ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Winter 2024

Pio Blieske

Inference of Nonlinear Causal Effects in Time Series in the Presence of Confounding

Submission Date: 17 March 2025

Co-Advisor Felix Schur
Advisor: Prof. Dr. Jonas Peters

In memory of my father, who never saw this adventure.

Acknowledgements

I would like to express my deepest appreciation to my advisor Jonas Peters for all the support, inspiration, and great advice. Without him, this thesis would not have been possible. I am very grateful to Felix Schur for all the discussions, all the help, and good ideas. He invested a lot of time and was always available when I ran into problems. Moreover, I would like to thank the members of the Seminar of Statistics at ETH Zürich, the Max Planck Institute for Evolutionary Biology, the Theoretical Biology Group at ETH, Sebastian Bonhoeffer, Roland Regös, Samuel Zweifel and Kwok Wai Lui for valuable comments, suggestions and discussions. I also thank Ada Stein for taking the time to read this thesis and offering invaluable feedback.

Code and data availability

All the code and data used can be found online in the GitHub repository together with a PDF file of the newest version of this thesis.

Contents

Abstract

Estimating causal effects between time series is of interest in many scientific fields such as climate science, epidemiology, and economics, but remains challenging due to possible confounding. This thesis focuses on the inference of nonlinear causal effects between two time series in the presence of a third unobserved, confounding time series. We assume sparsity of the confounder in the frequency domain, corresponding in applications, for example, to a seasonal periodicity. By developing a new transformation for the data, we leverage the sparsity assumption to reduce the confounding problem to an adversarial outlier problem, a technique known as deconfounding by robust regression (DecoR). We then use the robust regression algorithm Torrent to solve the adversarial outlier problem. To improve the estimation accuracy, we extend Torrent to a regularized version, which allows the incorporation of a smoothness penalty in DecoR, and provide upper bounds for the estimation error. For two different asymptotic settings, we prove the consistency of the nonlinear extensions of DecoR under suitable assumptions. We validate the nonlinear extensions of DecoR by a simulation study on synthetic data. In addition, we demonstrate its effectiveness with an application to a real-world example of environmental epidemiology.

Chapter 1: Introduction

1.1 Motivation and Problem Setting

Lots of scientific research is not just concerned with the question of correlation, but also seeks to understand the underlying causal system. Striving for answers in this matter led to the emergence of the field of causal inference (Pearl, 2009; Peters et al., 2017). Answering causal questions is especially difficult if one does not have the option of controlling the desired covariates, making well-established methods like random control trials unfeasible. This scenario often occurs when working with time series, as in epidemiology, medicine, climate science, or economics. There, one does not have the option to intervene in the system for the purpose of fundamental research.

To overcome these issues, we follow an approach inspired by Mahecha et al. (2010) that was first formalized by Schur and Peters (2024). The underlying assumption of this approach is that the influence of the confounding covariate on the outcome is sparse in the frequency domain, e.g. has a seasonal periodicity. This allows us to reduce the confounding problem to a robust regression problem with, for example, a Fourier transformation. While for a linear causal effect the transformation leads to a linear robust regression problem, this is no longer the case if the underlying causal effect (denoted by the function f) is not linear. This leads to the question of how to transform the data to be able to use the sparsity assumption for nonlinear effects, and whether the consistency results of Schur and Peters (2024) continue to hold. Since common nonparametric regression techniques like kernel regression, local polynomial, smoothing spline, or any kind of kernel tricks require the selection of a bandwidth or regularization parameter, this is a challenging problem as high-dimensional robust regression needs further assumptions or restrictions, see Filzmoser and Nordhausen (2021). Thus, an additional challenge is the reduction of the dimension, or finding alternative suitable restrictions or regularization techniques.

