &= \frac{n}{k}\left[1 - \widehat{F}_{\varepsilon_h}\left(\frac{X_{(n-k),2}}{\beta_{h\to 2}}\right)\right], \\ W_n = W_n^{(h)} &= \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\left\{\beta_{h\to 1}\varepsilon_{ih} > cX_{(n-k),2}, X_{i1} \le cX_{(n-k),2}\right\}. \end{aligned} \tag{A.2.8}$$

We show first that V_n converges in probability to p_{2h}, defined in (A.2.6). This is motivated by the fact that V_n is the empirical version of

$$t\mathbb{P}\left\{\beta_{h\to 2}\varepsilon_h > U_2(t)\right\} \to \beta_{h\to 2}/c_2 = p_{2h}, \quad t \to \infty,$$

where the limit follows from property (A.2.5). We will study the asymptotic properties of V_n. For $x > 0$ define

$$v_n(x) = \frac{n}{k}\left[1 - \widehat{F}_{\varepsilon_h}\left(\frac{xU_2\left(\frac{n}{k}\right)}{\beta_{h\to 2}}\right)\right] = \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\left\{\beta_{h\to 2}\varepsilon_{ih} > xU_2\left(\frac{n}{k}\right)\right\},$$

which is a nonincreasing function of x. Observe that on the set $B_{n\delta_1}$ defined in (A.2.4), we may bound the random variable V_n by

$$v_n(1+\delta_1) \le V_n = v_n\left(\frac{X_{(n-k),2}}{U_2\left(\frac{n}{k}\right)}\right) \le v_n(1-\delta_1). \tag{A.2.9}$$

We further compute the limits as $n \to \infty$

$$\begin{aligned} \mathbb{E}\{v_n(x)\} &= \frac{n}{k}\mathbb{P}\left[\beta_{h\to 2}\varepsilon_h > xU_2\left(\frac{n}{k}\right)\right] \to p_{2h}x^{-\alpha}, \\ \mathbb{V}\{v_n(x)\} &= \frac{n}{k^2}\mathbb{P}\left[\beta_{h\to 2}\varepsilon_h > xU_2\left(\frac{n}{k}\right)\right]\mathbb{P}\left[\beta_{h\to 2}\varepsilon_h \le xU_2\left(\frac{n}{k}\right)\right] = O(1/k) \to 0. \end{aligned}$$

An application of Chebyshev's inequality yields

$$v_n(1+\delta_1) \xrightarrow{P} p_{2h}(1+\delta_1)^{-\alpha}, \quad v_n(1-\delta_1) \xrightarrow{P} p_{2h}(1-\delta_1)^{-\alpha}, \quad n \to \infty. \tag{A.2.10}$$

For some $\tau > 0$, choose $\delta_1 > 0$ such that $p_{2h}(1+\delta_1)^{-\alpha} > p_{2h} - \tau$ and $p_{2h}(1-\delta_1)^{-\alpha} < p_{2h} + \tau$. Then with (A.2.9), (A.2.10) and the fact that $\mathbb{P}(B^c_{n\delta_1}) \to 0$ for $B_{n\delta_1}$ in (A.2.4), we conclude

$$\begin{aligned} \mathbb{P}(|V_n - p_{2h}| > \tau) \leq \mathbb{P}(B^c_{n\delta_1}) &+ \mathbb{P}\{v_n(1+\delta_1) < p_{2h} - \tau\} \\ &+ \mathbb{P}\{v_n(1-\delta_1) > p_{2h} + \tau\} \to 0, \quad n \to \infty, \end{aligned} \tag{A.2.11}$$

that is, V_n converges in probability to p_{2h}, as $n \to \infty$.

Furthermore, using (A.2.7) we will now show

$$\widehat{F}_1\left(cX_{(n-k),2}\right) \xrightarrow{P} 1. \tag{A.2.12}$$

Indeed, for any $\tau, \delta_1 > 0$, as $n \to \infty$

$$\begin{aligned} \mathbb{P}\left\{\left|\widehat{F}_1\left(cX_{(n-k),2}\right) - 1\right| > \tau\right\} \leq \mathbb{P}\left(B^c_{n\delta_1}\right) &+ \mathbb{P}\left[\widehat{F}_1\left\{c(1+\delta_1)U_2\left(\frac{n}{k}\right)\right\} > 1+\tau\right] \\ &+ \mathbb{P}\left[\widehat{F}_1\left\{c(1-\delta_1)U_2\left(\frac{n}{k}\right)\right\} < 1-\tau\right] \to 0, \end{aligned}$$

since $\widehat{F}_1(x)$ converges in probability to $F_1(x)$ for all $x \in \mathbb{R}$, and $U_2(n/k) \to \infty$, as $n \to \infty$. Moreover, with a similar argument as for $S_{2,n}$, one can show that W_n defined in (A.2.8) converges in probability to 0, as $n \to \infty$.

Putting everything together, using (A.2.7), (A.2.11) and (A.2.12) we conclude that $U^{(h)}_{1,n} \xrightarrow{P} p_{2h}$, $h \in A$, and thus

$$U_{1,n} \xrightarrow{P} \sum_{h\in A} p_{2h} = \frac{\sum_{h\in A} \beta^{\alpha}_{h\to 2}}{\sum_{h\in \mathrm{An}(2,G)} \beta^{\alpha}_{h\to 2}}. \tag{A.2.13}$$

Considering the term $U_{2,n}$, for each $h \in A_2$, define

$$U^{(h)}_{2,n} = \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\mathbf{1}\left\{\beta_{h\to 2}\varepsilon_{ih} > X_{(n-k),2}\right\}.$$

On the event $B_{n\delta_1}$, we can bound $U^{(h)}_{2,n}$ by

$$u_{2,n}(1+\delta_1) \leq U^{(h)}_{2,n} \leq u_{2,n}(1-\delta_1),$$

where we let, for all $x > 0$,

$$u_{2,n}(x) = \frac{1}{k}\sum_{i=1}^{n} \widehat{F}_1(X_{i1})\mathbf{1}\left\{\beta_{h\to 2}\varepsilon_{ih} > xU_2\left(\frac{n}{k}\right)\right\}.$$

Since X_1 is independent of ε_h, for $h \in A_2$, the values in the sum can be seen as $M(x) = M_n(x)$ random samples out of $\{1/n, \ldots, 1\}$ without replacement, where $M(x)$ is Binomial with success probability

$$\mathbb{P}\left\{\beta_{h\to 2}\varepsilon_h > xU_2\left(\frac{n}{k}\right)\right\} \sim \frac{kp_{2h}x^{-\alpha}}{n}, \quad n \to \infty.$$

Let Z_{ni} be random samples out of $\{1/n, \dots, 1\}$ without replacement, for all $i = 1, \dots, n$, $n \in \mathbb{N}$. Then $u_{2,n}(x)$ has the same distribution as

$$\frac{1}{k} \sum_{i=1}^{M(x)} Z_{ni}.$$

By a similar argument as in (A.2.10) the distribution of $M(x)$ satisfies for any fixed $x \in (0, \infty)$

$$\frac{M(x)}{m(x)} \xrightarrow{P} 1, \quad n \to \infty, \tag{A.2.14}$$

where $m(x) = m_n(x) = \lceil k p_{2h} x^{-\alpha} \rceil$. Thus, for any $\delta_2 > 0$ and any $x > 0$, the probability of the event

$$C_{n\delta_2,x} = \left\{ \left| \frac{M(x)}{m(x)} - 1 \right| < \delta_2 \right\}$$

converges to 1 as $n \to \infty$. Consider the quantity

$$\widetilde{u}_{2,n}(x) = \frac{1}{k} \sum_{i=1}^{m(x)} Z_{ni}. \tag{A.2.15}$$

Theorem 5.1 in Rosén (1965) states that the limit in probability of this sum of samples without replacement is the same as the corresponding sum of samples with replacement. Therefore, for any $x \in (0, \infty)$, we have the convergence in probability

$$\widetilde{u}_{2,n}(x) = \frac{m(x)}{k} \frac{1}{m(x)} \sum_{i=1}^{m(x)} Z_{ni} \xrightarrow{P} \frac{1}{2} p_{2h} x^{-\alpha}. \tag{A.2.16}$$

For $\tau > 0$, choose $\delta_1, \delta_2 > 0$ small enough such that

$$p_{2h}(1 - \delta_1)^{-\alpha}(1 + \delta_2)/2 < p_{2h}/2 + \tau. \tag{A.2.17}$$

Then we can bound the probability

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left(U_{2,n}^{(h)} - p_{2h}/2 > \tau \right) \\
&\quad \leq \limsup_{n\to\infty} \mathbb{P}\left\{ u_{2,n}(1 - \delta_1) > p_{2h}/2 + \tau \right\} + \mathbb{P}(B_{n\delta_1}^c) \\
&\quad \leq \lim_{n\to\infty} \mathbb{P}\left[\widetilde{u}_{2,n}\{(1 - \delta_1)(1 + \delta_2)^{-1/\alpha}\} > p_{2h}/2 + \tau \right] + \mathbb{P}(B_{n\delta_1}^c) + \mathbb{P}(C_{n\delta_2,1-\delta_1}^c) \\
&\quad = 0,
\end{aligned} \tag{A.2.18}$$

since $\mathbb{P}(B_{n\delta_1}^c)$ and $\mathbb{P}(C_{n\delta_2,1-\delta_1}^c)$ converge to 0 as $n \to \infty$, and the last term converges to 0 as a consequence of (A.2.16) and (A.2.17). Similarly, we can show that

$$\limsup_{n\to\infty} \mathbb{P}\left(U_{2,n}^{(h)} - p_{2h}/2 < -\tau \right) = 0,$$

and since $\tau > 0$ is arbitrary, $U_{2,n}^{(h)} \xrightarrow{P} p_{2h}/2$, for $h \in A_2$. Therefore,

$$U_{2,n} \xrightarrow{P} \sum_{h \in A_2} \frac{p_{2h}}{2} = \frac{1}{2} \frac{\sum_{h \in A_2} \beta_{h\to 2}^{\alpha}}{\sum_{h \in \mathrm{An}(2,G)} \beta_{h\to 2}^{\alpha}},$$

and $\widehat{\Gamma}_{21} \xrightarrow{P} \Gamma_{21}$.

□

A.2.4 Proof of Proposition 1.5

Proof. Let $s \in S := \{1, \dots, p\}$ and denote by $i_s \in V$ the node chosen by the algorithm *at* step s. Denote by

$$H_s = \begin{cases} \varnothing, & s = 1, \\ \{i_1, \dots, i_{s-1}\}, & s > 1, \end{cases}$$

the set of nodes chosen by the algorithm *before* step s. Let $G_s = (V_s, E_s)$ be the subgraph of G obtained by removing the nodes H_s that are already chosen, i.e., $V_s = V \setminus H_s$ and $E_s = E \cap (V_s \times V_s)$. Furthermore, define the score minimized by the algorithm to choose the node at step s,

$$M_i^{(s)} = \max_{j \in V_s \setminus \{i\}} \Gamma_{ji}, \quad \forall i \in V_s.$$

We want to show that EASE is a procedure that, for all $s \in S$, satisfies the statement

$$\Xi(s) := \left(i_s \in \arg\min_{i \in V_s} M_i^{(s)} \implies \mathrm{an}(i_s, G) \subseteq H_s \right). \tag{A.2.19}$$

We use strong induction. Namely, we prove that if $\Xi(s')$ holds for *all* natural numbers $s' < s$, then $\Xi(s)$ holds, too.

Fix $s \in S$ and suppose that for all $s' \in S$, with $s' < s$, $\Xi(s')$ holds. Assume $\varphi := i_s \in \arg\min_{i \in V_s} M_i^{(s)}$ and $\mathrm{an}(\varphi, G) \not\subseteq H_s$. Then, there exists a node $j \in V_s$ such that $j \in \mathrm{an}(\varphi, G)$, and by Theorem 1.3, $\Gamma_{j\varphi} = 1$. It follows $M_\varphi^{(s)} = 1$.

Also, since G_s is a DAG, there exists a node $\ell \in V_s$ such that $\mathrm{an}(\ell, G_s) = \varnothing$. If $\Xi(s')$ holds for every natural number $s' < s$, then $\mathrm{an}(\ell, G) \subseteq H_s$. Suppose not. Then, there exists a node $j \in V_s$ such that $j \in \mathrm{an}(\ell, G)$. Note that since $j \in \mathrm{an}(\ell, G)$ and $j \notin \mathrm{an}(\ell, G_s)$, there exists a directed path from j to ℓ in G that is absent in G_s. Thus, there exists a node $h \in H_s$ that lies on such path, and it follows that $j \in \mathrm{an}(h, G)$, which is a contradiction. Since $\mathrm{an}(\ell, G) \subseteq H_s$, by Theorem 1.3 it holds $M_\ell^{(s)} < 1 = M_\varphi^{(s)}$, which is a contradiction.

Since $s \in S$ was arbitrary, we have proved that $\Xi(s)$ holds for all $s \in S$. Furthermore, note that $\Xi(1)$ holds as a special case of the argument above. Hence, we conclude that, for all $s', s \in S$,

$$s' < s \implies i_s \notin \mathrm{an}(i_{s'}, G)$$

and therefore $\pi(i_s) = s$ is a causal order of G. □

A.2.5 Proof of Proposition 1.6

Proof. If $\widehat{\pi} \notin \Pi_G$, then there exists a node $i \in V$ that is chosen before one of its ancestors $u \in \mathrm{an}(i, G)$, i.e., $\widehat{\pi}(i) < \widehat{\pi}(u)$. Therefore, there exists a non-empty set $\tilde{V} \subseteq V$, with $i, u \in \tilde{V}$, such that

$$i \in \arg\min_{i' \in \tilde{V}} \max_{u' \in \tilde{V} \setminus \{i'\}} \widehat{\Gamma}_{u'i'}. \tag{A.2.20}$$

Furthermore, since G is a DAG, there exists a node $j \in \tilde{V}$ with no ancestors in $\tilde{V}$. Let $v \in \tilde{V} \setminus \{j\}$, and note that $v \notin \mathrm{An}(j, G)$. Thus, by (A.2.20), it follows that $\widehat{\Gamma}_{vj} - \widehat{\Gamma}_{ui} \geq 0$. Define $\widehat{\Delta}_{ij} := |\widehat{\Gamma}_{ij} - \Gamma_{ij}|$ and note that the event $\widehat{\Gamma}_{vj} - \widehat{\Gamma}_{ui} \geq 0$ can be bounded by

$$\left\{\widehat{\Delta}_{vj} \geq \frac{\Gamma_{ui} - \Gamma_{vj}}{2}\right\} \cup \left\{\widehat{\Delta}_{ui} \geq \frac{\Gamma_{ui} - \Gamma_{vj}}{2}\right\} \subseteq \left\{\widehat{\Delta}_{vj} \geq \frac{1-\eta}{2}\right\} \cup \left\{\widehat{\Delta}_{ui} \geq \frac{1-\eta}{2}\right\},$$

where $\Gamma_{ui} = 1$ because $u \in \mathrm{an}(i, G)$, and $\Gamma_{vj} \leq \eta < 1$. Therefore,

$$\begin{aligned}\mathbb{P}(\widehat{\pi} \notin \Pi_G) &\leq \sum_{j \in V,\ v \notin \mathrm{An}(j,G)} \mathbb{P}\left(\widehat{\Delta}_{vj} > \frac{1-\eta}{2}\right) + \sum_{i \in V,\ u \in \mathrm{an}(i,G)} \mathbb{P}\left(\widehat{\Delta}_{ui} > \frac{1-\eta}{2}\right) \\ &\leq p^2 \max_{i,j \in V : i \neq j} \mathbb{P}\left(\widehat{\Delta}_{ij} > \frac{1-\eta}{2}\right).\end{aligned}$$

This completes the proof of Proposition 1.6. □

A.2.6 Proof of Lemma 1.8

Proof. Recall that each variable X_h, $h \in V$, can be expressed as a weighted sum of the noise terms $\varepsilon_1, \ldots, \varepsilon_p$ belonging to the ancestors of X_h, as shown in (1.2.3). Therefore, we can write X_j and X_k as follows,

$$\begin{aligned}X_j &= \sum_{h \in A_{jk}} \beta_{h \to j} \varepsilon_h + \sum_{h \in A_{jk^*}} \beta_{h \to j} \varepsilon_h, \\ X_k &= \sum_{h \in A_{jk}} \beta_{h \to k} \varepsilon_h + \sum_{h \in A_{kj^*}} \beta_{h \to k} \varepsilon_h,\end{aligned}$$

where $A_{jk} = \mathrm{An}(j, G) \cap \mathrm{An}(k, G)$, $A_{jk^*} = \mathrm{An}(j, G) \cap \mathrm{An}(k, G)^c$ and similarly for A_{kj^*}.

We treat the two terms of (1.4.2) separately. Consider the first term,

$$\Psi^+_{jk} = \lim_{x \to \infty} \frac{1}{2} \mathbb{E}\left[\sigma(F_k(X_k)) \mid X_j > x\right], \quad j, k \in V.$$

Since X_j, $j \in V$, are regularly varying with index $\alpha > 0$, using similar arguments as in Lemma 1.2, we can write

$$\begin{aligned}&\mathbb{E}[\sigma(F_k(X_k)) \mathbf{1}\{X_j > x\}] \\ &\quad = \sum_{h \in \mathrm{An}(j,G)} \mathbb{E}\left[\sigma(F_k(X_k)) \mathbf{1}\left\{\beta_{h \to j} \varepsilon_h > x\right\}\right] + o\{\mathbb{P}(X_j > x)\} \\ &\quad = \sum_{h \in A_{jk}} \mathbb{E}\left[\sigma(F_k(X_k)) \mathbf{1}\left\{\beta_{h \to j} \varepsilon_h > x\right\}\right] \\ &\qquad + \sum_{h \in A_{jk^*}} \mathbb{E}\left[\sigma(F_k(X_k)) \mathbf{1}\left\{\beta_{h \to j} \varepsilon_h > x\right\}\right] + o\{\mathbb{P}(X_j > x)\}.\end{aligned} \tag{A.2.21}$$

Recall that σ is defined as

$$\sigma(x) = |2x - 1| = \begin{cases} 2x - 1, & x \geq 1/2 \\ 1 - 2x, & x < 1/2, \end{cases}$$

and that $0 \leq \sigma(F_k(X_k)) < 1$, $k \in V$.