1.2 Main Results

We extend the techniques of Schur and Peters (2024) to the setting where the causal relation between the covariate and the outcome is no longer linear. To this end, a suitable transformation of the data is derived such that the problem can be reduced to a regression problem with adversarial outliers due to the sparsity. The transformation is inspired by the Fourier transformation of function compositions of Bergner et al. (2006). By approximating the underlying nonlinear functions using an orthonormal, uniformly bounded basis of the L^2-space, we obtain a linear, robust regression problem with a bias. The bias vanishes with an increasing number of observations. We extend the results from Schur and Peters (2024) and show consistency in the L^2-norm under some smoothness assumption on f. This assumption is, for example, fulfilled for nonlinear functions f on a compact one-dimensional interval that are twice continuously differentiable. All the asymptotic results hold for two settings, namely for shorter time intervals, i.e. when the number of observations on a fixed time interval increases, and when we are obtaining observations over longer time horizons. We prove that this two settings are equivalent.

To solve the adversarial outlier problem, we use the robust regression algorithm Torrent

from Bhatia et al. (2015). It turns out that the number of basis functions used for the approximation is crucial. To reduce this dependence, we developed a regularized version of Torrent and derive statistical guarantees for the regularized version. When using the Fourier or cosine basis, the regularization allows us to directly penalize the smoothness of the underlying nonlinear function. For selecting the regularization parameter, we propose an out-of-bootstrap generalization error estimation scheme. A simulation study using synthetic data shows that regularization can improve the results in the presence of large noise or when there are only few observations. Furthermore, the simulations show that the nonlinear extensions of DecoR can even yield consistent estimations under model violations, even in the setting of non-vanishing fractions of outliers in the frequency domain.

Finally, we apply the nonlinear extension of DecoR to a real-world dataset from environmental epidemiology, investigating the influence of ozone on health. The application reveals that DecoR mainly removes low frequencies, indicating a seasonal confounding, and yields an estimator of similar magnitude as the state of the art method discussed in Bhaskaran et al. (2013).

1.3 Related Literature

Our work adds to the literature on causal inference (Rubin, 2005; Pearl, 2009; Peters et al., 2017) and particularly extends the work of Schur and Peters (2024). Similar approaches have been developed by Mahecha et al. (2010) when the support of the confounder in the frequency domain is known and by Sippel et al. (2019) for meteorological data where slowly varying frequencies are removed. We also add to the literature of (higher-dimensional) robust regression (Suggala et al., 2019; Filzmoser and Nordhausen, 2021; D'Orsi et al., 2021), mainly by using projections on subspaces or assuming sparsity. Extending the work of Bhatia et al. (2015) to a regularized version of Torrent, we build on smoothness assumptions. With this, we also provide an alternative to DecoR for generalized linear models of Shen and Sanghavi (2019) by not making assumptions about the underlying functional relationship. Studies on the selection of regularization parameters are scarce and focus on cross-validation for bandwidth selection in kernel regression (Leung, 2005; Čížek and Sadıkoğlu, 2020) using the ℓ_1- or Huber loss. We propose a new method by introducing an out-of-bootstrap estimation scheme that reduces the risk of bad splits. Inference of nonlinear functions using a Fourier approximation was studied by Popiński (1993). We formalized the observations in Bergner et al. (2006) about the Fourier transformation of function compositions and apply them to time series. Additionally, we add to the methodology of environmental epidemiology and provide an alternative to the standard method for seasonal deconfounding of Schwartz et al. (1996) and Bhaskaran et al. (2013) and review the effect of ozone on health (Nuvolone et al., 2018; World Health Organization and others, 2021).

This thesis is structured as follows. In Chapter 2 we state the problem setting and review the main results of Schur and Peters (2024). These results are then extended to nonlinear causal effects in Chapter 3. Chapter 4 introduces a regularized version of Torrent and discusses the application to DecoR. A simulation study confirming the theoretical results and demonstrating robustness for some model violations is provided in Chapter 5. In Chapter 6 we show an application of the nonlinear extensions to an example from environmental epidemiology. Chapter 7 discusses the results and possible directions of future research.

Chapter 2: Estimating Causal Effects in Time Series

In this chapter, we formalize the problem setting and discuss the connection to robust regression problems and causal inference. We present the main results for linear causal effects of Schur and Peters (2024) which will be extended to the nonlinear effects in Chapter 3.

Setting 1: *Let $d \in \mathbb{N}$ and $T \in \mathbb{R}_{>0}$ denote a fixed time horizon. Let $X = (X_t)_{t\in[0,T]}$ be a stochastic process on $[a, b]^d$, for some $a < b \in \mathbb{R}$, and $U = (U_t)_{t\in[0,T]}$ be a stochastic process in $\mathbb{R}$. Let $\eta = (\eta_t)_{t\in[0,T]}$ be a process of i.i.d. centered Gaussian random variables with constant variance $\sigma_\eta^2 \geq 0$ and independent of X. Let $f : [a, b]^d \to \mathbb{R}$ and $Y = (Y_t)_{t\in[0,T]}$ be a stochastic process that satisfies*

$$Y_t = f(X_t) + U_t + \eta_t \tag{1}$$

Fix $n \in \mathbb{N}$. We assume that we observe $X^n := (X_{T/n}, X_{2T/n}, \ldots, X_T)$ and $Y^n := (Y_{T/n}, Y_{2T/n}, \ldots, Y_T)$, where $(U_t)_{t\in[0,T]}$ is an unobserved confounder.

The main goal is to estimate the function f that gives the causal relationship between X_t and Y_t, see Section 2.2. Due to the confounding of U_t, standard regression methods such as kernel methods or smoothing splines fail because, in addition to f, they also infer for any $x \in [a, b]^d$ the confounding effect $\mathbb{E}[U_t \mid X_t = x]$, which is non-zero in general. In this chapter, we assume that f is a linear function as in Schur and Peters (2024). Later, we will study the case where f is nonlinear and make some mild regularity assumptions on f to guarantee consistency of our estimator of f in the L^2-space.

2.1 Sparse Representation

Consider an orthonormal basis $\phi = \{\phi_k\}_{k\in\mathbb{N}}$ of $L^2([0, T])$. We start by defining sparsity with respect to the orthonormal basis ϕ.

Definition 2.1 ((ϕ, G)-sparse process)**:** *Let $G \subseteq \mathbb{N}$ and $(U_t)_{t\in[0,T]}$ be a stochastic process in $\mathbb{R}$ satisfying $\mathbb{E}\left[\int_0^T U_t^2 dt\right] < \infty$. If for all $k \notin G$ almost surely*

$$\langle U, \phi_k \rangle_{L^2} = 0,$$

we call U a (ϕ, G)-sparse process.

Throughout the thesis, we assume that U is a (ϕ, G)-sparse process with G sufficiently small. In applications ϕ is often the Fourier or cosine basis. Thus, being sparse means having influence on only a few of the frequencies of Y.

Assumption 1: *The set $G \subseteq \mathbb{N}$ is such that U is a (ϕ, G)-sparse process.*

We can exploit the sparsity of U to estimate f. Let $\mathcal{S}$ be the set of $\mathbb{R}^d$-valued stochastic processes with domain $[0, T]$ and $\mathcal{C}$ the set of random variables on $\mathbb{R}^d$. For all $k \leq n$, define the transformation $T_k^{\phi,n} : \mathcal{S} \to \mathcal{C}$ as

$$T_k^{\phi,n}(V) := \begin{bmatrix} \frac{1}{n}\sum_{\ell=1}^{n}(V_{T\ell/n})_1\phi_k(T\ell/n) \\ \vdots \\ \frac{1}{n}\sum_{\ell=1}^{n}(V_{T\ell/n})_d\phi_k(T\ell/n) \end{bmatrix}. \tag{2}$$