Consider the summands in (A.2.21) where $h \in A_{jk}$. We distinguish two cases. When the ratio $c = \beta_{h\to k}/\beta_{h\to j} > 0$, we can bound the summand

$$\begin{aligned}
\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right) &\geq \mathbb{E}\left[\sigma(F_k(X_k))\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] \\
&\geq \mathbb{E}\left[\sigma(F_k(X_k))\mathbf{1}\left\{\beta_{h\to k}\varepsilon_h > cx, X_k > cx\right\}\right] \\
&\geq \sigma(F_k(cx))\mathbb{P}\left(\beta_{h\to k}\varepsilon_h > cx, X_k > cx\right) \\
&= \sigma(F_k(cx))\left[\mathbb{P}\left(\beta_{h\to k}\varepsilon_h > cx\right) - \mathbb{P}\left(\beta_{h\to k}\varepsilon_h > cx, X_k \leq cx\right)\right] \\
&= \mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right) + o\{\mathbb{P}(X_j > x)\}.
\end{aligned}$$

The last equality follows from Lemma A.3 and since $\sigma(F_k(cx)) \to 1$ as $x \to \infty$. When the ratio $c = \beta_{h\to k}/\beta_{h\to j} < 0$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right) &\geq \mathbb{E}\left[\sigma(F_k(X_k))\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] \\
&\geq \mathbb{E}\left[\sigma(F_k(X_k))\mathbf{1}\left\{\beta_{h\to k}\varepsilon_h < cx, X_k < cx\right\}\right] \\
&\geq \sigma(F_k(cx))\mathbb{P}\left(\beta_{h\to k}\varepsilon_h < cx, X_k < cx\right) \\
&= \sigma(F_k(cx))\left[\mathbb{P}\left(\beta_{h\to k}\varepsilon_h < cx\right) - \mathbb{P}\left(\beta_{h\to k}\varepsilon_h < cx, X_k \geq cx\right)\right] \\
&= \mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right) + o\{\mathbb{P}(X_j > x)\},
\end{aligned}$$

where the third inequality holds because, as $x \to \infty$, $F_k(X_k) < F_k(cx)$ implies $\sigma(F_k(X_k)) > \sigma(F_k(cx))$.

On the other hand, for all the summands in (A.2.21) where $h \in A_{jk^*}$ we have that X_k and ε_h are independent. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sigma(F_k(X_k))\mathbf{1}\left\{\beta_{h\to j}\varepsilon_h > x\right\}\right] &= \mathbb{E}\left[|2F_k(X_k) - 1|\right]\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right) \\
&= \frac{1}{2}\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right).
\end{aligned}$$

Consequently,

$$\begin{aligned}
\Psi^+_{jk} &= \lim_{x\to\infty}\frac{1}{2}\mathbb{E}\left[\sigma(F_k(X_k)) \mid X_j > x\right] \\
&= \lim_{x\to\infty}\frac{1}{2}\sum_{h\in A_{jk}}\frac{\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right)}{\mathbb{P}(X_j > x)} + \lim_{x\to\infty}\frac{1}{4}\sum_{h\in A_{jk^*}}\frac{\mathbb{P}\left(\beta_{h\to j}\varepsilon_h > x\right)}{\mathbb{P}(X_j > x)} \\
&= \frac{1}{4} + \frac{1}{4}\sum_{h\in A_{jk}}\lim_{x\to\infty}\frac{\mathbb{P}(\beta_{h\to j}\varepsilon_h > x)}{\mathbb{P}(X_j > x)} = \frac{1}{4} + \frac{1}{4}\frac{\sum_{h\in A_{jk}}|\beta_{h\to j}|^\alpha c^+_{hj}\ell(x)x^{-\alpha}}{\sum_{h\in \mathrm{An}(j,G)}|\beta_{h\to j}|^\alpha c^+_{hj}\ell(x)x^{-\alpha}} \\
&= \frac{1}{4} + \frac{1}{4}\frac{\sum_{h\in A_{jk}} c^+_{hj}|\beta_{h\to j}|^\alpha}{\sum_{h\in \mathrm{An}(j,G)} c^+_{hj}|\beta_{h\to j}|^\alpha}.
\end{aligned}$$

Similarly, the second term can be shown to be

$$\Psi^-_{jk} = \frac{1}{4} + \frac{1}{4}\frac{\sum_{h\in A_{jk}} c^-_{hj}|\beta_{h\to j}|^\alpha}{\sum_{h\in \mathrm{An}(j,G)} c^-_{hj}|\beta_{h\to j}|^\alpha}.$$

Putting everything together yields the desired form of $\Psi_{jk} = \Psi^+_{jk} + \Psi^-_{jk}$. □

A.3 Example of the EASE algorithm

Consider the DAG G in Figure A.1, with vertex set $V = \{1, 2, 3, 4\}$. The set of causal orders of G is $\Pi_G = \{(2, 1, 4, 3), (2, 1, 3, 4), (2, 3, 1, 4)\}$. In Figure A.2, we display the state-space tree of the EASE algorithm, i.e., the set of states that the algorithm can visit to find a causal order π of G. Each state represents the status of the vector π^{-1} during the algorithm evaluation. A state is red if all the paths below it lead to wrong causal order. The green states represent the causal orders of G.

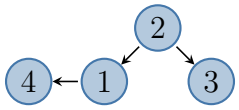


Figure A.1: DAG G.

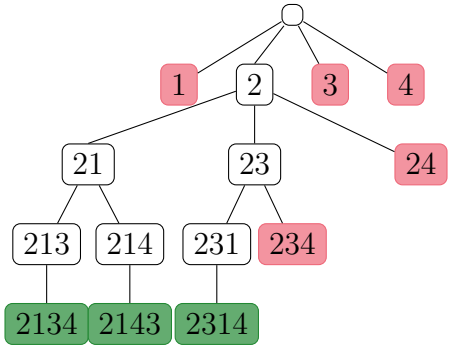


Figure A.2: Extremal ancestral search (EASE) for the DAG shown in Figure A.1.

A.4 Experimental settings for the simulation study

The parameters of the simulation are the following.

- Distribution: Student's t, with degrees of freedom $\alpha \in \{1.5, 2.5, 3.5\}$.
- Number of observations: $n \in \{500, 1000, 10000\}$.
- Number of variables: $p \in \{4, 7, 10, 15, 20, 30, 50\}$.

The settings that we consider are,

- Linear SCM,
- Linear SCM with hidden confounders,
- Nonlinear SCM,
- Linear SCM and uniform transformation of each variable.

For each combination of n, p, and α and each setting, we generate $n_{exp} = 50$ random SCMs. Each SCM is built as follows.

1. Generate a random DAG.
   (a) Take a random permutation π of the nodes $V = \{1, \dots, p\}$ that defines the causal order.
   (b) For each $i \in V$ such that $\pi(i) > 1$, sample the number of parents
   $$n_{\text{pa}} \sim \text{Bin}(\pi(i) - 1, q),$$
   from a binomial distribution. We set $q = \min\{5/(p-1), 1/2\}$ so that, on average, there are 2.5 edges per node, when $p > 10$.
   (c) Sample without replacement n_{pa} from $\{j \in V : \pi(j) < \pi(i)\}$ and name the resulting set $\text{pa}(i)$.
2. Sample uniformly from $\{-0.9, -0.1\} \cup \{0.1, 0.9\}$ the coefficients β_{ij}, where $i \in V$ and $j \in \text{pa}(i, G)$.
3. In the case of hidden confounders,
   (a) Sample the number of confounding variables,
   $$n_{\text{conf}} \sim \text{Bin}\left(\frac{p(p-1)}{2}, q\right),$$
   from a binomial distribution. We set $q = 2/(3p-3)$ so that, on average, there is one hidden confounder for every three nodes.
   (b) Sample without replacement n_{conf} unordered pairs from $\{\{i, j\} : i, j \in V\}$ and name the resulting set C.
   (c) Update the parents of each node i as $\text{pa}(i) := \text{pa}(i) \cup C_i$, where $C_i \subseteq C$ is the set of hidden confounders affecting node $i \in V$. Similarly, for each hidden confounder $c \in C$, set $\text{pa}(c) = \varnothing$.
   (d) Sample uniformly from $\{-0.9, -0.1\} \cup \{0.1, 0.9\}$ the coefficients β_{ic}, where $i \in V$ and $c \in C_i$.
4. Let $\widetilde{V} = V \cup C$[1]. For all $i \in \widetilde{V}$, sample n i.i.d. copies of noise $\varepsilon_i \sim$ Student's t with df $=$

[1] If there are no hidden confounders, then $C = \varnothing$.

α.

5. For each node $i \in \widetilde{V}$, generate

$$X_i := \sum_{j \in \mathrm{pa}(i)} \beta_{ij}\ f(X_j) + \varepsilon_i, \quad \text{where}$$

   (a) in case of linear SCM, $f(X_j) = X_j$,

   (b) in case of nonlinear SCM, $f(X_j) = X_j\ \mathbf{1}\{\widehat{F}_j(X_j) > 0.95\}$, where $\widehat{F}_j$ is the empirical cdf of X_j.

6. In case of uniform margins, transform each variable by $X_i := \widehat{F}_i(X_i)$, where $\widehat{F}_i$ is the empirical cdf of X_i, $i \in \widetilde{V}$.

A.5 Additional Figures and Tables

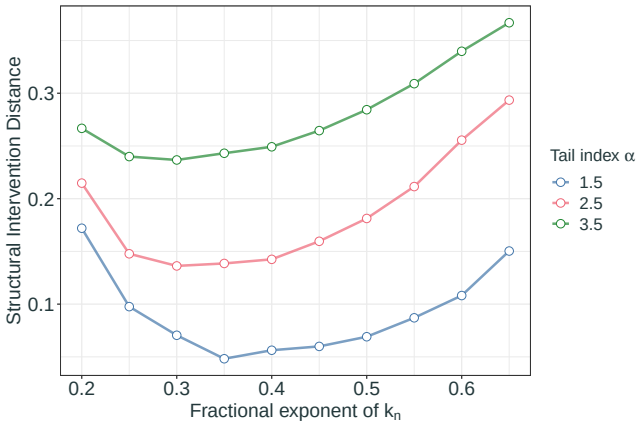


Figure A.3: The figure refers to Section 1.5.1. It shows the structural intervention distance (SID) of the EASE algorithm for different fractional exponents $\nu \in [0.2, 0.7]$ of $k_n = \lfloor n^\nu \rfloor$ and different tail indices $\alpha \in \{1.5, 2.5, 3.5\}$. Each point represents the SID measure averaged over 10 random samples for different sample sizes $n \in \{500, 1000, 10000\}$ and dimensions $p \in \{4, 7, 10, 15, 20, 30, 50\}$ in a linear SCM. In the experiments of Section 1.5.1, we set $k_n = \lfloor n^{0.4} \rfloor$.

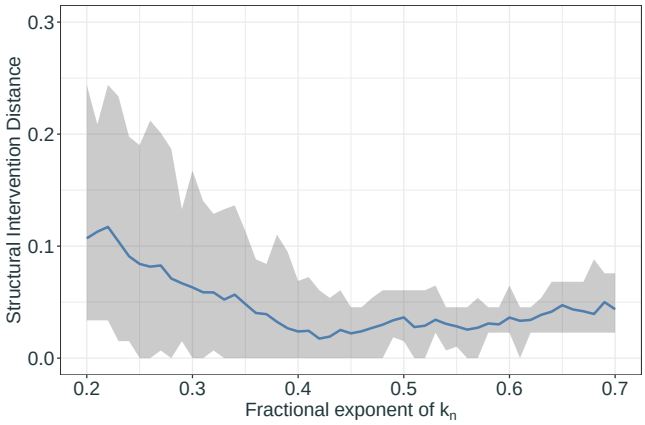


Figure A.4: The figure refers to Section 1.5.3. It shows the robustness of the structural intervention distance (SID) for varying fractional exponents $\nu \in [0.2, 0.7]$ of $k_n = \lfloor n^\nu \rfloor$. Each point represents an average over 50 SID evaluations for the EASE algorithm, after bootstrapping the original dataset. The shaded interval corresponds to the 90% bootstrap confidence interval.

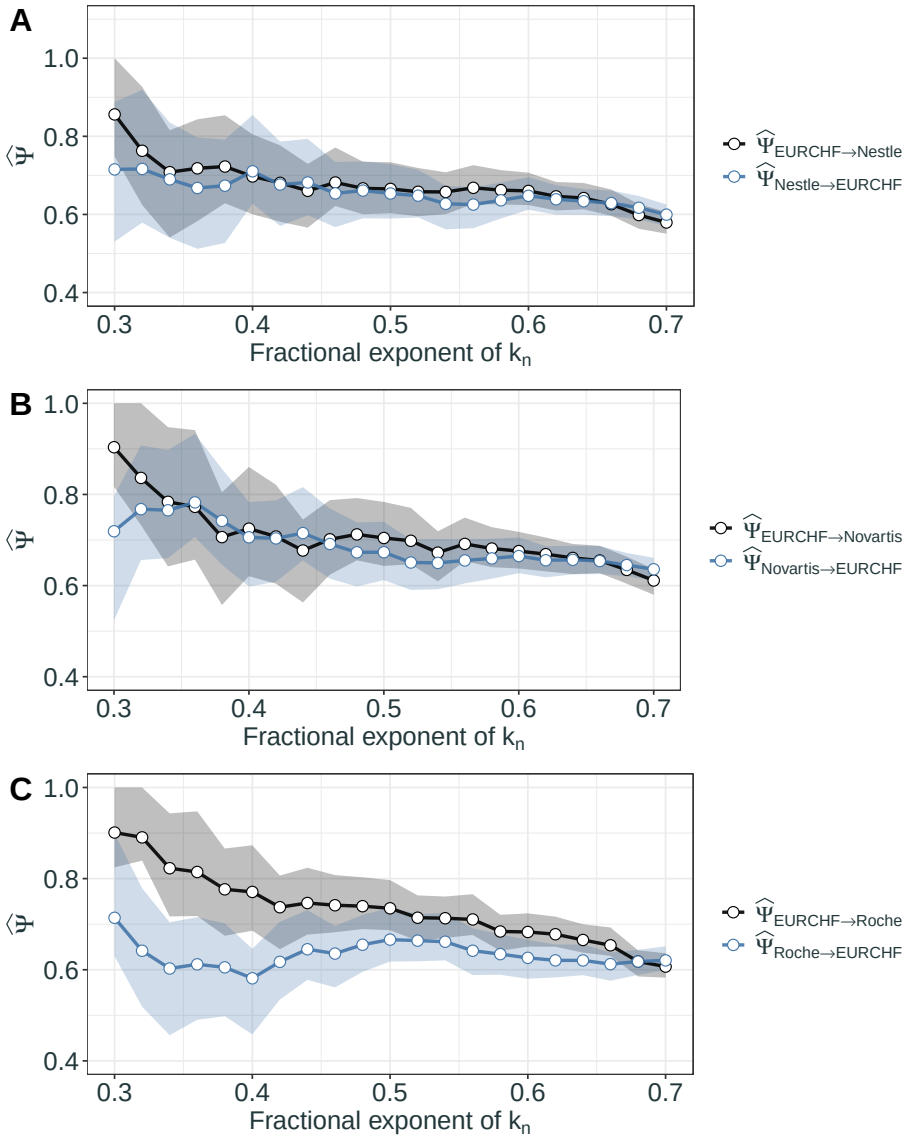


Figure A.5: The figure refers to Section 1.5.2. It shows the variability of the estimated coefficients $\widehat{\Psi}$ for different fractional exponents $\nu \in [0.2, 0.7]$ of $k_n = \lfloor n^\nu \rfloor$. Each point represents the estimates of $\widehat{\Psi}$ based on the full dataset. The shaded intervals correspond to the 90% bootstrap confidence intervals over 1000 repetitions.

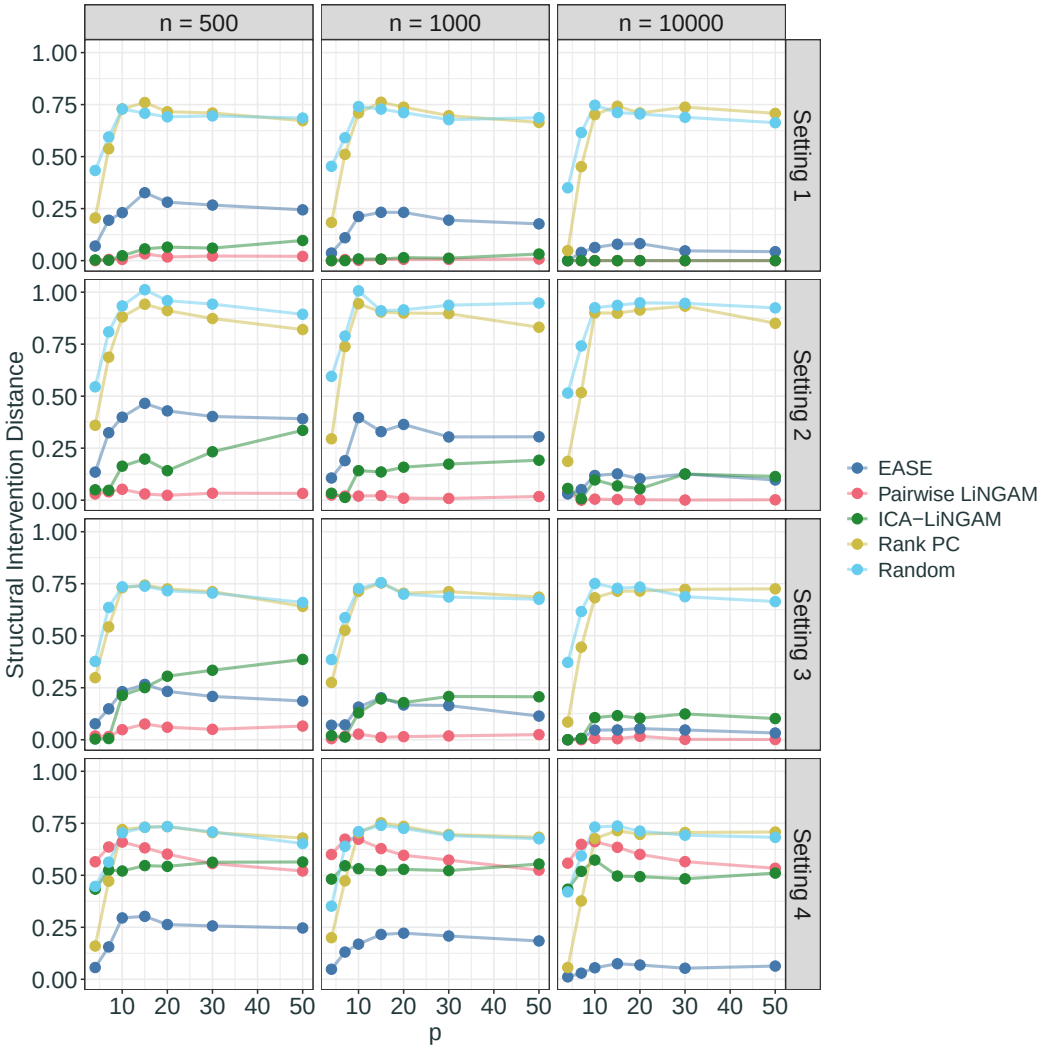


Figure A.6: The figure refers to Section 1.5.1. It shows the SID averaged over 50 simulations, for each method, setting, sample size n and dimension p, when the tail index is $\boldsymbol{\alpha = 2.5}$. Each row of the figure corresponds to one setting. In order, the settings are: (1) Linear SCM; (2) Linear SCM with hidden confounders; (3) Nonlinear SCM; (4) Linear SCM where each variable is transformed to a uniform margin.

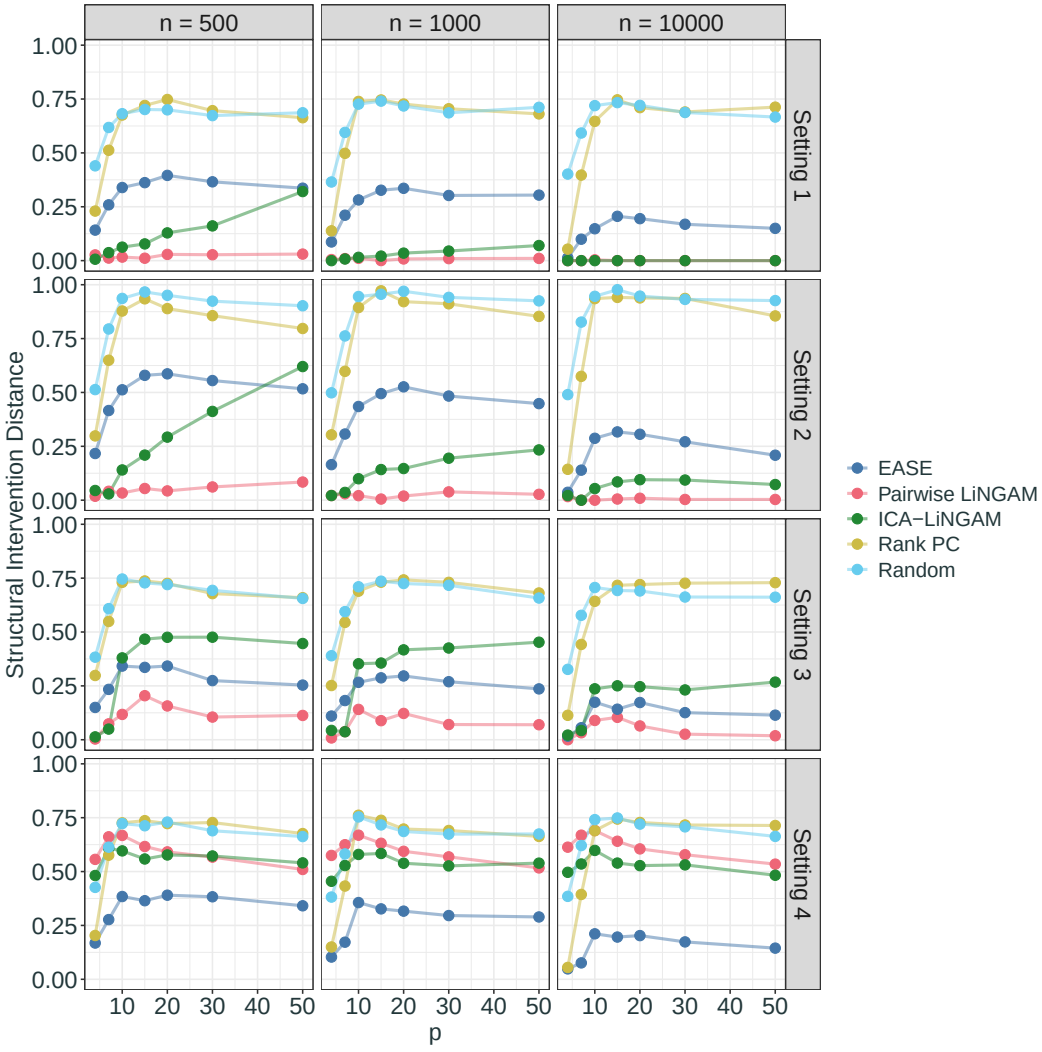


Figure A.7: The figure refers to Section 1.5.1. It shows the SID averaged over 50 simulations, for each method, setting, sample size n and dimension p, when the tail index is $\boldsymbol{\alpha = 3.5}$. Each row of the figure corresponds to one setting. In order, the settings are: (1) Linear SCM; (2) Linear SCM with hidden confounders; (3) Nonlinear SCM; (4) Linear SCM where each variable is transformed to a uniform margin.

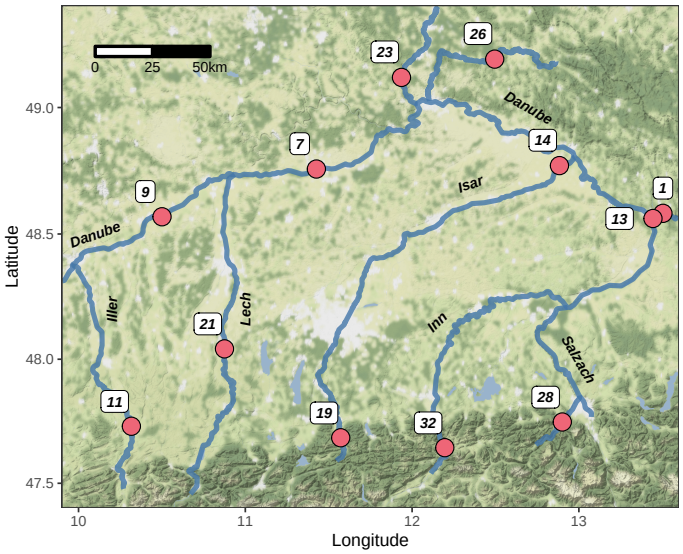


Figure A.8: The figure refers to Section 1.5.3. It represents the map of the upper Danube basin. The plot is created with the ggmap `R` package developed by Kahle and Wickham (2013). The background is taken from maps.stamen.com.

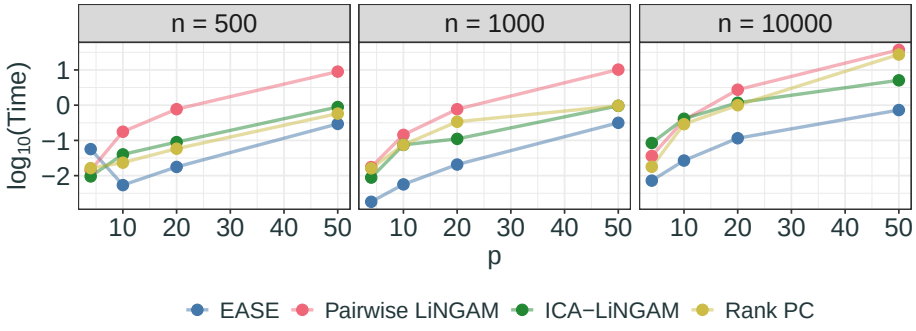


Figure A.9: The figure refers to Section 1.5.1. It shows the base-10 logarithm of the computational time (in seconds) averaged over 10 simulations, for each method, sample size n and dimension p.

Table A.5.1: The table refers to Section 1.5.1. It displays the average SID and the standard error (SE) over 50 simulations, for ICA-LiNGAM and Pairwise LiNGAM. For each dimension p we simulate $n = 10000$ observations from a non-linear SCM with Student's-t noise with $\alpha = 3.5$ degrees of freedom (see Section 1.5.1). We run each method both on the full dataset ('Full dataset') and on the partial dataset, where we keep only the observations in the tails of the distribution ('Keep tails'). As dimension p increases, the number of extreme observations in all their coordinates decreases exponentially. Therefore, for a given dimension p, we say that an observation $x \in \mathbb{R}^p$ lies in the tails of the distribution if at least $\lfloor \sqrt{p} \rfloor$ of its coordinates are below (or above) the 10% (or 90%) quantile.

		Full dataset		*Keep tails*	
		SID	*SE*	*SID*	*SE*
ICA-LiNGAM	$p = 10$	0.236	0.029	0.110	0.021
	$p = 20$	0.246	0.022	0.150	0.017
	$p = 30$	0.231	0.017	0.148	0.014
	$p = 50$	0.268	0.014	0.216	0.012
Pairwise LiNGAM	$p = 10$	0.090	0.024	0.003	0.001
	$p = 20$	0.064	0.013	0.003	0.001
	$p = 30$	0.026	0.007	0.005	0.003
	$p = 50$	0.019	0.004	0.006	0.002

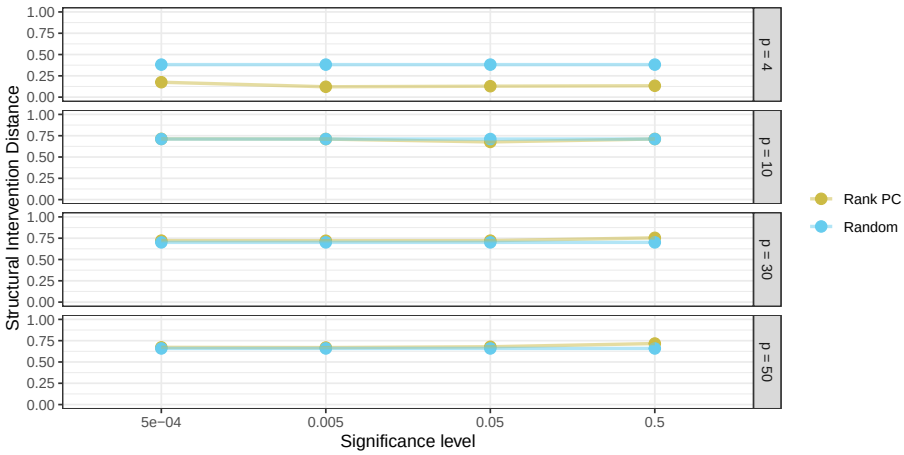


Figure A.10: The figure refers to Section 1.5.1. It shows the average SID of the Rank PC algorithm over 50 simulations for different significance levels of the independent test. For each dimension p we simulate $n = 1000$ observations from a linear SCM with Student's-t noise with $\alpha = 3.5$ degrees of freedom (see Section 1.5.1). For comparison, we also display the SID associated with random guessing.

A.6 Financial application

Starting from the financial application of Section 1.5.2, we study the dynamic evolution of the $\widehat{\Psi}$ coefficient across time. By looking at the time series of the EURCHF in Figure A.11,

we notice two periods of extremely high positive and negative returns, namely, August 2011 and January 2015. The second event was due to an unexpected decision of the Swiss National Bank (SNB) to remove the peg of 1.20 Swiss francs per Euro.

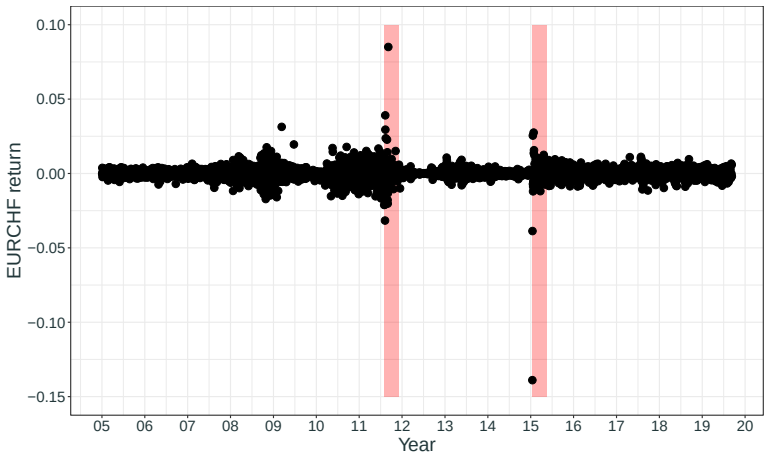


Figure A.11: Daily returns of the EURCHF exchange rate. The red sections represent the two major turmoil events. They occurred, respectively, in the month of August 2011 and in January 2015. The latter event was caused by the unexpected decision of the Swiss National Bank (SNB) to remove the peg of 1.20 Swiss francs per Euro.

The idea is to estimate the $\widehat{\Psi}$ coefficients between the EURCHF and the three stocks on a rolling window of 1500 days. In this case, we use a threshold $k = 10$ which corresponds approximately to a fractional exponent $\nu = 0.3$, where we define $k = \lfloor n^{\nu} \rfloor$ and n is the number of observations in the sample. In Figure A.12, we notice that during turmoil periods, highlighted in red, the $\widehat{\Psi}$ coefficient is higher in the direction that goes from the EURCHF to the stocks. This is an example where the causal structure becomes clearer during extreme events. It is also interesting to observe that the $\widehat{\Psi}$ coefficient is quite stable during calmer market periods. For example, the stable black lines between 2012 and 2015 correspond to the currency peg maintained by the SNB. As a last note, the drop in the black lines in mid 2017 is due to the fact that the extreme event of 2011 'exits' the rolling window.

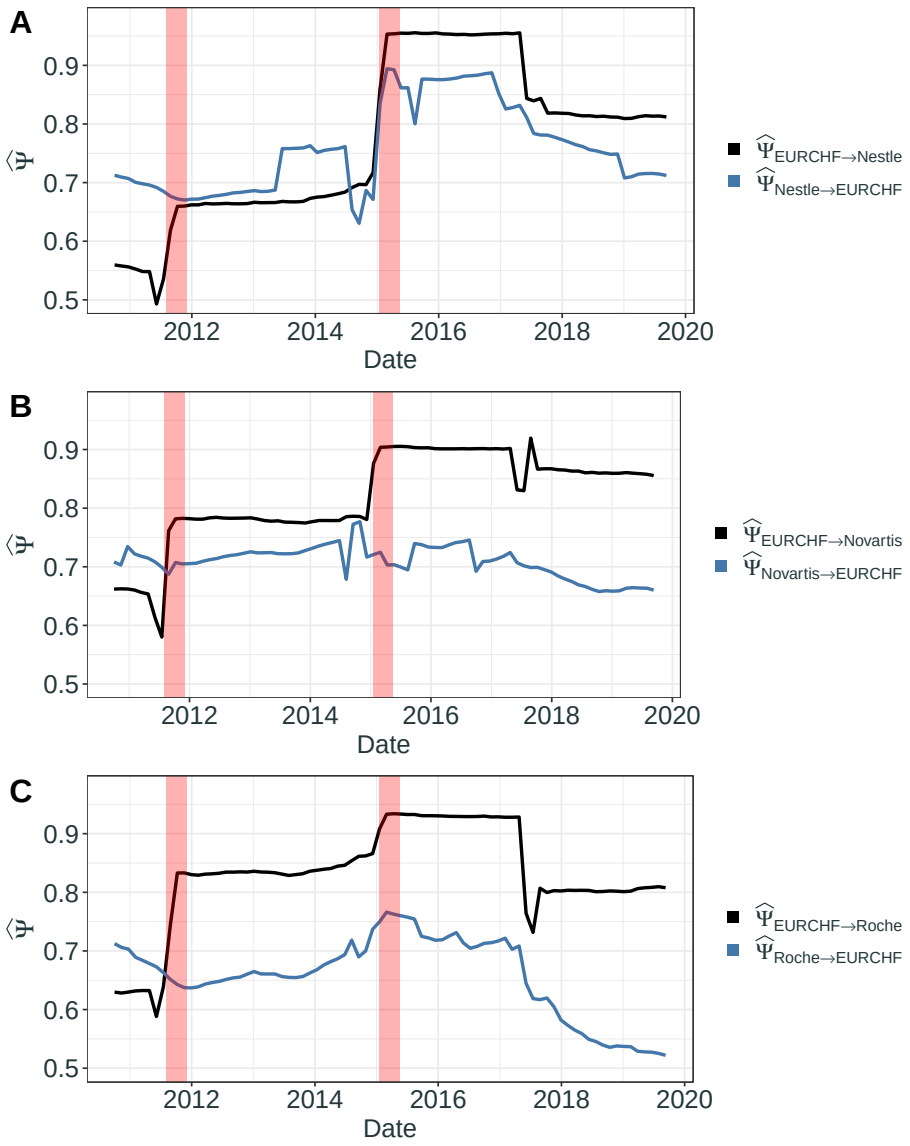


Figure A.12: Estimated coefficients $\widehat{\Psi}$ on a rolling window of 1500 days. The threshold used to estimate the coefficient is $k = 10$.

Appendix B

Extremal Random Forests

B.1 Proof of Theorem 2.4

Given the data generating process of Assumption 2.1 in the main text, define the random variable $Z = (Y - Q_X(\tau_0))_+$. We then have the stochastic representation

$$(X, Z, 1\{Z > 0\}) \stackrel{d}{=} (X, VP, P), \tag{B.1.1}$$

where V follows a GPD with parameter vector $\theta(X)$, and $P \sim \text{Bernoulli}(1 - \tau_0)$, independent of X and V. Similarly, for the training data (X_i, Y_i) we may use a an analogous representation with $(X_i, V_i P_i, P_i)$ as in (B.1.1), $i = 1, \ldots, n$. With this we can rewrite the weighted (negative) log-likelihood function in (2.3.3) as

$$L_n(\theta; x) = \sum_{i=1}^{n} w_n(x, X_i)\ell_\theta(V_i)P_i,$$

Moreover, for a fixed predictor value $x \in \text{Int } \mathcal{X}$ let V^* denote a GPD with parameter vector $\theta(x)$ and define $L(\theta; x) = \mathbb{E}[\ell_\theta(V^*)P]$, where $\theta \in (0, \infty)^2$. To prove our result we rely on Theorem 5.7 of van der Vaart (1998), which we state here adapted to our setting.

Theorem B.1. *Let $\theta \mapsto L_n(\theta; x)$ be random functions, and let $\theta \mapsto L(\theta; x)$ be a fixed function such that, for $x \in \text{Int } \mathcal{X}$, it holds*

$$\sup_{\theta \in \Theta} |L_n(\theta; x) - L(\theta; x)| \xrightarrow{\mathbb{P}} 0, \tag{B.1.2}$$

$$L(\theta(x); x) < \inf \{L(\theta; x) : \|\theta - \theta(x)\|_2 \geq \delta,\ \theta \in \Theta\}, \text{ for all } \delta > 0. \tag{B.1.3}$$

Then any sequence of estimators $\hat{\theta}(x)$ with $L_n(\hat{\theta}(x); x) \leq L_n(\theta(x); x) + o_P(1)$ converges in probability to $\theta(x)$.

We can now prove our Theorem 2.4.

Proof of Theorem 2.4. First, notice that $\theta(x) \in \theta(\mathcal{X}) \subset \Theta$, where Θ is compact. Therefore, from (2.3.4) in the main text, we have that $L_n(\hat{\theta}(x); x) \leq L_n(\theta(x); x)$ for all $n > 0$. Furthermore, a standard argument using Kullback–Leibler divergence implies the true parameter $\theta(x)$ is a minimizer for $\theta \mapsto L(\theta; x)$. Since the GPD is identifiable, the true parameter is a unique minimizer, satisfying condition (B.1.3). Moreover, from Lemma B.2, condition (B.1.2) is satisfied.

Therefore, from Theorem B.1, the estimator $\hat{\theta}(x) \to \theta(x)$ in probability as $n \to \infty$. □

Lemma B.2. *Under the assumptions of Theorem 2.4, it holds that* $\sup_{\theta\in\Theta}|L_n(\theta;x) - L(\theta;x)| \xrightarrow{\mathbb{P}} 0$.

Proof. We have that

$$\begin{aligned} L_n(\theta;x) &= \sum_{i=1}^{n} w_n(x,X_i)\ell_\theta(V_i)P_i \\ &= \sum_{i=1}^{n} w_n(x,X_i)\ell_\theta(V_i^*)P_i + \sum_{i=1}^{n} w_n(x,X_i)\left(\ell_\theta(V_i) - \ell_\theta(V_i^*)\right)P_i \\ &= S_{1,n}(\theta) + S_{2,n}(\theta), \end{aligned}$$

where we couple the random variables V_i and V_i^* through $V_i = F^{-1}_{\theta(X_i)}(U_i)$, $V_i^* = F^{-1}_{\theta(x)}(U_i)$, $U_i \overset{iid}{\sim} \text{Unif}[0,1]$, and F_θ^{-1} is the inverse of the GPD function with parameter $\theta \in \Theta$. By Lemma B.3 and B.7, the claim follows. □

Lemma B.3. *Under the assumptions of Theorem 2.4, it holds that* $\sup_{\theta\in\Theta} |S_{1,n}(\theta) - L(\theta;x)| \xrightarrow{\mathbb{P}} 0$.