Applying the transformation to Y and using linearity implies that for all $k \leq n$

$$\begin{aligned} T_k^{\phi,n}(Y) &= T_k^{\phi,n}(f(X)) + T_k^{\phi,n}(U) + T_k^{\phi,n}(\eta) \\ &\overset{\text{a.s.}}{=} \begin{cases} T_k^{\phi,n}(f(X)) + T_k^{\phi,n}(U) + T_k^{\phi,n}(\eta), & \text{if } k \in G, \\ T_k^{\phi,n}(f(X)) + T_k^{\phi,n}(\eta), & \text{else.} \end{cases} \end{aligned} \tag{3}$$

Here we made use of a technical assumption (see Assumption 2 for the details). Thus, the transformation gives us a new sample in the frequency domain where only few observations are confounded, namely the one in G. After the transformation, the problem can be seen as a robust regression problem with adversarial outliers. We have to remove the observations where $T_k^{\phi,n}(U) \neq 0$ to get rid of the confounding. Thus, the first main difficulty is to find or obtain a good estimator for G. The second challenge lies in finding a suitable expression of $T_k^{\phi,n}(f(X))$, that is allowing the inference of the underlying function f.

2.2 Connection to Causal Inference

In this section, we make the connection to causal inference using the notion of structural causal models (see Appendix A.1 for a short introduction). Let $g : \mathbb{R}^2 \to [0, T]$ be an arbitrary function and $f : [a, b] \to \mathbb{R}$ be continuously differentiable. Consider the structural causal model $\mathfrak{C}$ given by the assignments

$$\begin{aligned} X_t &:= g(U_t, N_t) \\ Y_t &:= f(X_t) + U_t + \eta_t \end{aligned}$$

for all $t \in [0, T]$, with $(\eta_t)_{t\in[0,T]}$ an stochastic process independent of $(U_t)_{t\in[0,T]}$ and $(N_t)_{t\in[0,T]}$. Then, the causal effect form X_t to Y_t is given by

$$\mathbb{E}^{\mathfrak{C}}_{do(X_t:=x_t)}[Y_t] - \mathbb{E}[Y_t] = f(x_t) - \mathbb{E}[f(X_t)]$$

or in terms of derivatives

$$\frac{\partial}{\partial x_t}\mathbb{E}^{\mathfrak{C}}_{do(X_t:=x_t)}[Y_t] = f'(x_t).$$

Hence f describes how an intervention on X_t from outside changes the outcome of Y_t. Thus, f is of particular interest if we do not only want to learn about the correlation, but also about the underlying causal relationship. The directed acyclic graph of $\mathfrak{C}$ is given in Figure 2.1 below.

2.3 Robust Regression by Hard Thresholding – Torrent

Before solving the deconfounding problem, we first present a robust regression algorithm – Torrent – which will be essential in tackling the deconfounding problem. To this end, we introduce the setting of the linear model with adversarial outliers.

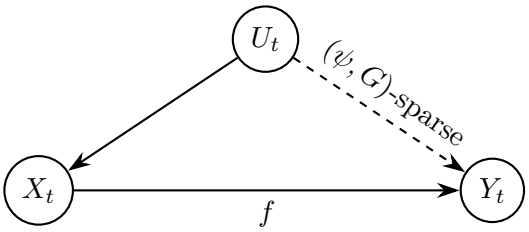


Figure 2.1: Directed graph of the underlying structural causal model $\mathfrak{C}$.

Setting 2: *Let $d \in \mathbb{N}$ and $\beta \in \mathbb{R}^d$. For all $n \geq d$, let $\epsilon \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$ and $G_n \subseteq \{1, \ldots, n\}$. Suppose that $o \in \mathbb{R}^n$ is such that $\forall i \notin G_n : o_i = 0$. Define $Y \in \mathbb{R}^n$ by*

$$Y := X\beta + \epsilon + o.$$

We call G_n the (potential) outliers, G_n^c the inliers and observe X and Y. The goal will be to estimate the coefficients β. For any subset $S \subseteq \{1, \ldots, n\}$ we denote by $X_S \in \mathbb{R}^{|S| \times d}$ the submatrix containing only the rows with indices in S, and similar Y_S, ϵ_S, o_S the subvectors induced by the rows in S. If $X_S^\top X_s$ has full rank, we denote by

$$\hat{\beta}_{OLS}^S(X, Y) := \left(X_S^\top X_S\right)^{-1} X_S^\top Y_S$$

the ordinary least-square estimator, minimizing the ℓ_2-loss $\|X_S\beta - Y_S\|_2^2$. Next, for all $v \in \mathbb{R}^n$ let $s_1, \ldots, s_n$ be a permutation of $\{1, \ldots, n\}$ such that $v_{s_1} \leq v_{s_2} \leq \cdots \leq v_{s_n}$ and introduce for $a \in \{1, \ldots, n\}$ the hard-threshold

$$HT(v, a) := \{s_1, \ldots, s_a\},$$

the indices corresponding to the a smallest elements of v. Given this definition, we can state the linear robust regression algorithm Torrent.

Algorithm 1 Torrent (Bhatia et al., 2015)

Require: $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$, $a \in \{1, \ldots, n\}$

```
S_0 ← {1, ..., n}, e ← Y, err ← ∞, t ← 0
while ‖e_{S_t}‖_2 < err do
    t ← t + 1
    err ← ‖e_{S_{t-1}}‖_2
    β̂^t_Tor ← β̂^{S_{t-1}}_OLS(X, Y)
    e ← |Y − X β̂^t_Tor|
    S_t ← HT(e, a)
end while
return β̂^{n,a}_Tor := β̂^t_Tor
```

Summarized, Torrent tries to iteratively detect the outliers in a greedy fashion. The following lemma guarantees that Torrent will terminate after finite many steps.

Lemma 2.2: *For all $n \in \mathbb{N}$ there is a constant $N \in \mathbb{N}$ such that Algorithm 1 converges in less than N steps.*

The proof can be found in Appendix B.1. The next theorem gives a non-asymptotic bound on the error in the 2-norm of the estimated coefficients $\hat{\beta}_{Tor}^{n,a}$ obtained by Torrent.

Theorem 2.3: *Assume Setting 2 and that there exists a known $c_n \in \mathbb{N}$ such that $|G_n| \leq c_n$. Let S_t be the estimated subset in the final iteration of Torrent executed on the data X, Y using the threshold parameter $a_n = n - c_n$. Define for $S \subseteq \{1, \dots, n\}$*

$$V(S) := (S \cup G_n^c) \setminus (G_n^c \cap S),$$

the symmetric difference between S and G_n^c. Furthermore, assume that

$$\eta := \max_{S \subseteq \{1,\dots,n\} \text{ s.t. } |S|=a_n} \frac{\|X_{V(S)}\|_2}{\sqrt{\lambda_{min}\left(X_S^T X_S\right)}} < \frac{1}{\sqrt{2}}. \tag{4}$$

Then,

$$\left\|\hat{\beta}_{Tor}^{n,a_n} - \beta\right\|_2 \leq \left\|\left(X_{S_t}^\top X_{S_t}\right)^{-1} X_{S_t}^\top \epsilon_{S_t}\right\|_2 + \frac{\sqrt{2}\left\|X_{V(S_t)}\right\|_2 \left\|\left(X_{S_t}^\top X_{S_t}\right)^{-1} X_{S_t}^\top \epsilon_{S_t}\right\|_2 + \sqrt{2}\left\|\epsilon_{V(S_t)}\right\|_2}{\sqrt{\lambda_{min}\left(X_{S_t}^\top X_{S_t}\right)}(1-\sqrt{2}\eta)}.$$

Defining

$$\gamma_{a_n}(X) := \min_{S \subseteq \{1,\dots,n\} \text{ s.t. } |S|=a_n} \sqrt{\lambda_{\min}\left(X_S^\top X_S\right)},$$

we can use this result to derive an upper bound for the estimation error in the sub-Gaussian noise setting.