Proof. Corollary 2.2 of Newey (1991) provides sufficient conditions for uniform convergence.

1. (Compactness): Θ is compact.
2. (Pointwise convergence): For each $\theta \in \Theta$, $S_{1,n}(\theta) - L(\theta;x) = o_P(1)$.
3. (Stochastic Equicontinuity): There exists $C_n = O_P(1)$ such that for all $\theta, \theta' \in \Theta$, $|S_{1,n}(\theta) - S_{1,n}(\theta')| \leq C_n\|\theta - \theta'\|_2$.
4. (Continuity): The map $\theta \mapsto L(\theta;x)$ is continuous.

Condition 1 holds by assumption. The remaining conditions are shown in Lemmas B.4, B.5, and B.6, respectively. □

Lemma B.4. *For each* $\theta \in \Theta$, *it holds that* $S_{1,n}(\theta) - L(\theta;x) = o_P(1)$.

Proof. For each $\theta \in \Theta$, recall that $L(\theta;x) = \mathbb{E}[\ell_\theta(V^*)P]$, where $V^* \sim \text{GPD}(\theta(x))$. Furthermore, we have that

$$S_{1,n}(\theta) = \sum_{i=1}^{n} w_n(x,X_i)\ell_\theta(V_i^*)P_i = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} w_{n,b}(x,X_i)\ell_\theta(V_i^*)P_i = \frac{1}{B}\sum_{b=1}^{B} T_{n,b}(x,\theta),$$

where $T_{n,b}(x,\theta)$ is the output of a regression gradient tree (Athey et al., 2019) with response $\ell_\theta(V_i^*)P_i$ independent of $w_{n,b}(x,X_i)$, $i = 1,\ldots,n$. Consider a tree $T_{n,b}(x,\theta)$, $b = 1,\ldots,B$, of the generalized random forest. Its expectation writes

$$\begin{aligned} \mathbb{E}\left(T_{n,b}(x,\theta)\right) &= \sum_{i=1}^{n}\mathbb{E}\left(w_{n,b}(x,X_i)\ell_\theta(V_i^*)P_i\right) = \sum_{i=1}^{n}\mathbb{E}\left(w_{n,b}(x,X_i)\right)\mathbb{E}\left(\ell_\theta(V_i^*)P_i\right) \\ &= \mathbb{E}\left\{\sum_{i=1}^{n} w_{n,b}(x,X_i)\right\}\mathbb{E}\left(\ell_\theta(V^*)P\right) = \mathbb{E}\left(\ell_\theta(V^*)P\right) = L(\theta;x), \end{aligned} \tag{B.1.4}$$

since the weights sum to one. Therefore, $\mathbb{E}\left(S_{1,n}(\theta)\right) = L(\theta; x)$. Concerning the variance of the thinning $\ell_\theta(V^*)P$ we have that

$$\mathbb{V}\left(\ell_\theta(V^*)P\right) = \mathbb{E}\left(\ell_\theta(V^*)^2\right)\mathbb{E}(P^2) - \mathbb{E}\left(\ell_\theta(V^*)\right)^2\mathbb{E}\left(P\right)^2 < +\infty, \tag{B.1.5}$$

since P is a Bernoulli variable and $\ell_\theta(V^*)$ has exponential tail. Therefore, the variance of $T_{n,b}(x,\theta)$ writes

$$\begin{aligned}
\mathbb{V}\left(T_{n,b}(x,\theta)\right) &= \mathbb{E}\left\{\left(T_{n,b}(x,\theta) - L(\theta;x)\right)^2\right\} = \mathbb{E}\left\{\left(\sum_{i=1}^n w_{n,b}(x,X_i)\left(\ell_\theta(V_i^*)P_i - L(\theta;x)\right)\right)^2\right\} \\
&= \mathbb{E}\left(\sum_{i=1}^n w_{n,b}(x,X_i)^2\left(\ell_\theta(V_i^*)P_i - L(\theta;x)\right)^2\right. \\
&\quad \left. + \sum_{i\neq j} w_{n,b}(x,X_i)w_{n,b}(x,X_j)(\ell_\theta(V_i^*)P_i - L(\theta;x))(\ell_\theta(V_j^*)P_j - L(\theta;x))\right) \\
&= \mathbb{V}\left(\ell_\theta(V^*)P\right)\mathbb{E}\left(\sum_{i=1}^n w_{n,b}(x,X_i)^2\right) \leq \mathbb{V}(\ell_\theta(V^*)P) < +\infty,
\end{aligned} \tag{B.1.6}$$

where the fourth equality holds because (V_i^*, P_i) are i.i.d. , the second last inequality holds because $0 \leq w_{n,b}(x,X_i) \leq 1$, and the last inequality follows from (B.1.5). Using results about U-statistics (Hoeffding, 1948), Wager and Athey (2018) show that the variance of a forest is at most s/n times the variance of a tree, that is

$$\limsup_{n\to\infty} \frac{n}{s}\frac{\mathbb{V}\left(S_{1,n}(\theta)\right)}{\mathbb{V}\left(T_{n,b}(x,\theta)\right)} \leq 1. \tag{B.1.7}$$

where $s < n$ denotes the subsample size. From Assumption 2.3, we have that $s/n \to 0$, therefore (B.1.6) and (B.1.7) imply that $\mathbb{V}\left(S_{1,n}(\theta)\right) \to 0$ as $n \to 0$. The result follows from Markov's inequality. □

Lemma B.5. *There exists $C_n = O_P(1)$ such that for all θ, $\theta' \in \Theta$, $|S_{1,n}(\theta) - S_{1,n}(\theta')| \leq C_n\|\theta - \theta'\|_2$.*

Proof. The negative log-likelihood $\theta \mapsto \ell_\theta(z)$ is defined for each $z \geq 0$ and $\theta \in (0,\infty)^2$ as

$$\ell_\theta(z) = \log\sigma + \left(1 + \frac{1}{\xi}\right)\log\left(1 + \frac{\xi}{\sigma}z\right).$$

Therefore, its partial derivatives can be bounded by

$$\begin{aligned}
\left|\partial_\xi \ell_\theta(z)\right| &\leq \frac{1}{\xi^2}\log\left(1 + \frac{\xi}{\sigma}z\right) + \frac{1 + \frac{1}{\xi}}{\xi}, \\
\left|\partial_\sigma \ell_\theta(z)\right| &\leq \frac{1}{\sigma} + \frac{1 + \frac{1}{\xi}}{\sigma},
\end{aligned} \tag{B.1.8}$$

for any $\theta = (\sigma, \xi) \in (0,\infty)^2$. The bounds from (B.1.8) are continuous on the compact set $\Theta \subset (0,\infty)^2$, and therefore, from an application of the dominated convergence theorem,

$$g(z) := \sup\left\{\left|\partial_\xi \ell_\theta(z)\right| : \theta \in \Theta\right\} + \sup\left\{\left|\partial_\sigma \ell_\theta(z)\right| : \theta \in \Theta\right\} \tag{B.1.9}$$

is integrable with respect to a GPD with parameter vector $\theta(x)$. Moreover, for any $\theta, \theta' \in \Theta$, the mean-value theorem and the Cauchy–Schwarz inequality imply

$$|\ell_\theta(z) - \ell_{\theta'}(z)| = \left|\nabla \ell_{\widetilde{\theta}}(z)(\theta - \theta')\right| \leq \|\nabla \ell_{\widetilde{\theta}}(z)\|_2 \|\theta - \theta'\|_2, \tag{B.1.10}$$

where $\widetilde{\theta} = c\theta + (1-c)\theta'$ for some $0 < c < 1$, and $z \geq 0$. Furthermore, from (B.1.9), we have that

$$\|\nabla \ell_{\widetilde{\theta}}(z)\|_2 \leq \left|\partial_\xi \ell_{\widetilde{\theta}}(z)\right| + \left|\partial_\sigma \ell_{\widetilde{\theta}}(z)\right| \leq g(z). \tag{B.1.11}$$

From equations (B.1.10) and (B.1.11) it follows that $\ell_\theta(z)$ is Lipschitz in $\theta \in \Theta$ with constant $g(z)$, $z \geq 0$. Therefore,

$$\begin{aligned} |S_{1,n}(\theta) - S_{1,n}(\theta')| &= \left|\sum_{i=1}^n w_n(x, X_i) \left(\ell_\theta(V_i^*) - \ell_{\theta'}(V_i^*)\right) P_i\right| \leq \sum_{i=1}^n w_n(x, X_i) P_i \left|\ell_\theta(V_i^*) - \ell_{\theta'}(V_i^*)\right| \\ &\leq \left(\sum_{i=1}^n w_n(x, X_i) g(V_i^*) P_i\right) \|\theta - \theta'\|_2 =: C_n \|\theta - \theta'\|_2. \end{aligned}$$

For every $n \in \mathbb{N}$ and $i = 1, \ldots, n$, V_i^* is independent of $w_n(x, X_i)$ and P_i. Therefore, since $z \mapsto g(z)$ is integrable with respect to a GPD with parameter vector $\theta(x)$ it follows that $\mathbb{E}[C_n] < +\infty$. Hence, $C_n = O_P(1)$. □

Lemma B.6. *The map $\theta \mapsto L(\theta; x)$ is continuous.*

Proof. For any $\theta \in \Theta$, recall that

$$L(\theta; x) = \mathbb{E}[\ell_\theta(V^*)P] = \left\{\log \sigma + \left(1 + \frac{1}{\xi}\right) \mathbb{E}\left(\log\left[1 + \frac{\xi}{\sigma} V^*\right]\right)\right\} (1 - \tau_0).$$

The maps $\theta \mapsto \log \sigma$ and $\theta \mapsto (1 + 1/\xi)$ are continuous for $\theta \in \Theta$. Also, by an application of the dominated convergence theorem, the map $\theta \mapsto \mathbb{E}\left[\log\left(1 + \frac{\xi}{\sigma} V^*\right)\right]$ is continuous for $\theta \in \Theta$. □

Lemma B.7. *Under the assumptions of Theorem 2.4, it holds that* $\sup_{\theta \in \Theta} |S_{2,n}(\theta)| \xrightarrow{\mathbb{P}} 0$.

Proof. We have that

$$\begin{aligned} 0 \leq \sup_{\theta \in \Theta} |S_{2,n}(\theta)| &= \sup_{\theta \in \Theta} \left|\sum_{i=1}^n w_n(x, X_i) P_i \left(\ell_\theta \circ F^{-1}_{\theta(X_i)}(U_i) - \ell_\theta \circ F^{-1}_{\theta(x)}(U_i)\right)\right| \\ &\leq \sup_{\theta \in \Theta} \sum_{i=1}^n w_n(x, X_i) P_i \left|\ell_\theta \circ F^{-1}_{\theta(X_i)}(U_i) - \ell_\theta \circ F^{-1}_{\theta(x)}(U_i)\right| \\ &\leq \sup_{\theta \in \Theta} \sum_{i=1}^n w_n(x, X_i) P_i K(\theta, U_i) \|X_i - x\|_2 \\ &\leq \sup\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\} \sum_{i=1}^n w_n(x, X_i) P_i \sup_{\theta \in \Theta} K(\theta, U_i) \\ &= o_P(1), \end{aligned} \tag{B.1.12}$$

where the second last inequality follows from Lemma B.8.a) and the last equality follows from Lemmas B.9 and B.8.b). □

Lemma B.8. *Let $x \in \operatorname{Int} \mathcal{X}$, $U \sim \operatorname{Unif}[0,1]$, and $\theta \in \Theta$.*

a) *Then, there exists a function $K(\theta, U) < +\infty$ such that for any $y \in \mathcal{X}$,*

$$\left|\ell_\theta \circ F^{-1}_{\theta(y)}(U) - \ell_\theta \circ F^{-1}_{\theta(x)}(U)\right| \leq K(\theta, U)\|y - x\|_2.$$

b) *Then, under the assumptions of Theorem 2.4, it holds that*

$$\sum_{i=1}^{n} w_n(x, X_i) P_i \sup_{\theta \in \Theta} K(\theta, U_i) = O_P(1).$$

Proof.

a) Let $U \sim \operatorname{Unif}[0,1]$, and $\theta \in \Theta$. For any $y \in \mathcal{X}$ define

$$g(y; \theta, W) := \ell_\theta \circ F^{-1}_{\theta(y)}(1 - 1/W) = \log \sigma + \left(1 + \frac{1}{\xi}\right) \log\left(1 + \frac{\xi}{\sigma}\frac{\sigma(y)}{\xi(y)}\left\{W^{\xi(y)} - 1\right\}\right), \tag{B.1.13}$$

where $W := 1/(1-U) \sim \text{Pareto}(1)$ with support $[1, \infty)$. The map $y \mapsto g(y; \theta, W)$ admits partial derivatives with respect to y_j, $j = 1, \ldots, p$, i.e.,

$$\begin{aligned}\partial_{y_j} g(y; \theta, W) = &\left(1 + \frac{1}{\xi}\right)\left(1 + \frac{\xi}{\sigma}\frac{\sigma(y)}{\xi(y)}\left\{W^{\xi(y)} - 1\right\}\right)^{-1} \frac{\xi}{\sigma} \\ &\times \left(\frac{\sigma_j'(y)\xi(y) - \sigma(y)\xi_j'(y)}{\xi(y)^2}\left\{W^{\xi(y)} - 1\right\} + \frac{\sigma(y)}{\xi(y)}\left\{W^{\xi(y)} \log W\right\}\xi_j'(y)\right),\end{aligned} \tag{B.1.14}$$

where σ_j', and ξ_j' are the jth partial derivatives of $y \mapsto \sigma(y)$ and $y \mapsto \xi(y)$, respectively. From Assumption 2.2 in the main text, we know that $y \mapsto \partial_{y_j} g(y; \theta, W)$ are continuous on the interior of $\mathcal{X}$. Thus, for $x \in \operatorname{Int} \mathcal{X}$ and $y \in \mathcal{X}$, the mean-value theorem and the Cauchy–Schwarz inequality imply

$$|g(y; \theta, W) - g(x; \theta, W)| \leq \|\nabla g(x'; \theta, W)\|_2 \|y - x\|_2,$$

where $x' = cy + (1-c)x$ for some $c \in (0,1)$. Moreover, Assumption 2.2 ensures that the partials derivatives of $y \mapsto g(y; \theta, W)$ exist on the compact set $\mathcal{X}$. Thus, we can define $K(\theta, U) := \sum_{j=1}^{p} \sup\{|\partial_{y_j} g(y; \theta, W)| : y \in \mathcal{X}\}$ and obtain

$$\left|\ell_\theta \circ F^{-1}_{\theta(y)}(U) - \ell_\theta \circ F^{-1}_{\theta(x)}(U)\right| \leq K(\theta, U)\|y - x\|_2.$$

b) From Part a), we have that $K(\theta, U) = \sum_{j=1}^{p} \sup\{|\partial_{y_j} g(y; \theta, W)| : y \in \mathcal{X}\}$, where $\theta \in \Theta$, and $W \geq 1$ follows a standard Pareto distribution. For every $j = 1, \ldots, p$ it holds that

$$\begin{aligned}\sup_{y \in \mathcal{X}} |\partial_{y_j} g(y; \theta, W)| \leq &\sup_{y \in \mathcal{X}} \left(1 + \frac{1}{\xi}\right)\left(1 + \frac{\xi}{\sigma}\frac{\sigma(y)}{\xi(y)}\left\{W^{\xi(y)} - 1\right\}\right)^{-1} \frac{\xi}{\sigma} \\ &\times \left(\frac{|\sigma_j'(y)\xi(y) - \sigma(y)\xi_j'(y)|}{\xi(y)^2}\left\{W^{\xi(y)} - 1\right\} + \frac{\sigma(y)}{\xi(y)}|\xi_j'(y)|\left\{W^{\xi(y)} \log W\right\}\right) \\ =: &\sup_{y \in \mathcal{X}} \left(1 + \frac{1}{\xi}\right)\frac{\xi}{\sigma}\left(\frac{M_{1j}(y)\left\{W^{\xi(y)} - 1\right\}}{1 + M(y, \theta)\left\{W^{\xi(y)} - 1\right\}} + \frac{M_{2j}(y)\left\{W^{\xi(y)} \log W\right\}}{1 + M(y, \theta)\left\{W^{\xi(y)} - 1\right\}}\right),\end{aligned}$$

where $M_{1j}(y) = |\sigma_j'(y)\xi(y) - \sigma(y)\xi_j'(y)|/\xi(y)^2 \geq 0$, $M_{2j}(y) = \sigma(y)|\xi_j'(y)|/\xi(y) \geq 0$, and $M(y,\theta) = (\sigma(y)\xi)/(\xi(y)\sigma) > 0$. Notice that almost surely

$$0 \leq \frac{M_{1j}(y)\left\{W^{\xi(y)} - 1\right\}}{1 + M(y,\theta)\left\{W^{\xi(y)} - 1\right\}} \leq \frac{M_{1j}(y)}{M(y,\theta)},$$
$$0 \leq \frac{M_{2j}(y)W^{\xi(y)}}{1 + M(y,\theta)\left\{W^{\xi(y)} - 1\right\}} \leq \max\left\{M_{2j}(y), \frac{M_{2j}(y)}{M(y,\theta)}\right\}.$$

Therefore, for every $j = 1, \ldots, p$, we have

$$\begin{aligned}\sup_{\theta\in\Theta,\ y\in\mathcal{X}} |\partial_{y_j} g(y;\theta,W)| &\leq \sup_{\theta\in\Theta,\ y\in\mathcal{X}} \left(1 + \frac{1}{\xi}\right)\frac{\xi}{\sigma}\left(\frac{M_{1j}(y)}{M(y,\theta)} + \max\left\{M_{2j}(y), \frac{M_{2j}(y)}{M(y,\theta)}\right\}\log W\right)\\ &\leq \left(1 + \frac{1}{\xi^-}\right)\frac{\xi^+}{\sigma^-}\left(\frac{M_{1j}}{M} + \max\left\{M_{2j}, \frac{M_{2j}}{M}\right\}\log W\right)\\ &= \left(1 + \frac{1}{\xi^-}\right)\frac{\xi^+}{\sigma^-}\left(\frac{M_{1j}}{M} + \frac{M_{2j}}{M}\log W\right),\end{aligned}$$

where $M_{hj} := \sup\{M_{hj}(y) : y \in \mathcal{X}\}$, for $h = 1, 2$, $M := \inf\{M(y,\theta) : \theta \in \Theta,\ y \in \mathcal{X}\} < 1$, and σ^+, ξ^+ (σ^-, ξ^-) are the maxima (minima) of the parameter values over the compact set Θ, respectively. Since $W \sim$ Pareto(1) with support $[1, \infty)$, it follows that $\log W \sim$ Exp(1). Therefore, by taking expectation we obtain

$$\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U)\right) \leq \left(1 + \frac{1}{\xi^-}\right)\frac{\xi^+}{\sigma^-}\sum_{j=1}^{p}\left(\frac{M_{1j} + M_{2j}}{M}\right) =: M^* < \infty.$$