Corollary 2.4 (Sub-Gaussian noise)**:** *Assume the setting of Theorem 2.3 with i.i.d. zero-mean sub-Gaussian noise ϵ with variance proxy σ^2 and assume $c_n < n/2$. Then, there exists a constant $K > 0$ such that for all $\delta > 0$ with probability at least $1 - \delta$*

$$\begin{aligned}\left\|\hat{\beta}_{Tor}^{n,a_n} - \beta\right\|_2 \leq & \frac{\sigma}{\gamma_{a_n}(X)}\left(1 + \frac{\sqrt{2}}{1-\sqrt{2}\eta}\right)\left(\sqrt{d} + \sqrt{2c_n K \log(2en/c_n\delta)}\right) \\ & + \frac{2\sigma\sqrt{c_n}\left(1 + \sqrt{K\log(2en/c_n\delta)}\right)}{\gamma_{a_n}(X)(1-\sqrt{2}\eta)}.\end{aligned}$$

In particular, if the rows of X are i.i.d. standard Gaussian random vectors and $c_n \in o(n/\log(n))$, then $\hat{\beta}_{Tor}^{n,a_n}$ is consistent.

2.4 Deconfounding by Robust Regression – DecoR

Having introduced the robust regression algorithm Torrent, we now return to the original setting and assume that the causal relation is linear.

Setting 1′ : *Let $\beta \in \mathbb{R}^d$. Assume a modified version of Setting 1 where we replace Equation (1) by*

$$Y_t = X_t^\top \beta + U_t + \eta_t$$

and lift the restriction that for all $t \in [0, T]$, $X_t \in [a, b]^d$.

Recall that for $n \in \mathbb{N}$ we observe the variables $X^n := (X_{T/n}, X_{2T/n}, \dots, X_T)$ and $Y^n := (Y_{T/n}, Y_{2T/n}, \dots, Y_T)$. Consider the transformations $T_k^{\phi,n}$, see Equation (2), and write $X_\phi^n := (T_1^{\phi,n}(X), \dots, T_n^{\phi,n}(X))^\top$. Let $k \leq n$, due to the linearity of the transformation $T_k^{\phi,n}$, Equation (3) becomes

$$T_k^{\phi,n}(Y) \overset{\text{a.s.}}{=} \begin{cases} T_k^{\phi,n}(X)^\top \beta + T_k^{\phi,n}(U) + T_k^{\phi,n}(\eta), & \text{if } k \in G, \\ T_k^{\phi,n}(X)^\top \beta + T_k^{\phi,n}(\eta), & \text{else.} \end{cases} \tag{5}$$

With this decomposition, we can use a linear robust regression algorithm to estimate the coefficients β. This leads to the deconfounding by robust regression algorithm, short DecoR, outlined below. DecoR consists of two main steps. First, we transform the problem into a robust linear regression problem. Second, we apply a robust regression algorithm $\mathcal{A}$ to obtain an estimation of the coefficients β.

Algorithm 2 DecoR (Schur and Peters, 2024)

Require: $X^n \in \mathbb{R}^{n\times d}$, $Y^n \in \mathbb{R}^n$, orthonormal basis ϕ, robust linear regression algorithm $\mathcal{A}$

$X_\phi^n \leftarrow (T_1^{\phi,n}(X), \dots, T_n^{\phi,n}(X))^\top$

$Y_\phi^n \leftarrow (T_1^{\phi,n}(Y), \dots, T_n^{\phi,n}(Y))^\top$

$\hat{\beta}_{DecoR}^{\phi,n} \leftarrow \mathcal{A}(X_\phi^n, Y_\phi^n)$

return $\hat{\beta}_{DecoR}^{\phi,n}$

The theoretical results presented in this section are based on two types of assumptions. One assumption ensures that G_n does not grow too quickly with n (clearly, if $|G_n| \geq n/2$, for example, there is no consistent estimator for β). Another set of assumptions contains technical conditions on the transformations $T_k^{\phi,n}$.