Let $\varepsilon > 0$ and consider $M_\varepsilon = (M^* + 1)/\varepsilon > 0$. Then, for any $n \in \mathbb{N}$, it holds that

$$\mathbb{P}\left(\sum_{i=1}^{n} w_n(x, X_i)P_i \sup_{\theta\in\Theta} K(\theta, U_i)) > M_\varepsilon\right) \leq \frac{\mathbb{E}\left(\sum_{i=1}^n w_n(x, X_i)P_i \sup_{\theta\in\Theta} K(\theta, U_i)\right)}{M_\varepsilon}$$
$$= \frac{\sum_{i=1}^n \mathbb{E}\left(w_n(x, X_i)\right)\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U_i)\right)\mathbb{E}\left(P_i\right)}{M_\varepsilon} = \frac{\mathbb{E}\left(\sum_{i=1}^n w_n(x, X_i)\right)\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U)\right)\mathbb{E}\left(P\right)}{M_\varepsilon}$$
$$= \frac{\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U)\right)(1 - \tau_0)}{M_\varepsilon} \leq \frac{M^*(1 - \tau_0)}{M_\varepsilon} < \varepsilon.$$

□

Lemma B.9. *Under the assumptions of Theorem 2.4, it holds that*

$$\sup\left\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\right\} = o_P(1).$$

Proof. This result follows from Lemma 2 of Wager and Athey (2018) which states that $\operatorname{diam}(L_b(x)) = o_P(1)$. It does not require the random forest to be honest; i.e., we can assume that we use the same observations to place the splits and make predictions. For each tree $b = 1, \ldots, B$ of the forest, we subsample $\mathcal{S}_b \subset \{1, \ldots, n\}$ observations from the training data, with $|\mathcal{S}_b| = s < n$. Denote by $L_b(x) \subset \mathcal{X}$ the leaf containing the fixed predictor value $x \in \mathcal{X}$. Define the diameter $\operatorname{diam}(L_b(x)) := \sup_{z,y\in L_b(x)}\|z - y\|_2$ of the

leaf $L_b(x)$ as the length of the longest segment contained inside $L_b(x)$. Recall that the weights of a (not necessarily honest) random forest are defined as

$$w_n(x, X_i) = \frac{1}{B}\sum_{i=1}^{B} w_{n,b}(x, X_i) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{S}_b\}}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}.$$

Also, note that

$$\begin{aligned}&\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\} = \{\|X_i - x\|_2 : \exists b = 1, \ldots, B, X_i \in L_b(x), i \in \mathcal{S}_b\} \\ &= \cup_{b=1}^{B} \{\|X_i - x\|_2 : X_i \in L_b(x), i \in \mathcal{S}_b\} \subset \cup_{b=1}^{B} \{\|y - x\|_2 : y \in L_b(x)\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\sup\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\} \leq \sup \cup_{b=1}^{B} \{\|y - x\|_2 : y \in L_b(x)\} \\ &= \max_{b=1}^{B} \sup\{\|y - x\|_2 : y \in L_b(x)\} \leq \max_{b=1}^{B} \operatorname{diam}(L_b(x)).\end{aligned}$$

Thus, for every $\varepsilon > 0$

$$\begin{aligned}0 &\leq \mathbb{P}\left(\sup\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\} > \varepsilon\right) \\ &\leq \mathbb{P}\left(\max_{b=1}^{B} \operatorname{diam}(L_b(x)) > \varepsilon\right) \leq \sum_{b=1}^{B} \mathbb{P}\left(\operatorname{diam}(L_b(x)) > \varepsilon\right) \to 0.\end{aligned}$$

□

B.2 Partial Derivative on the Boundary

For any function $f : [0,1]^p \to \mathbb{R}$, we define the first order partial derivative on the boundary by

$$\partial_{x_j} f(x) := \begin{cases} \lim_{h \downarrow 0} \frac{f(x + h e_j)}{h}, & \text{if } x \in [0,1]^p, x_j = 0, \\ \lim_{h \downarrow 0} \frac{f(x) - f(x - h e_j)}{h}, & \text{if } x \in [0,1]^p, x_j = 1. \end{cases}$$

B.3 Weight Function Estimation

In quantile regression tasks, the weight function $(x, y) \mapsto w_n(x, y)$ estimated by GRF measures the similarity between x and y according to their conditional distribution.

Figure B.1 shows the localizing weights $w_n(x, X_i)$, $x, X_i \in \mathbb{R}^p$, for two test predictors x with $x_1 = -0.2, 0.5$, respectively. The data is generated according to Example 2.1, with $n = 2000$ observations and $p = 40$ predictors. In the left panel of Figure B.1, the observations (X_i, Y_i) with $X_{i1} < 0$ are the ones influencing most the test predictor x with $x_1 = -0.2$. This is because they share the same conditional distribution. A similar argument holds for the right panel of Figure B.1.

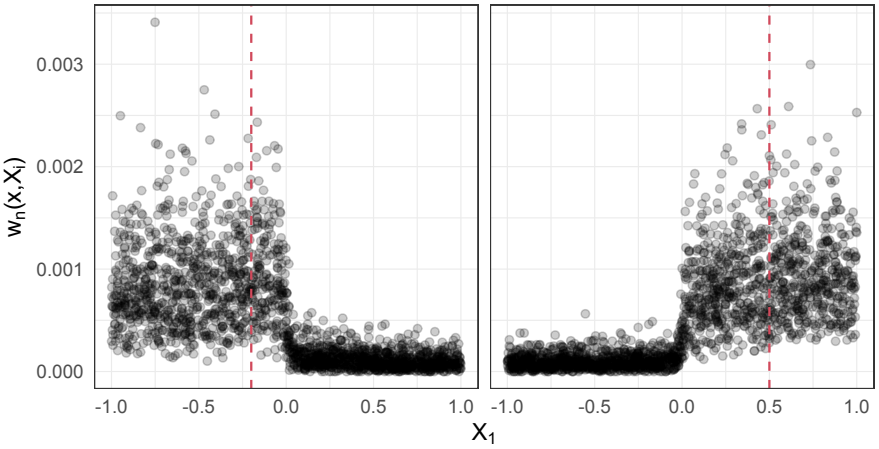


Figure B.1: The height of the points represents the localizing weights $w_n(x, X_i)$ between a test predictor $x \in \mathbb{R}^p$ and each training observation $X_i \in \mathbb{R}^p$. The dashed line indicates the first coordinate of the test predictor values.

B.4 Additional Material for Simulation Study

B.4.1 Sensitivity of Intermediate Threshold Level

Figure B.2 shows the square root MISE of predicted quantiles as a function of the intermediate threshold τ_0 for different quantile levels τ and different shape parameters ξ of the noise variable. Even though the threshold choice has an influence on the prediction accuracy, from the scales of the square root MISE it can be seen that this influence is not too strong. The optimal choice will depend on the properties of the data such as the tail heaviness of the response; for details see de Haan and Ferreira (2006, Section 3.2). In applications, there are numerous data-driven methods for choosing the threshold such as the mean excess plot (see Embrechts et al., 2012, Section 6.2.2).

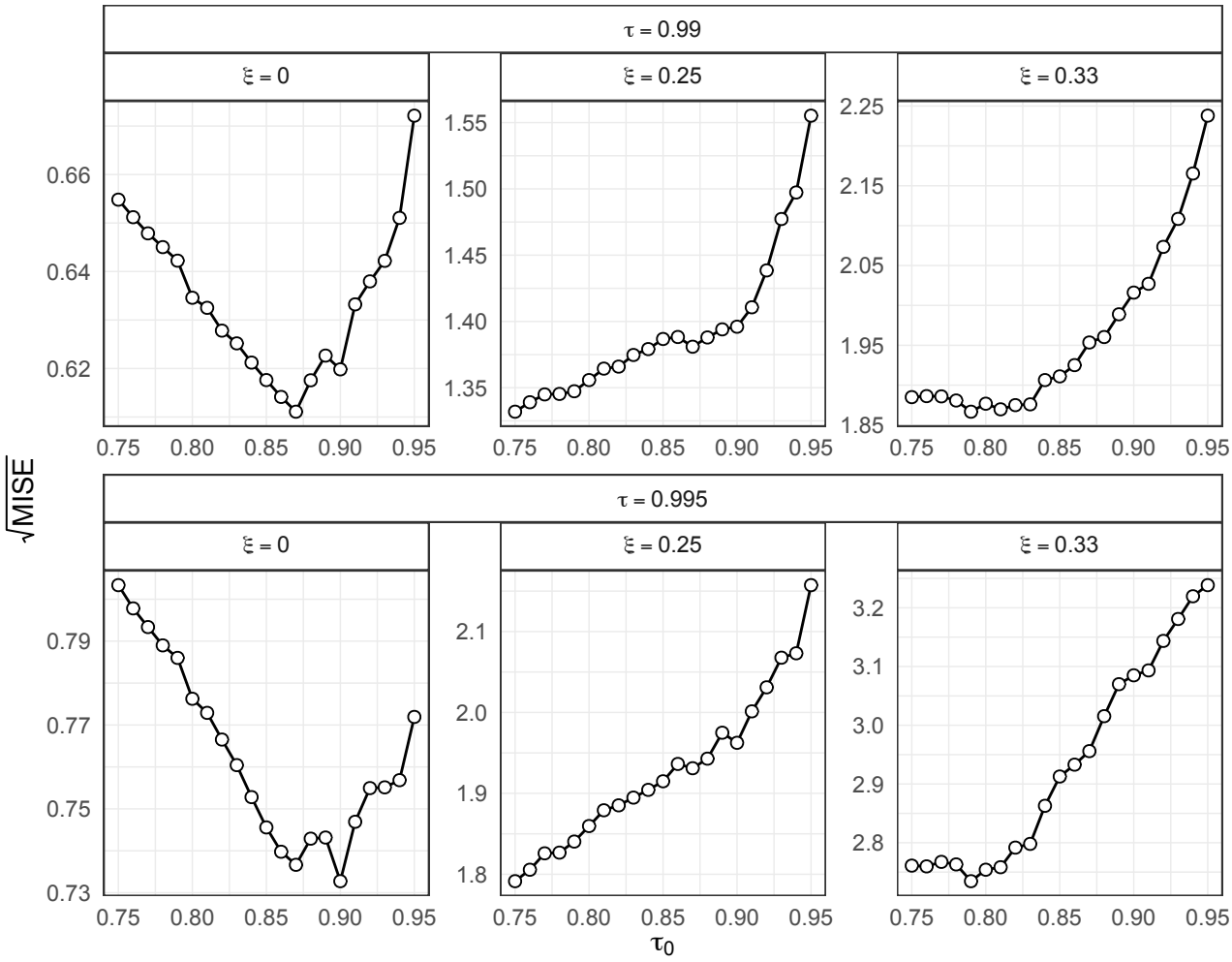


Figure B.2: Square root MISE of predicted quantiles as a function of the intermediate threshold τ_0 for different quantile levels τ and different shape parameters ξ of the noise variable. Each point is an average over $m = 100$ repetitions. The data is generated according to Example 2.1 in the main text, where we set the dimension of the predictor space to $p = 5$.

B.4.2 Experiment 3

In the last experiment mentioned in Section 2.4, we consider more complex regression functions depending on more signal variables both in the scale and shape parameters. While the predictor variables X are uniform distributed on $[-1, 1]^p$ with $p = 10$, the conditional response follows three different models

$$(Y \mid X = x) \sim s_j(x) T_{\nu(x)}, \quad j = 1, 2, 3,$$

where we allow both degrees of freedom $\nu(x)$ and the scale $s_j(x)$ of the Student's t distribution to depend on the predictors. In particular, we model the degrees of freedom as a decreasing function of the first predictor as $\nu(x) = 3[2 + \tanh(-2x_1)]$, and the different scale functions as

$$\begin{aligned} s_1(x) &= [2 + \tanh(2x_1)](1 + x_2/2), \\ s_2(x) &= 4 - (x_1^2 + 2x_2^2), \\ s_3(x) &= 1 + 2\pi\varphi(2x_1, 2x_2), \end{aligned}$$

where φ denotes a centered bivariate Gaussian density with unit variance and correlation coefficient equal to 0.75. The first scale function $s_1(x)$ is non-linear with respect to the first predictor and contains an interaction effect between the first two predictors. The function $s_2(x)$ is quadratic and decreasing in the first two dimensions. The third scale function $s_3(x)$ is non-linear in the first two predictors and contains an interaction effect. The sample size is $n = 5000$.

In this experiment we compare ERF, GRF, GBEX, EGP Tail and the unconditional method. We leave out EGAM because we observed it performs poorly in the scenarios considered here. Figure B.3 shows the boxplots of $\sqrt{\text{ISE}}$ over $m = 50$ simulations over different models, methods, and quantile levels. For better visualization, we remove large outliers of GRF, QRF, and EGP Tail. We observe that ERF and GBEX generally outperform the other methods over all models and quantile levels, where GBEX has a slight advantage in high quantiles for Models 2 and 3. GRF and QRF seems to deteriorate completely for very large quantiles.

B.5 Additional Material for U.S. Wage Analysis

B.5.1 Additional Figure

Figure B.4 shows that estimated GPD parameters $\hat{\theta}(x)$ for the original response as a function of age for groups with less or more than 15 years of education.

B.5.2 Analysis with Log-Transformed Response

Following Angrist et al. (2009), we consider here the natural logarithm of the wage as response variable for quantile regression. We perform the same analysis as in Section 2.5 again with this log-transformed response since it highlights several interesting properties of the ERF algorithm. Figure B.5 shows the GPD parameters $\hat{\theta}^{\log}(x)$ estimated by ERF as a function of years of education when the response is $\log(Y)$. We notice that the log-transformation makes the response lighter-tailed, with estimated shape parameters $\hat{\xi}^{\log}(x)$ fairly close to 0. The scale parameters $\hat{\sigma}^{\log}(x)$ still show a certain structure, but they vary on a much smaller scale compared to $\hat{\sigma}(x)$ estimated on the original response; see Figure 2.6 in the main text. These observations are consistent with theory since it is well-known that the log-transformation renders heavy-tailed data into light-tailed (Embrechts et al., 2012, Example 3.3.33). Moreover, the shape parameter on the original data then essentially acts as a scale parameter in the GPD approximation of the log-transformed data, explaining the smaller variation of $\hat{\sigma}^{\log}(x)$.

Figure B.6 in the main text shows the (exponentiated) predicted quantiles $\exp\{\hat{Q}_x^{\log}(\tau)\}$ of the different methods as a function of years of education when the response is $\log(Y)$; we removed again all quantiles above 6,000 predicted by GRF. By construction, GRF is invariant to the log-transformation, while the methods based on extrapolation may produce predictions that differ from $\hat{Q}_x(\tau)$ in Figure 2.7 fitted on the original data. The reason is that the approximation by the GPD is done on heavy-tailed data on the original scale and on much lighter-tailed data on the log-scale. We observe in Figure B.6 that the flexible methods ERF and GBEX have the desirable property that the predictions do not change much under marginal transformations. The unconditional method on the other hand seems to be sensitive to marginal transformation and works better on the log-transformed data as it captures a larger variability of the conditional quantiles even for

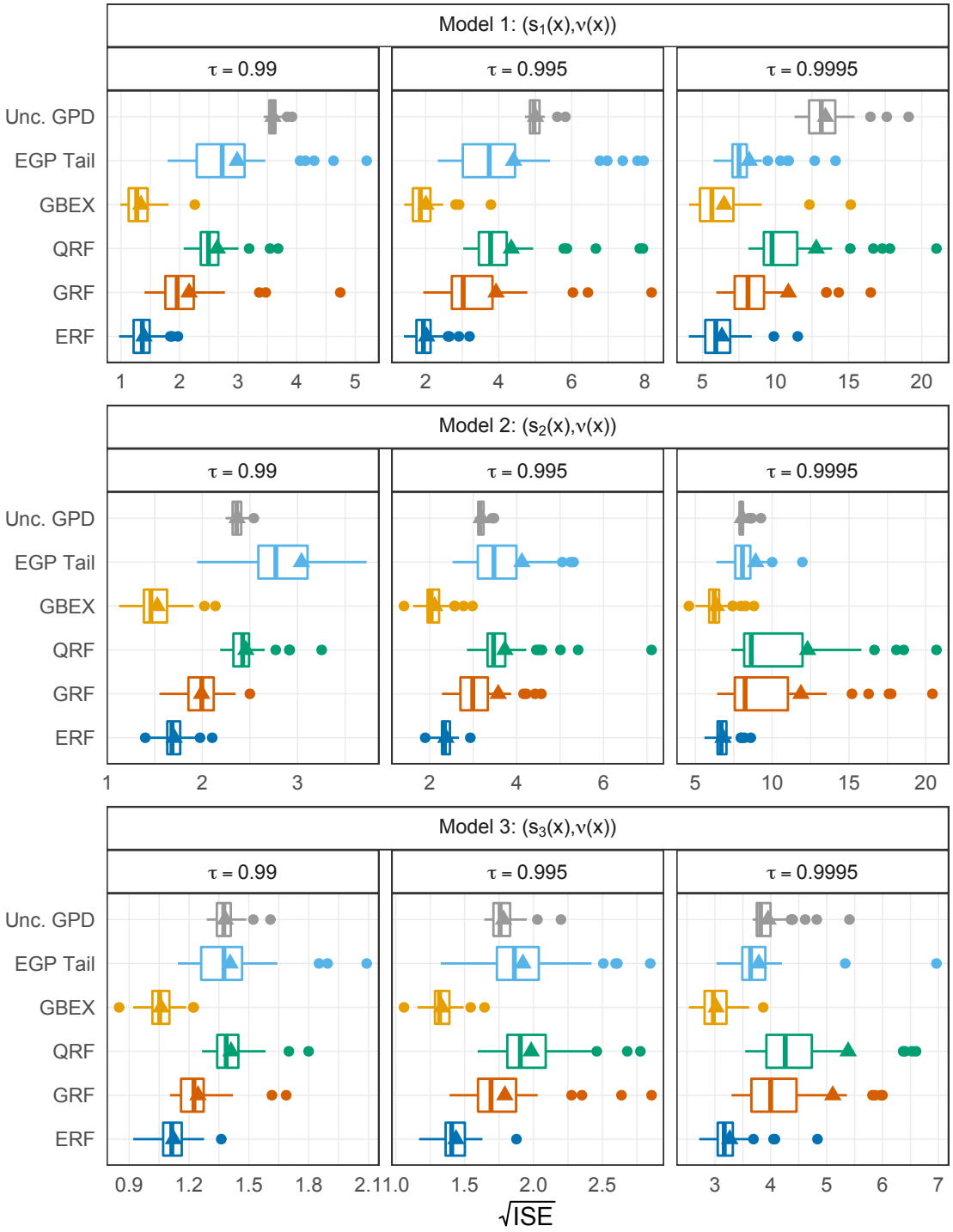


Figure B.3: Boxplots of $\sqrt{\text{ISE}}$ over $m = 50$ simulations for different generative models (rows) and quantile levels (columns). The predictor space dimension is set to $p = 10$. Triangles represent the average values.

high τ. This is confirmed by Figure B.7 where we observe that the unconditional method has a smaller loss especially for higher quantiles, while all other methods have a similar performance as on the original data. To better understand this behavior, we recall the

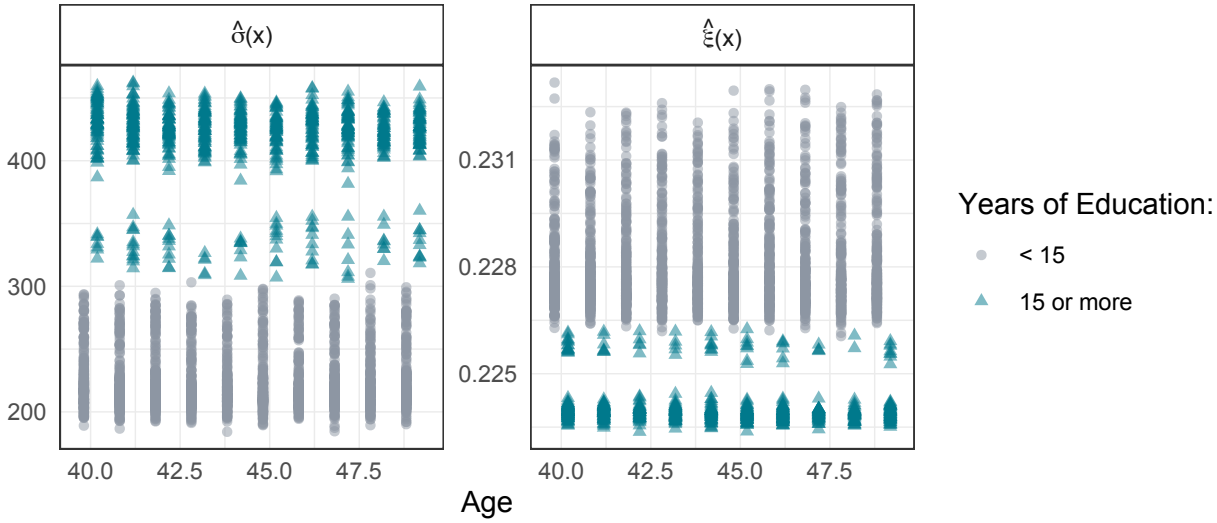


Figure B.4: Estimated GPD parameters $\hat{\theta}(x)$ as a function of age for groups with less (circles) or more (triangles) than 15 years of education.

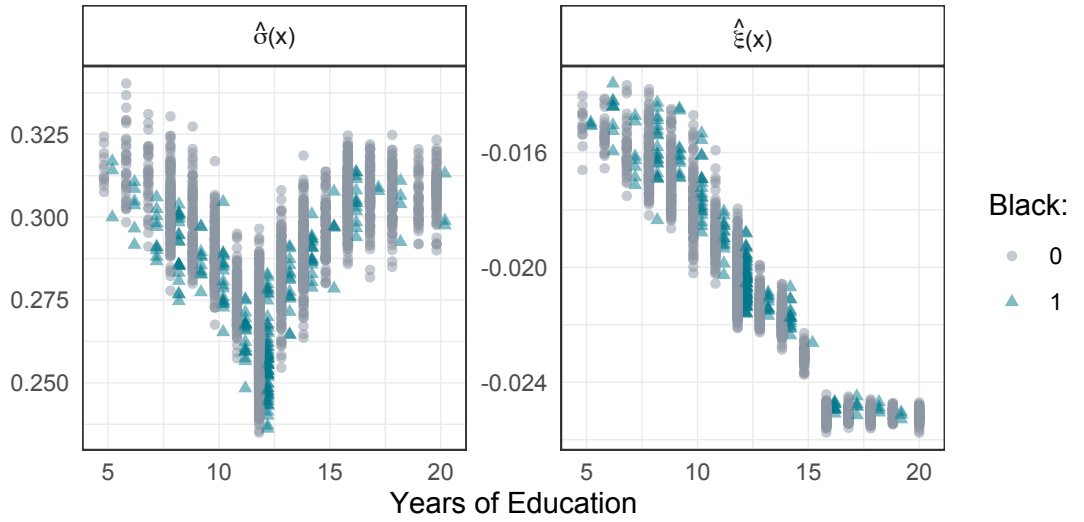


Figure B.5: Estimated GPD parameters $\hat{\theta}(x)$ for the log-response as a function of the years of education for the black (triangles) and white (circles) subgroups.

GPD approximation from (2.1.1) for large quantiles estimated on the original response as

$$\hat{Q}_x(\tau) \approx \hat{Q}_x(\tau_0) + G^{-1}\left(\frac{\tau - \tau_0}{1 - \tau_0}; \hat{\theta}(x)\right), \tag{B.5.1}$$

where G^{-1} is the inverse of the distribution function (2.2.2) of the GPD; see Figure 2.7 in the main text. On the other hand, first estimating the quantiles of the log-transformed data with a similar approximation and then exponentiating these estimates results in

$$\exp\{\hat{Q}_x^{\log}(\tau)\} \approx \hat{Q}_x(\tau_0) \exp\left\{G^{-1}\left(\frac{\tau - \tau_0}{1 - \tau_0}; \hat{\theta}^{\log}(x)\right)\right\}, \tag{B.5.2}$$

where $\hat{\theta}^{\log}(x)$ is the parameter vector of the GPD fitted for the response $\log(Y)$; see Figure B.6. We note that $\hat{Q}_x(\tau_0)$ is the same in both approximations since it is fitted using

quantile GRF, which is invariant under marginal transformations. Comparing (B.5.1) and (B.5.2) shows that the intermediate quantiles have an additive and multiplicative influence on the extreme quantiles, respectively. This explains why using the unconditional method for the GPD with $\hat{\theta}^{\log}(x) \equiv \hat{\theta}^{\log}$ seems to work better on the log-transformed data. Indeed, the different multiplicative scalings observed for ERF and GBEX in Figure 2.7 in the main text cannot be represented by (B.5.1) with unconditional GPD, but they can be represented by (B.5.2) if the intermediate quantile already carries the structure.

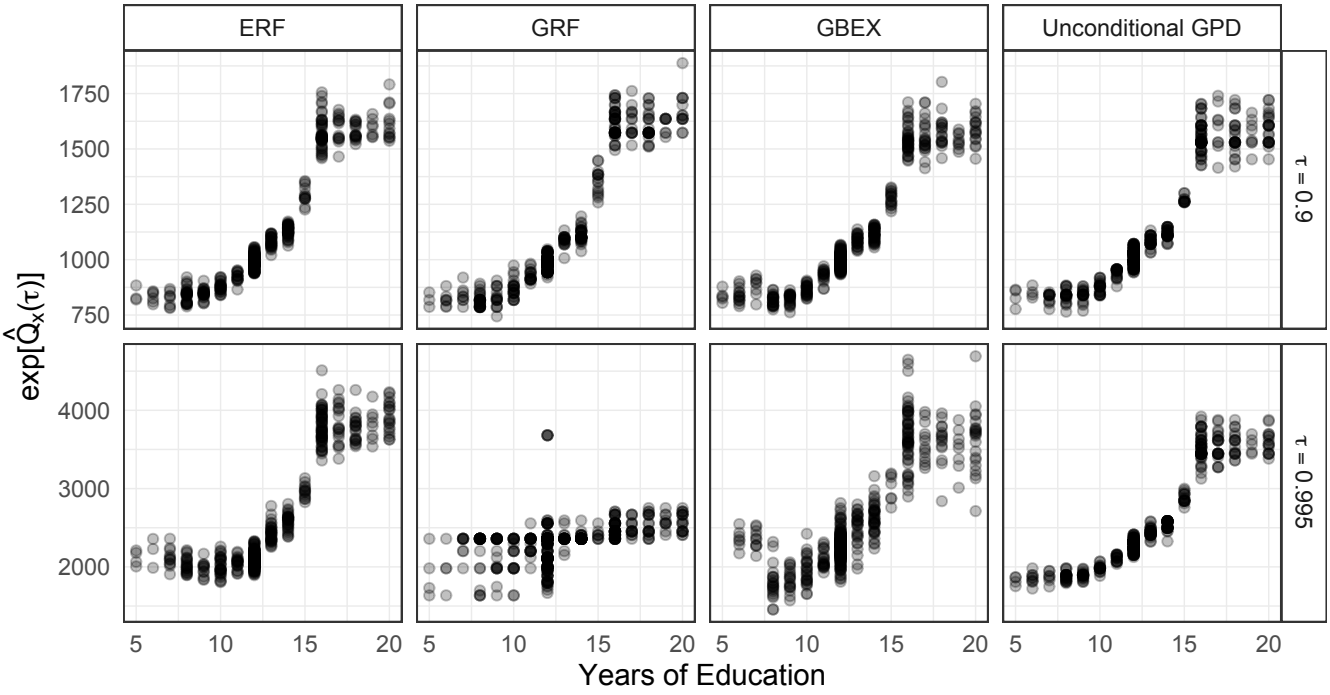


Figure B.6: Predicted quantiles at levels $\tau = 0.9, 0.995$ for ERF, GRF, GBEX, and the unconditional method fitted on the log-response.

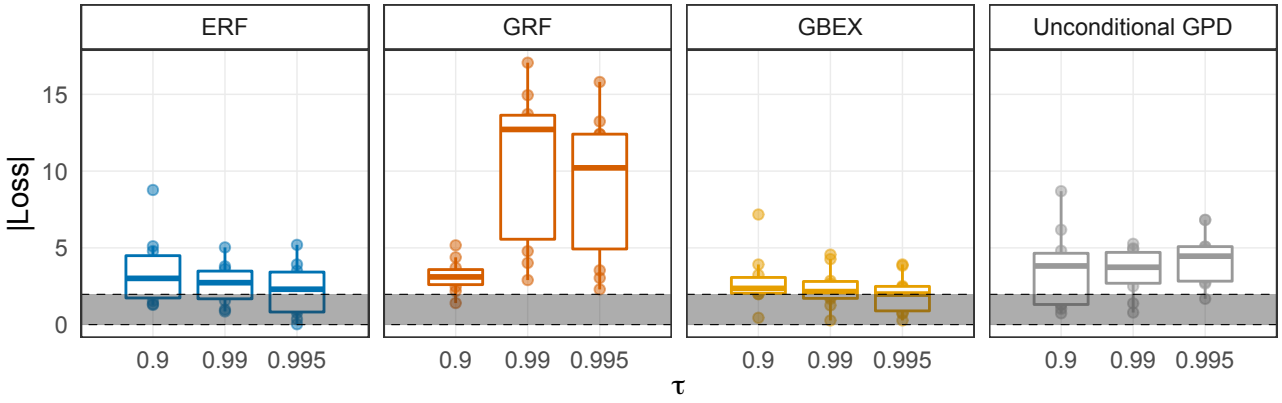


Figure B.7: Absolute value of the loss (2.5.1) for the different methods fitted on the log-response of the U.S. wage data. The shaded area represents the 95% interval of the absolute value of a standard normal distribution.

Appendix C

Distribution generalization in semi-parametric models: A control function approach

C.1 Proofs

C.1.1 Proof of Proposition 3.4

Proof.

We first show that $\mathcal{J} \subseteq \mathcal{I}$. Let $f \in \mathcal{J}$. Let $\delta \in \ker(M_0^T)$ such that $f(x) = f_0(x) + \delta^T x$, for all $x \in \mathbb{R}^p$. Notice that,

$$Y - f(X) = U - \delta^T X = U - \delta^T(M_0 E + V) = U - \delta^T V,$$

because $\delta \in \ker(M_0^T)$. Since $(U, V) \perp\!\!\!\perp E$, it follows that $Y - f(X) \perp\!\!\!\perp E$. Thus, $f \in \mathcal{I}$.

We now show that there exists a function $f \in \mathcal{I}$ such that $f \notin \mathcal{J}$. Fix some measurable non-linear function $g : \mathbb{R} \to \mathbb{R}$ such that $f(x) = f_0(x) + g(\delta^T x)$, for all $x \in \mathbb{R}^p$. Since $\delta \in \ker(M_0^T)$, it holds that $g(\delta^T X) = g(\delta^T V)$. Notice that

$$Y - f(X) = U - g(\delta^T V) \perp\!\!\!\perp E.$$

Thus, $f \in \mathcal{I}$. Since g is non-linear, it holds that $f \notin \mathcal{J}$. □

C.1.2 Proof of Proposition 3.5

Proof. First, the error term $V = X - \mathbb{E}[X \mid E]$ is identified from the observational distribution $P_{\mathcal{C}_0}$. For almost every (a.e.) $x, v \in \mathbb{R}^p$, it holds that

$$\begin{aligned} \mathbb{E}[Y \mid X = x, V = v] &= \mathbb{E}[f_0(X) + U \mid X = x, V = v] = f_0(x) + \mathbb{E}[U \mid X = x, V = v] \\ &= f_0(x) + \mathbb{E}[U \mid V = v] = f_0(x) + \gamma_0^T v. \end{aligned} \tag{C.1.1}$$

We now show the two implications.

($\Rightarrow$) Let $f \in \mathcal{J}$. Let $\delta \in \ker(M_0^T)$ such that $f(x) = f_0(x) + \delta^T x$, for all $x \in \mathbb{R}^p$. For a.e. $x, v \in \mathbb{R}^p$, there exists $e \in \mathbb{R}^r$ such that $x = M_0 e + v$. Thus, $\delta^T x = \delta^T(M_0 e + v) = \delta^T v$. Fix $\gamma = \gamma_0 - \delta \in \mathbb{R}^p$. Then, for a.e. $x, v \in \mathbb{R}^p$,

$$f(x) + \gamma^T v \ = f_0(x) + \delta^T x - \delta^T v + \gamma_0^T v = f_0(x) + \gamma_0^T v,$$

which satisfies (C.1.1).

($\Leftarrow$) Let $f : \mathbb{R}^p \to \mathbb{R}$ and $\gamma \in \mathbb{R}^p$ such that $\mathbb{E}[Y \mid X = x, V = v] = f(x) + \gamma^T v$, for a.e. $x, v \in \mathbb{R}^p$. Fix $h = f - f_0$ so that, for a.e. $x, v \in \mathbb{R}^p$,

$$\mathbb{E}[Y \mid X = x, V = v] = f_0(x) + h(x) + \gamma^T v.$$

From (C.1.1), it follows that

$$h(x) = (\gamma_0 - \gamma)^T v, \quad \text{for a.e. } x, v \in \mathbb{R}^p. \tag{C.1.2}$$

Let $e \in \mathbb{R}^r$, $v \in \mathbb{R}^p$, and fix $x = M_0 e + v \in \mathbb{R}^p$. From (C.1.2), it follows that

$$h(M_0 e + v) = (\gamma_0 - \gamma)^T v. \tag{C.1.3}$$

Since (C.1.3) holds for all $e \in \mathbb{R}^r$, $v \in \mathbb{R}^p$, we have in particular

$$h(v) = (\gamma_0 - \gamma)^T v, \quad \text{for all } v \in \mathbb{R}^p. \tag{C.1.4}$$

Since $x = M_0 e + v \in \mathbb{R}^p$, (C.1.4) implies

$$h(M_0 e + v) = (\gamma_0 - \gamma)^T (M_0 e + v). \tag{C.1.5}$$

Equating (C.1.3) and (C.1.5), we conclude that $(\gamma_0 - \gamma)^T M_0 e = 0$ for all $e \in \mathbb{R}^r$, and hence $(\gamma_0 - \gamma) \in \ker(M_0^T)$. Therefore, $f(\cdot) = f_0(\cdot) + (\gamma_0 - \gamma)^T \cdot \in \mathcal{J}$. □

C.1.3 Proof of Proposition 3.6

Proof. Notice that $f_0 \in \mathcal{J}$. Moreover, for any $\delta \in \ker(M_0^T)$ we have that

$$Y - f_0(X) - \delta^T X = U - \delta^T (M_0 E + V) = U - \delta^T V. \tag{C.1.6}$$

Thus,

$$\begin{aligned} &\min_{\delta \in \ker(M_0^T)} \mathbb{E}[(Y - f_0(X) - \delta^T X)^2] = \min_{\delta \in \ker(M_0^T)} \mathbb{E}[(U - \delta^T V)^2] \\ &= \min_{\delta \in \ker(M_0^T)} \mathbb{E}[(\gamma_0^T V + \epsilon_U - \delta^T V)^2] = \min_{\delta \in \ker(M_0^T)} \mathbb{E}[(\gamma_0^T V - \delta^T V)^2] + \mathbb{E}[\epsilon_U^2], \end{aligned} \tag{C.1.7}$$

where the cross product term in the last equality vanishes since $\epsilon_U \perp\!\!\!\perp V$. Notice that any vector $\delta \in \ker(M_0^T)$ can be written as $\delta = R\alpha$ for some $\alpha \in \mathbb{R}^{p-r}$. Therefore,(C.1.7) writes

$$\begin{aligned} \min_{a \in \mathbb{R}^{p-r}} \mathbb{E}[(V^T \gamma_0 - V^T R\alpha)^2] &= \min_{a \in \mathbb{R}^{p-r}} (\gamma_0 - R\alpha)^T S (\gamma_0 - R\alpha) \\ &= \min_{a \in \mathbb{R}^{p-r}} \gamma_0^T S \gamma_0 - 2\alpha^T R^T S \gamma_0 + \alpha^T R^T S R \alpha. \end{aligned} \tag{C.1.8}$$

Differentiating with respect to α and using the fact that $R^T S R$ is invertible yields

$$\delta_0 := R\alpha_0 = R(R^T S R)^{-1} R^T S \gamma_0.$$

Therefore, the optimal function is $f^*(x) = f_0(x) + \delta_0^T x \in \mathcal{J}$, for all $x \in \mathbb{R}^p$.