Assumption 2: *(i) For all $n \in \mathbb{N}$ and $\ell, k \leq n$ it holds that*

$$\frac{1}{n}\sum_{j=1}^{n} \phi_\ell(Tj/n)\phi_k(Tj/n) = \mathbb{1}_{\{\ell=k\}}.$$

(ii) For every $\delta > 0$ there exists $c' > 0$ and $\overline{n} \in \mathbb{N}$ such that for all $n \geq \overline{n}$ there exists $S_n' \subseteq \{1, \dots, n\}$ with $|S_n'| = 2c_n + d$ such that for all $S'' \subseteq S_n'$ with $|S''| = d$ it holds that with probability at least $1 - \delta$

$$\lambda_{min}\left((X_\phi^n)_{S''}^\top (X_\phi^n)_{S''}\right) \geq c'. \tag{6}$$

Condition (i) guarantees that the orthonormal basis remains orthogonal and normalized when we apply it to discretized observations. This is satisfied, for example, for the cosine basis or the Haar basis (see Appendix A.1) and implies Equation (2). Condition (ii) ensures that the transformed features X_ϕ^n are not sparse in the frequency domain compared to the confounder. This condition holds for many of the commonly used stochastic processes, including Ornstein-Uhlenbeck processes, Brownian motions, and band-limited processes; see Section 4.1 in Schur and Peters (2024).

Theorem 2.5 (Convergence of DecoR-Tor)**:** *Let c_n be a known sequence of natural numbers such that $|G_n| \leq c_n$ and assume that Assumption 2 is satisfied. Suppose that for all $\delta > 0$ there exists $\overline{n} \in \mathbb{N}$ such that for all $n \geq \overline{n}$ it holds that*

$$\mathbb{P}\left[\max_{S\subseteq\{1,\dots,n\} \text{ s.t. } |S|=n-c_n} \frac{\left\|(X_\phi^n)_{V(S)}^\top\right\|_2}{\sqrt{\lambda_{min}\left((X_\phi^n)_S^\top (X_\phi^n)_S\right)}} < \frac{1}{\sqrt{2}}\right] \geq 1 - \delta. \tag{7}$$

If DecoR is run with Torrent and the sequence $a_n = n - c_n$ of threshold parameters, then

$$\left\| \hat{\beta}^{\phi,n}_{DecoR} - \beta \right\|_2 \in \mathcal{O}_{\mathbb{P}} \left(\sigma_\eta \sqrt{\frac{c_n \log(n/c_n)}{n}} \right).$$

Thus, if the assumptions are satisfied and there is no noise, $\sigma_\eta = 0$, or the number of confounded frequencies grows slower than the number of observations n, i.e. $c_n \in o(n)$, Theorem 2.5 states that DecoR executed with Torrent yields a consistent estimator of β. If $c_n \sim n$ and there is noise present, the estimation error is bounded by the noise variance σ_n up to a constant factor.

Chapter 3: Nonlinear Extensions of DecoR

Building on the ideas from Chapter 2, we are going to derive a suitable deconfounding method for nonlinear causal effects using the robust regression algorithm Torrent. Moreover, we prove the consistency in the L^2-norm for nonlinear extensions and show the equivalence of two different asymptotic settings.

3.1 Basisexpansions of Function Compositions

We focus on the case where the function f is defined on a bounded, compact domain $[a,b]^d$ for some $a < b \in \mathbb{R}$ and $d \in \mathbb{N}$. Alternatively, one can assume periodicity of f. We first study the transformation of compositions. Proposition 3.1 is formalized version of the observations of Bergner et al. (2006) in the study of spectral analysis of function composition for image synthesis.