We now prove that the optimal function f^* is well-defined, i.e., it does not depend on the choice of the representative function $\tilde{f} \in \mathcal{J}$. Let $\tilde{f} \in \mathcal{J}$. From the definition of $\mathcal{J}$ in (3.3.4), there exists some fixed $\overline{\delta} \in \ker(M_0^T)$ such that for all $x \in \mathbb{R}^p$

$$\tilde{f}(x) = f_0(x) + \overline{\delta}^T x, \tag{C.1.9}$$

so, it holds $\mathbb{P}_{\mathcal{C}_0}$-almost surely that

$$Y = f_0(X) + U = \tilde{f}(X) - \overline{\delta}^T X + U = \tilde{f}(X) - \overline{\delta}^T V + U.$$

Using the same argument as in (C.1.7) and (C.1.8), with $\tilde{f}$ and $(\gamma_0 - \overline{\delta})$ in place of f_0 and γ_0, respectively, the optimal element in the kernel writes

$$\tilde{\delta} := R(R^T S R)^{-1} R^T S(\gamma_0 - \overline{\delta}). \tag{C.1.10}$$

Write $\overline{\delta} = R\alpha$ for some $\alpha \in \mathbb{R}^{p-r}$. Therefore, using (C.1.9) and (C.1.10), the optimal function writes

$$\begin{aligned} f^*(x) = \tilde{f}(x) + \tilde{\delta}^T x &= f_0(x) + \overline{\delta}^T x + (\gamma_0 - \overline{\delta})^T S^T R(R^T S R)^{-1} R^T x \\ &= f_0(x) + \alpha^T R^T x + (\gamma_0 - R\alpha)^T S^T R(R^T S R)^{-1} R^T x \\ &= f_0(x) + \alpha^T R^T x + \gamma_0^T S^T R(R^T S R)^{-1} R^T x - \alpha^T R^T S^T R(R^T S R)^{-1} R^T x \\ &= f_0(x) + \alpha^T R^T x + \delta_0^T x - \alpha^T R^T x = f_0(x) + \delta_0^T x, \text{ for all } x \in \mathbb{R}^p. \end{aligned}$$

□

C.1.4 Proof of Lemma 3.7

Proof. For ease of notation, define $M := M_0$. For $\varepsilon > 0$, let $h^* \in \mathcal{F}$ such that

$$\sup_{e \in \mathbb{R}^r} F(h^*, e) < \inf_{h \in \mathcal{F}} \sup_{e \in \mathbb{R}^r} F(h, e) + \varepsilon.$$

For any $e \in \mathbb{R}^r$, it holds that

$$\begin{aligned} F(h^*, e) - \mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] &\geq 2\ (\gamma_0 - \delta_0)^T \mathbb{E}[VV^T\delta_0 - Vh^*(Me + V)] \\ &= 2\ (\gamma_0 - \delta_0)^T \left\{S\delta_0 - \mathbb{E}[Vh^*(Me + V)]\right\} \\ &= 2(\gamma_0 - \delta_0)^T S\delta_0 - 2(\gamma_0 - \delta_0)^T \mathbb{E}[Vh^*(Me + V)]. \end{aligned} \tag{C.1.11}$$

From Proposition 3.6, the optimal element in the kernel can be written as $\delta_0 = P\gamma_0 := R(R^T S R)^{-1} R^T S\gamma_0 \in \mathbb{R}^p$, where $R \in \mathbb{R}^{p \times (p-r)}$ denotes an orthonormal basis for $\ker(M^T)$. Using the fact that $SP = P^T SP$, we obtain

$$(\gamma_0 - \delta_0)^T S\delta_0 = \gamma_0^T (I - P^T) SP\gamma_0 = \gamma_0^T (SP - P^T SP)\gamma_0 = 0. \tag{C.1.12}$$

Thus, from (C.1.11) and (C.1.12), it follows that

$$\sup_{e \in \mathbb{R}^r} F(h^*, e) - \mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] \geq -2(\gamma_0 - \delta_0)^T \mathbb{E}[Vh^*(Me + V)],\ e \in \mathbb{R}^r. \tag{C.1.13}$$

Suppose h^* is bounded. By Lemma C.4, for any $\eta > 0$ there exists $\tilde{e} \in \mathbb{R}^r$ such that

$$-2(\gamma_0 - \delta_0)^T \mathbb{E}[Vh^*(M\tilde{e} + V)] > -\eta. \tag{C.1.14}$$

Hence, from (C.1.13) and (C.1.14), it follows that

$$\sup_{e \in \mathbb{R}^r} F(h^*, e) > \mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] - \eta.$$

Since $\eta > 0$ was arbitrary, it follows that

$$\mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] \leq \sup_{e\in\mathbb{R}^r} F(h^*, e) < \inf_{h\in\mathcal{F}} \sup_{e\in\mathbb{R}^r} F(h, e) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this completes the proof of Lemma 3.7 for bounded $h^* \in \mathcal{F}$.

Suppose $h^* \in \mathcal{F}$ is unbounded. Define the function $F_A : \mathcal{F} \times \mathbb{R}^r \to [0, \infty)$ for all $h \in \mathcal{F}$, $e \in \mathbb{R}^r$ and $A \subseteq \mathbb{R}^p$ by

$$F_A(h, e) \coloneqq \int_A (v^T\gamma_0 - h(M_0e + v))^2\phi_S(v)\, dv,$$

where ϕ_S denotes the density of V. Fix $K > 0$ and consider the compact $I_K = T([-K, K]^p) \subseteq \mathbb{R}^p$, where $T : \mathbb{R}^p \to \mathbb{R}^p$ is the invertible map $\epsilon \mapsto T(\epsilon) = S^{1/2}\epsilon$. Define $m_K \coloneqq \max\left\{|v^T\gamma_0| : v \in I_K\right\}$, and consider the bounded function $h_K(v) = \min\{|h^*(v)|, m_K\}\operatorname{sign}(h^*(v))$. Then, it holds that

$$\sup_{e\in\mathbb{R}} F(h^*, e) \geq \sup_{e\in\mathbb{R}} F_{I_K}(h^*, e) \geq \sup_{e\in\mathbb{R}} F_{I_K}(h_K, e). \tag{C.1.15}$$

Since $V \sim N(0, S)$, it holds that $V \overset{d}{=} S^{1/2}\epsilon$, where $\epsilon \sim N(0, I_p)$. Thus,

$$\begin{aligned} F_{I_K^c}(h_K, e) &= \int_{I_K^c} (v^T\gamma_0 - h_K(M_0e + v))^2\phi_S(v)\, dv \\ &= \int_{\mathbb{R}^p\setminus[-K,K]^p} (\epsilon^T S^{1/2}\gamma_0 - h_K(M_0e + S^{1/2}\epsilon))^2\phi_I(\epsilon)\, d\epsilon, \end{aligned}$$

where ϕ_I is the density of ϵ. Moreover, for any $\eta > 0$, there exists K_0 such that for all $K > K_0$,

$$0 \leq \sup_{e\in\mathbb{R}} F_{I_K^c}(h_K, e) \leq \int_{\mathbb{R}^p\setminus[-K,K]^p} \left(\epsilon^T S^{1/2}\gamma_0 + m_K \operatorname{sign}(\epsilon^T S^{1/2}\gamma_0)\right)^2 \phi_I(\epsilon)\, d\epsilon < \eta.$$

since all moments exist. Recall that h_K is bounded for all $K > K_0$, and thus

$$\mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] \leq \sup_{e\in\mathbb{R}} F(h_K, e). \tag{C.1.16}$$

At the same time,

$$\mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] \leq \sup_{e\in\mathbb{R}} F(h_K, e) = \sup_{e\in\mathbb{R}} \left\{F_{I_K}(h_K, e) + F_{I_K^c}(h_K, e)\right\} < \sup_{e\in\mathbb{R}} F_{I_K}(h_K, e) + \eta.$$

Since $\eta > 0$ was arbitrary, it follows that $\liminf_{K\to\infty} \sup_{e\in\mathbb{R}} F_{I_K}(h_K, e) = \mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2]$. Using (C.1.15) it follows that

$$\mathbb{E}[(V^T\gamma_0 - V^T\delta_0)^2] \leq \sup_{e\in\mathbb{R}} F(h^*, e) \leq \inf_{h\in\mathcal{F}} \sup_{e\in\mathbb{R}} F(h, e) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this completes the proof of Lemma 3.8 for unbounded $h^* \in \mathcal{F}$.

□

C.1.5 Proof of Theorem 3.8

Proof. We treat the two terms in (3.3.9) separately. Consider the term on the right-hand side. Let $f \in \mathcal{F}$ and $e \in \mathbb{R}^r$. Since the environment vector $E \in \mathbb{R}^r$ is exogenous, the do-intervention on E is the same as the conditional expectation

$$\mathbb{E}[(Y - f(X))^2 \mid \mathrm{do}(E := e)] = \mathbb{E}[(Y - f(X))^2 \mid E = e]. \tag{C.1.17}$$

Moreover, we can write $f = f_0 + h$, for some $h \in \mathcal{F}$. Substituting the definitions of Y and f we obtain

$$Y - f(X) = f_0(X) + U - f_0(X) - h(X) = U - h(X) = V^T\gamma_0 + \epsilon_U - h(X). \tag{C.1.18}$$

By combining (C.1.17) and (C.1.18), we obtain

$$\begin{aligned} &\mathbb{E}[(Y - f(X))^2 \mid \mathrm{do}(E := e)] = \mathbb{E}[(V^T\gamma_0 + \epsilon_U - h(X))^2 \mid E = e] \\ &= \mathbb{E}[(V^T\gamma_0 + \epsilon_U - h(M_0 e + V))^2] = \mathbb{E}[(V^T\gamma_0 - h(M_0 e + V))^2] + \mathbb{E}[\epsilon_U^2], \end{aligned} \tag{C.1.19}$$

where the cross product term in the last equation vanishes since $\epsilon_U \perp\!\!\!\perp V$.

Consider now the term on the left-hand side of 3.3.9. Let $\overline{f} \in \mathcal{J}$. Then, there exists some $\delta \in \ker(M_0^T)$ such that $\overline{f}(x) = f_0(x) + \delta^T x$ for all $x \in \mathbb{R}^p$. Moreover, recall that $\delta^T X = \delta^T V$. Substituting the definitions of Y and $\overline{f}$, we obtain

$$\begin{aligned} \mathbb{E}[(Y - \overline{f}(X))^2] &= \mathbb{E}[(f_0(X) + V^T\gamma_0 + \epsilon_U - f_0(X) - \delta^T X)^2] \\ &= \mathbb{E}[(V^T\gamma_0 + \epsilon_U - X^T\delta)^2] \\ &= \mathbb{E}[(V^T\gamma_0 - V^T\delta)^2] + \mathbb{E}[\epsilon_U^2], \end{aligned} \tag{C.1.20}$$

where the cross product term in the last equation vanishes since $\epsilon_U \perp\!\!\!\perp V$. Since $f \in \mathcal{F}$, $\overline{f} \in \mathcal{J}$ and $e \in \mathbb{R}^r$ were arbitrary, using (C.1.19) and (C.1.20), we get

$$\begin{aligned} \min_{f \in \mathcal{J}} \mathbb{E}[(Y - f(X))^2] &= \min_{\delta \in \ker(M_0^T)} \mathbb{E}[(V^T\gamma_0 - V^T\delta)^2] + \mathbb{E}[\epsilon_U^2] \\ &= \min_{h \in \mathcal{F}} \sup_{e \in \mathbb{R}^r} \mathbb{E}[(V^T\gamma_0 - h(M_0 e + V))^2] + \mathbb{E}[\epsilon_U^2] \\ &= \min_{f \in \mathcal{F}} \sup_{e \in \mathbb{R}^r} \mathbb{E}[(Y - f(X))^2 \mid \mathrm{do}(E := e)], \end{aligned} \tag{C.1.21}$$

where the second equality follows from Lemma 3.7. □

C.1.6 Proof of Corollary 3.9

Proof. Since $J \subsetneq I$ from Proposition 3.4, we have that

$$\min_{f \in \mathcal{J}} \mathbb{E}\left[(Y - f(X))^2\right] \geq \min_{f \in \mathcal{I}} \mathbb{E}\left[(Y - f(X))^2\right]. \tag{C.1.22}$$

For any $f \in \mathcal{I}$, since $Y - f(X) \perp\!\!\!\perp E$, it holds that

$$\mathbb{E}\left[(Y - f(X))^2 \mid E = e\right] = \mathbb{E}\left[(Y - f(X))^2\right], \text{ for all } e \in \mathbb{R}^r. \tag{C.1.23}$$

Therefore, using Theorem 3.8 and the fact that $\mathcal{I} \subseteq \mathcal{F}$, we obtain

$$\begin{aligned}\min_{f\in\mathcal{J}} \mathbb{E}\left[(Y-f(X))^2\right] &= \min_{f\in\mathcal{F}} \sup_{e\in\mathbb{R}^r} \mathbb{E}[(Y-f(X))^2 \mid \mathrm{do}(E:=e)] \\ &\leq \min_{f\in\mathcal{I}} \sup_{e\in\mathbb{R}^r} \mathbb{E}[(Y-f(X))^2 \mid \mathrm{do}(E:=e)] \\ &= \min_{f\in\mathcal{I}} \sup_{e\in\mathbb{R}^r} \mathbb{E}[(Y-f(X))^2 \mid E=e] \\ &= \min_{f\in\mathcal{I}} \mathbb{E}\left[(Y-f(X))^2\right].\end{aligned} \tag{C.1.24}$$

□

C.2 Further lemmas

We first define the main objects needed in the following lemmas.

Definition C.1. *Let*

1. $M \in \mathbb{R}^{p\times r}$ *be a matrix such that* $\mathrm{rank}(M) = r$,
2. $S \in \mathbb{R}^{p\times p}$ *be a positive definite matrix,*
3. $R \in \mathbb{R}^{p\times(p-r)}$ *denote an orthonormal basis for* $\ker(M^T)$,
4. $Q \in \mathbb{R}^{p\times r}$ *denote an orthonormal basis for* $\mathrm{span}(M)$,
5. $B_k = S + kMM^T$, *for any* $k > 0$.

We will also use the following facts.

Remark C.1. The matrix MM^T admits the spectral decomposition

$$MM^T = R\Lambda_0 R^T + Q\Lambda_1 Q^T,$$

where $0 = \lambda_1 = \cdots = \lambda_{p-r}$ are the eigenvalues in Λ_0, and $0 < \lambda_{p-r+1} \leq \cdots \leq \lambda_p$ are the eigenvalues in Λ_1. ◁

Remark C.2. For any $k > 0$, the matrix kMM^T has the same eigenvectors as MM^T and eigenvalues $0 = k\lambda_1 = \cdots = k\lambda_{p-r} < k\lambda_{p-r+1} \leq \cdots \leq k\lambda_p$. So, we can write

$$kMM^T = R\Lambda_0 R^T + Qk\Lambda_1 Q^T.$$

◁

Remark C.3. Let $k > 0$ and consider $B_k = S + kMM^T$. Denote by $\tilde{\lambda}_j$ its eigenvalues, $j = 1, \ldots, p$. By Weyl's inequality (Weyl, 1912), we have that

$$\lambda_1^S + k\lambda_j \leq \tilde{\lambda}_j \leq \lambda_p^S + k\lambda_j, \quad \text{for } j = 1, \ldots, p,$$

where $0 < \lambda_j^S$ are the eigenvalues of S positive definite. Thus, we have that

$$\begin{aligned}&0 < \lambda_1^S = \lambda_1^S + \lambda_j \leq \tilde{\lambda}_j \leq \lambda_p^S + \lambda_j = \lambda_p^S, \quad j = 1, \ldots, p-r, \\ &0 < k\lambda_j < k\lambda_j + \lambda_1^S \leq \tilde{\lambda}_j \leq \lambda_p^S + k\lambda_j, \quad j = p-r+1, \ldots, p,\end{aligned} \tag{C.2.1}$$

Therefore, we can write

$$B_k = R_k \tilde{\Lambda}_{0k} R^T + Q_k \tilde{\Lambda}_{1k} Q_k^T, \tag{C.2.2}$$

where $R_k \in \mathbb{R}^{p\times(p-r)}$ and $Q_k \in \mathbb{R}^{p\times r}$ are orthonormal matrices. ◁

Lemma C.2. *It holds that*

$$\|B_k^{-1/2} - RR^T B_k^{-1/2}\|_F \to 0, \text{ as } k \to \infty.$$

Proof. Using the spectral decomposition of B_k given in (C.2.2) and using results from Lemma C.6, we have that

$$\begin{aligned}\|(I - RR^T)B_k^{-1/2}\|_F &\leq \|(I - RR^T)R_k\tilde{\Lambda}_{0k}^{-1/2}R^T\|_F + \|(I - RR^T)Q_k\tilde{\Lambda}_{1k}^{-1/2}Q_k^T\|_F \\ &= \|(I - RR^T)R_k\tilde{\Lambda}_{0k}^{-1/2}\|_F + \|(I - RR^T)Q_k\tilde{\Lambda}_{1k}^{-1/2}\|_F \\ &\leq \|(I - RR^T)R_k\|_F \ \|\tilde{\Lambda}_{0k}^{-1/2}\|_F + \|(I - RR^T)\|_F \ \|Q_k\|_F \ \|\tilde{\Lambda}_{1k}^{-1/2}\|_F.\end{aligned} \tag{C.2.3}$$

We treat the terms separately. First, notice that

$$\|(I - RR^T)\|_F = \|Q_k\|_F = \sqrt{r}. \tag{C.2.4}$$

Also, from (C.2.1) in Remark C.3, it holds that

$$\begin{aligned}\|\tilde{\Lambda}_{0k}^{-1/2}\|_F &\leq \sqrt{\frac{p-r}{\lambda_1^S}}, \\ \|\tilde{\Lambda}_{1k}^{-1/2}\|_F &\leq \sqrt{\frac{r}{k\lambda_{p-r+1}}} \to 0, \text{ as } k \to \infty.\end{aligned} \tag{C.2.5}$$

Furthermore, using the fact that $(I - R_kR_k^T) = Q_kQ_k^T$, it holds that

$$\begin{aligned}\|(I - RR^T)R_k\|_F^2 &= \operatorname{tr}(R_k^T(I - RR^T)R_k) \\ &= \operatorname{tr}(R_k^TR_k - R_k^TRR^TR_k) \\ &= \operatorname{tr}(R^TR - R^TR_kR_k^TR) \\ &= \operatorname{tr}(R^T(I - R_kR_k^T)R) \\ &= \operatorname{tr}(R^TQ_kQ_k^TR) \\ &= \|Q_k^TR\|_F^2 \to 0,\end{aligned} \tag{C.2.6}$$

as $k \to \infty$, using Lemma C.3.

Putting together (C.2.3), (C.2.4), (C.2.5), and (C.2.6), we conclude that

$$\|B_k^{-1/2} - RR^T B_k^{-1/2}\|_F \to 0, \quad \text{as } k \to \infty.$$

□

The following lemma is an adaptation of the Davis–Kahan theorem (Davis and Kahan, 1970).

Lemma C.3. *It holds that*

$$\|Q_k^TR\|_F \to 0, \quad \text{as } k \to \infty.$$

Proof. Let $k > 0$. From Remark C.2, we can write

$$kMM^TR = R\Lambda_0 + Qk\Lambda_1Q^TR = R\Lambda_0 = 0 \in \mathbb{R}^{p\times(p-r)},$$

since $0 = \lambda_1 = \cdots = \lambda_{p-r}$. Therefore, $SR = kMM^TR + SR = (kMM^T + S)R = B_kR$. At the same time, from (C.2.2) in Remark C.3, it holds that $Q_k^T B_k = \tilde{\Lambda}_{1k} Q_k^T$. Thus, we have that $Q_k^T SR = Q_k^T B_k R = \tilde{\Lambda}_{1k} Q_k^T R$. It follows that

$$\begin{aligned} \|Q_k^T R\|_F &= \|\tilde{\Lambda}_{1k}^{-1}\tilde{\Lambda}_{1k} Q_k^T R\|_F \leq \|\tilde{\Lambda}_{1k}^{-1}\|_F \, \|\tilde{\Lambda}_{1k} Q_k^T R\|_F = \|\tilde{\Lambda}_{1k}^{-1}\|_F \, \|Q_k^T SR\|_F \\ &\leq \|\tilde{\Lambda}_{1k}^{-1}\|_F \, \|Q_k\|_F \, \|SR\|_F. \end{aligned} \tag{C.2.7}$$

We treat each term separately. First, from (C.2.1) in Remark C.3, it holds that

$$\|\tilde{\Lambda}_{1k}^{-1}\|_F \leq \frac{r}{k\lambda_{p-r+1}} \to 0, \text{ as } k \to \infty. \tag{C.2.8}$$

Also, it holds that

$$\|Q_k\|_F = \sqrt{r}, \ \|SR\|_F \leq \sqrt{p(p-r)\lambda_p^S}. \tag{C.2.9}$$

Putting (C.2.7), (C.2.8), and (C.2.9), the claim follows. □

Lemma C.4. *Let $V \sim N(0, S)$ where $S \in \mathbb{R}^{p\times p}$ is positive definite. Let $h \in \mathcal{F}$ be a bounded function. Then, for any $\eta > 0$ there exists $\tilde{e} \in \mathbb{R}^r$ such that*

$$(\gamma_0 - \delta_0)^T \mathbb{E}[Vh(M\tilde{e} + V)] < \eta.$$

Proof. Let $h \in \mathcal{F}$ be a bounded function and let $k > 0$. Consider

$$\mathbb{E}[Vh(ME_k + V)] = \int_{\mathbb{R}^r} \mathbb{E}[Vh(Me + V)]\phi_{kI}(e)\, de,$$

where $E_k \sim N(0, kI)$ with $E_k \perp\!\!\!\perp V$, and $\phi_{kI} : \mathbb{R}^r \to [0, \infty)$ denotes the multivariate normal density with mean zero and covariance $kI \in \mathbb{R}^{r\times r}$. Let $W_k := ME_k + V \sim N(0, B_k)$, where $B_k := S + kMM^T$ is positive definite. From the properties of Gaussian distribution, the conditional expectation of E_k given W_k is

$$\mathbb{E}[E_k \mid W_k] \ = \mathbb{E}[E_k W_k^T]\mathbb{E}[W_k W_k^T]^{-1} W_k = kM^T B_k^{-1} W_k. \tag{C.2.10}$$

Furthermore, notice that

$$I - kMM^T B_k^{-1} = (S + kMM^T - kMM^T)B_k^{-1} = SB_k^{-1}. \tag{C.2.11}$$

Therefore, from (C.2.10) and (C.2.11), it holds that

$$\begin{aligned} \mathbb{E}[Vh(ME_k + V)] &= \mathbb{E}[(W_k - ME_k)h(W_k)] = \mathbb{E}[(W_k - M\mathbb{E}[E_k \mid W_k])h(W_k)] \\ &= \mathbb{E}[\{W_k - kMM^T B_k^{-1} W_k\}h(W_k)] \\ &= \mathbb{E}[\{I - kMM^T B_k^{-1}\}W_k h(W_k)] \\ &= SB_k^{-1}\mathbb{E}[W_k h(W_k)]. \end{aligned} \tag{C.2.12}$$

For $\lambda \sim N(0, I)$ with $I \in \mathbb{R}^{p\times p}$ we have that $W_k \overset{d}{=} B_k^{1/2}\lambda$. Thus, we can write (C.2.12) as

$$\mathbb{E}[Vh(ME_k + V)] = SB_k^{-1}\mathbb{E}[W_k h(W_k)] = SB_k^{-1/2}\, \mathbb{E}[\lambda h(B_k^{1/2}\lambda)]. \tag{C.2.13}$$

Since h is bounded, there exists $M > 0$ such that

$$|h(v)| \leq M, \quad v \in \mathbb{R}^p. \tag{C.2.14}$$

Moreover, from Proposition 3.6 it holds that

$$\delta_0 = P\gamma_0 = R(R^T S R)^{-1} R^T S \gamma_0 \in \mathbb{R}^p, \tag{C.2.15}$$

where $R \in \mathbb{R}^{p\times(p-r)}$ denotes an orthonormal basis for $\ker(M^T)$. Therefore, from (C.2.13), (C.2.14), and (C.2.15), and using Cauchy–Schwarz, it follows that

$$\begin{aligned} \Big|(\gamma_0 - \delta_0)^T \mathbb{E}[Vh(ME_k + V)]\Big| &= \Big|\gamma_0^T (I-P)^T S B_k^{-1/2}\ \mathbb{E}[\lambda h(B_k^{1/2}\lambda)]\Big| \\ &\leq \|\gamma_0^T (I-P)^T S B_k^{-1/2}\|_2\ \|\mathbb{E}(|\lambda|)M\|_2. \end{aligned} \tag{C.2.16}$$

Notice that $\|\mathbb{E}(|\lambda|)M\|_2 < \infty$. Moreover, using results from Lemma C.6, we have that

$$\begin{aligned} \frac{\|\gamma_0^T (I-P)^T S B_k^{-1/2}\|_2}{\|\gamma_0\|_2} &\leq \|(I-P)^T S B_k^{-1/2}\|_F \\ &= \|(I-P)^T S R R^T B_k^{-1/2} + (I-P)^T S B_k^{-1/2} - (I-P)^T S R R^T B_k^{-1/2}\|_F \\ &\leq \|(I-P)^T S R R^T B_k^{-1/2}\|_F + \|(I-P)^T S (B_k^{-1/2} - R R^T B_k^{-1/2})\|_F \\ &\leq \|(I-P)^T S R\|_F\ \|R^T B_k^{-1/2}\|_F + \|(I-P)^T S\|_F\ \|B_k^{-1/2} - R R^T B_k^{-1/2}\|_F. \end{aligned} \tag{C.2.17}$$

Notice that

$$\|(I-P)^T S R\|_F = \|SR - SR(R^T S R)^{-1} R^T S R\|_F = 0.$$

Furthermore, from Lemma C.5, we have that $\|R^T B_k^{-1/2}\|_F < \infty$. Also, from Lemma C.2, we have that $\|B_k^{-1/2} - R R^T B_k^{-1/2}\|_F \to 0$ as $k \to \infty$. So, using (C.2.17), it follows that

$$\|\gamma_0^T (I-P)^T S B_k^{-1/2}\|_2 \to 0, \quad \text{as } k \to \infty,$$

which in turn implies, using (C.2.16), that

$$\Big|(\gamma_0 - \delta_0)^T \mathbb{E}[Vh(ME_k + V)]\Big| \to 0, \quad \text{as } k \to \infty.$$

Therefore, for any $\eta > 0$, there exists a $k_0 > 0$ such that for all $k \geq k_0$ it holds that

$$-\eta < (\gamma_0 - \delta_0)^T \mathbb{E}[Vh(ME_k + V)] = \int_{\mathbb{R}^r} (\gamma_0 - \delta_0)^T \mathbb{E}[Vh(Me + V)]\phi_{kI}(e)\, de < \eta.$$

Therefore, there exists $\tilde{e} \in \mathbb{R}^r$ such that

$$(\gamma_0 - \delta_0)^T \mathbb{E}[Vh(M\tilde{e} + V)] < \eta.$$

□

Lemma C.5. *For any $k > 0$, it holds that $\|R^T B_k^{-1/2}\|_F < \infty$.*

Proof. Let $k > 0$. It holds that

$$\begin{aligned} \|R^T B_k^{-1/2}\|_F \leq \|R\|_F \|B_k^{-1/2}\|_F &= \|R\|_F \|R_k \tilde{\Lambda}_{0k}^{-1/2} R^T + Q_k \tilde{\Lambda}_{1k}^{-1/2} Q_k^T\|_F \\ &\leq \|R\|_F \left(\|R_k \tilde{\Lambda}_{0k}^{-1/2}\|_F + \|Q_k \tilde{\Lambda}_{1k}^{-1/2}\|_F\right) \\ &\leq \|R\|_F \left(\|R_k\|_F\ \|\tilde{\Lambda}_{0k}^{-1/2}\|_F + \|Q_k\|_F\ \|\tilde{\Lambda}_{1k}^{-1/2}\|_F\right) \\ &\leq \sqrt{p-r}\left(\frac{p-r}{\sqrt{\lambda_1^S}} + \frac{r}{\sqrt{k\lambda_{p-r+1}}}\right) < \infty. \end{aligned}$$

□

The next lemma provides useful results about the Frobenius norm.

Lemma C.6. *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$ and $v \in \mathbb{R}^n$. Let $C \in \mathbb{R}^{q\times n}$ such that $C^TC = I_n$. Then, it holds that*

$$\begin{aligned} \|Av\|_2 &\leq \|A\|_F\|v\|_2, \\ \|AB\|_F &\leq \|A\|_F \ \|B\|_F, \\ \|A+B\|_F &\leq \|A\|_F + \|B\|_F, \\ \|AC^T\|_F &= \|CA^T\|_F = \|A\|_F. \end{aligned}$$

where $\|A\|_F := \sqrt{\mathrm{tr}(A^TA)}$.

Bibliography

Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1576–1584.

Angrist, J. D., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74(2):539–563.

Angrist, J. D., Chernozhukov, V., and Fernández-Val, I. (2009). Replication data for: Quantile regression under misspecification, with an application to the U.S. wage structure. https://doi.org/10.7910/DVN/JNEOLQ.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *Annals of Applied Statistics*, 9(4):2023–2050.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Balkema, A. A. and de Haan, L. (1974). Residual Life Time at Great Age. *The Annals of Probability*, 2(5):792 – 804.

Basrak, B. and Segers, J. (2009). Regularly varying multivariate time series. *Stochastic processes and their applications*, 119(4):1055–1080.

Beirlant, J., Dierckx, G., and Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots. *Bernoulli*, 11(6):949 – 970.

Beirlant, J., Wet, T. D., and Goegebeur, Y. (2004). Nonparametric estimation of extreme conditional quantiles. *Statistical Computation and Simulation*, 74(8):567–580.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(38):1063–1095.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons Inc., New York, NY, USA.

Breiman, L. (2001). Random forests. *Machine Learning, 45, 5–32*.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. *Wadsworth and Brooks,Monterey, California*.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Bücher, A., Lilienthal, J., Kinsvater, P., and Fried, R. (2020). Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis. *Extremes*, pages 1–24.

Bücher, A. and Segers, J. (2017). On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes*, 20(4):839–872.

Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science*, 35(3):404–426.

Bühlmann, P., Peters, J., and Ernest, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics*, 42:2526–2556.

Cai, J. J., Einmahl, J. H. J., de Haan, L., and Zhou, C. (2015). Estimation of the marginal expected shortfall: the mean when a related variable is extreme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:417–442.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):207–222.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, 33(2):806 – 839.

Christiansen, R., Pfister, N., Jakobsen, M., Gnecco, N., and Peters, J. (2021). A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 1–1.

Claassen, T., Mooij, J., and Heskes, T. (2013). Learning sparse causal models is not np-hard. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 172–181. ACM Press.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, New York, NY, USA.

Coles, S. G. and Dixon, M. J. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23.

Coles, S. G. and Tawn, J. A. (1996). A bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(4):463–478.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*, volume 67 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, USA.

Daouia, A., Gardes, L., Girard, S., and Lekina, A. (2011). Kernel estimators of extreme level curves. *Test, Spanish Society of Statistics and Operations Research/Springer*, 20(2):311–333.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*. Springer, New York, NY, USA.

de Haan, L., Sinha, A. K., et al. (1999). Estimating the probability of a rare event. *The Annals of Statistics*, 27(2):732–759.

de Zea Bermudez, P. and Turkman, M. A. (2003). Bayesian approach to parameter estimation of the generalized pareto distribution. *Test*, 12(1):259–277.

Deuber, D., Li, J., Engelke, S., and Maathuis, M. H. (2021). Estimation and inference of extremal quantile treatment effects for heavy-tailed distributions. *arXiv preprint arXiv:2110.06627*.

Dombry, C. (2015). Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli*, 21(1):420 – 436.

Dombry, C. and Ferreira, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli*, 25(3):1690–1723.

Drees, H., Ferreira, A., and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index. *Ann. Appl. Probab.*, 14(3):1179–1201.

Dunker, F. (2021). Adaptive estimation for some nonparametric instrumental variable models with full independence. *Electronic Journal of Statistics*, 15(2):6151–6190.

Dunker, F., Florens, J.-P., Hohage, T., Johannes, J., and Mammen, E. (2014). Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics*, 178:444–455.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events: for Insurance and Finance*. Springer, New York, NY, USA.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2012). *Modelling Extremal Events for Insurance and Finance*. Stochastic Modelling and Applied Probability. Springer Heidelberg New York Dordrecht London, 9^{th} edition.

Engelke, S., de Fondeville, R., and Oesting, M. (2019). Extremal behaviour of aggregated data with an application to downscaling. *Biometrika*, 106:127–144.

Engelke, S. and Hitz, A. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82:871–932. Discussion paper.

Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8. To appear.

Engelke, S. and Volgushev, S. (2020). Structure learning for extremal tree models. *arXiv preprint arXiv:2012.06179*.

Entner, D. and Hoyer, P. O. (2010). Discovering unconfounded causal relationships using linear non-gaussian models. In *Proceedings of the 2010 international conference on New Frontiers in Artificial Intelligence*, pages 181–195. Springer.

Farkas, S., Lopez, O., and Thomas, M. (2020). Cyber claim analysis through generalized pareto regression trees with applications to insurance pricing and reserving. Preprint at https://hal.archives-ouvertes.fr/hal-02118080v2.

Fawcett, L. and Walshaw, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics*, 18(2):173–188.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Volume II*. John Wiley & Sons Inc., New York, NY, USA.

Ferreira, A., de Haan, L., and Zhou, C. (2012). Exceedance probability of the integral of a stochastic process. *J. Multivariate Anal.*, 105:241–257.

Forbes, K. J. and Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance*, 57(5):2223–2261.

Friedman, J. H. (2001a). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. (2001b). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378.

Gardes, L. and Stupfler, G. (2019). An integrated functional Weissman estimator for conditional extreme quantiles. *REVSTAT*, 17(1):109–144.

Gissibl, N. and Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A):2693–2720.

Gissibl, N., Klüppelberg, C., and Lauritzen, S. (2020). Identifiability and estimation of recursive max-linear models. *Scandinavian Journal of Statistics*.

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

Greene, W. H. (2003). *Econometric analysis*. Pearson Education, Upper Saddle River, New Jersey, USA.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM*, 7(12):701–702.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(1):3365–3383.

Hastie, T. J., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, NY, USA, second edition.

Heagerty, P. J. and Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 13(5):1163–1174.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008). Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152.

Imbens, G. (2014). Instrumental variables: an econometrician's perspective. Technical report, National Bureau of Economic Research.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, New York, NY, USA.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239 – 262.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Journal of the Econometric Society*, 46(1):33–50.

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *International journal of computer mathematics*, 2(1-4):157–168.

Lauritzen, S. L. (1996). *Graphical models.* Oxford Statistical Science Series. A Clarendon Press Publication. Oxford University Press, New York, NY, USA.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H. G. (1990). Independence properties of directed markov fields. *Networks*, 20(5):491–505.

Martins-Filho, C., Yao, F., and Torero, M. (2015). High-order conditional quantile estimation based on nonparametric models of regression. *Econometric Reviews*, 34(6 - 10):907–958.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Meinshausen, N. (2018). Causality from a distributional robustness point of view. In *IEEE Data Science Workshop*, pages 6–10.

Mhalla, L., Chavez-Demoulin, V., and Dupuis, D. J. (2020). Causal mechanism of extreme river discharges in the upper danube basin network. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(4):741–764.

Misra, N. and Kuruoglu, E. E. (2016). Stable graphical models. *Journal of Machine Learning Research*, 17(168):1–36.

Naveau, P., Ribes, A., Zwiers, F., Hannart, A., Tuel, A., and Yiou, P. (2018). Revising return periods for record events in a climate event attribution context. *Journal of Climate*, 31(9):3411–3422.

Nestlé (2019). Financial statements 2019. *Online release*, page 89.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Newey, W. K., Powell, J. L., and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603.

Ng, S. and Pinkse, J. (1995). Nonparametric-two-step estimation of unknown regression functions when the regressors and the regression error are not independent. *Cahier de recherche*, 9551.

Novartis (2019). Annual report 2019. *Online release*, pages F–25–26.

Pearl, J. (2009a). *Causality*. Cambridge University Press, New York, USA, 2nd edition.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition.

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

Peters, J. and Bühlmann, P. (2015). Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799.

Peters, J., Janzing, D., Gretton, A., and Schölkopf, B. (2009). Detecting the direction of causal time series. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 801–808. ACM Press.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053.

Pickands, J. I. (1975). Statistical inference using extreme value order statistics. *Annals of Statististics*.

Poirier, A. (2017). Efficient estimation in models with independence restrictions. *Journal of Econometrics*, 196(1):1–22.

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., and Schwaighofer, A. (2009). *Dataset shift in machine learning*. Mit Press.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer, New York, NY, USA.

Roche (2019). Finance report 2019. *Online release*, page 160.

Rodriguez, J. C. (2007). Measuring financial contagion: A copula approach. *Journal of Empirical Finance*, 14(3):401–423.

Rosén, B. (1965). Limit theorems for sampling from finite populations. *Arkiv för Matematik*, 5:383–424.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246.

Schlather, M. and Tawn, J. A. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716 – 1741.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3-4):125–161.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248.

Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Smith, R. L. (1990). Regional estimation from spatially dependent data. Unpublished.

Smith, R. L. and Naylor, J. (1987). A comparison of maximum likelihood and bayesian estimators for the three-parameter weibull distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):358–369.

Spirtes, P., Glymour, C. N., and Scheines (2000). *Causation, Prediction, and Search*. MIT press, Cambridge, MA, USA.

Stone, C. J. (1980). Optimal Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, 8(6):1348 – 1360.

Stone, C. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053.

Taillardat, M., Fougères, A.-L., Naveau, P., and Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34(3):617–634.

Tashiro, T., Shimizu, S., Hyvärinen, A., and Washio, T. (2014). Parcelingam: A causal ordering method robust against latent confounders. *Neural Computation*, 26(1):57–83.

Taylor, J. W. (1999). A quantile regression approach to estimating the distribution of multiperiod returns. *The Journal of Derivatives*, 7(1):64–78.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311.

Tibshirani, J., Athey, S., Sverdrup, E., and Wager, S. (2021). *grf: Generalized Random Forests*. R package version 2.0.2.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Velthoen, J., Cai, J.-J., Jongbloed, G., and Schmeits, M. (2019). Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2021). Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, H. and Tsai, C.-L. (2009). Tail index regression. *Journal of the American Statistical Association*, 104(487):1233–1240.

Wang, H. J. and Li, D. (2013). Estimation of extreme conditional quantiles through power transformation. *American Statistical Association*, pages 1062–1074.

Wang, H. J., Li, D., and He, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *American Statistical Association*, pages 1453–1464.

Wang, Y. S. and Drton, M. (2020). Causal discovery with unobserved confounding and non-gaussian data. *arXiv preprint arXiv:2007.11131*.

Weyl, H. (1912). Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, MA.

Yang, S. (1999). Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association*, 94(445):137–145.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for u.s. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 52(3):331–350.

Zhang, J. (2008). Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(Jul):1437–1474.

Zou, N., Volgushev, S., and Bücher, A. (2021). Multiple block sizes and overlapping blocks for multivariate time series extremes. *The Annals of Statistics*, 49(1):295–320.