Proposition 3.1: *Let $\{\psi_k\}_{k\in\mathbb{N}}$, $\{\phi_k\}_{k\in\mathbb{N}}$ be orthonormal, uniformly bounded bases of $L^2([a,b]^d)$ and $L^2([0,T])$, respectively. Let $f \in L^2([a,b]^d)$, define $d_k \coloneqq \langle f, \psi_k\rangle$ and assume that $\sum_{k=0}^{\infty} |d_k| < \infty$. Let $g \in L^2([0,T])$ with $a \le g \le b$, then $\psi_\ell \circ g \in L^2([0,T])$ and $f \circ g \in L^2([0,T])$. Furthermore*

$$\langle f \circ g, \phi_k\rangle_{L^2([0,T])} = \sum_{\ell=0}^{\infty} d_\ell \langle \psi_\ell \circ g, \phi_k\rangle_{L^2([0,T])}.$$

The proof for Proposition 3.1 can be found in Appendix B.2. The conditions are satisfied, for example, if we choose the Fourier basis and $f \in C^2([a,b])$, see Lemma 3.3. With Proposition 3.1 we have found a representation of $T_k^{\phi,n}(f(X))$ and can go back to our original problem in Setting 1. For this, assume that we are given $X^n = (X_{T/n}, \dots, X_T)$ and $Y^n = (Y_{T/n}, \dots, Y_T)$ and suppose that for some $f \in L^2([a,b])$ it holds that

$$Y_t = f(X_t) + U_t + \eta_t$$

for all $t \in [0,T]$. For $k \le n$ we can now compute the transformation $T_k^{\phi,n}$ of $f(X)$. For $\ell \in \mathbb{N}$ let the coefficients be defined as $d_\ell \coloneqq \langle f, \psi_\ell\rangle$. Since for all $i \in \{1,\dots,n\}$ we have

$$f\left(X_{iT/n}\right) \overset{a.s.}{=} \sum_{\ell=0}^{\infty} d_\ell \psi_\ell\left(X_{iT/n}\right),$$

by using the dominated convergence theorem it follows that

$$\begin{aligned} T_k^{\phi,n}(f(X)) &\overset{a.s.}{=} \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=0}^{\infty} d_\ell \psi_\ell(X_{iT/n})\phi_k(iT/n) \\ &\overset{a.s.}{=} \sum_{\ell=0}^{\infty} d_\ell \frac{1}{n}\sum_{i=1}^{n} \psi_\ell(X_{iT/n})\phi_k(iT/n) \\ &\overset{a.s.}{=} \sum_{\ell=0}^{\infty} d_\ell P_{\psi,\phi}^n(\ell,k) \end{aligned}$$

with $P_{\psi,\phi}^n(\ell,k)$ given by

$$P_{\psi,\phi}^n(\ell,k) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \psi_\ell(X_{iT/n})\phi_k(iT/n).$$

We also denote by $P^n_{\psi,\phi} \in \mathbb{R}^{L\times n}$ the matrix with elements $(P^n_{\psi,\phi})_{\ell,k} = P^n_{\psi,\phi}(\ell,k)$. Combining everything with Equation (3) and using the linearity of the transformation, we obtain

$$T^{\phi,n}_k(Y) \overset{a.s.}{=} \begin{cases} \sum_{\ell=0}^{\infty} d_\ell P^n_{\psi,\phi}(\ell,k) + T^{\phi,n}_k(U) + T^{\phi,n}_k(\eta), & \text{if } k \in G \\ \sum_{\ell=0}^{\infty} d_\ell P^n_{\psi,\phi}(\ell,k) + T^{\phi,n}_k(\eta), & \text{otherwise.} \end{cases} \tag{8}$$

We can approximate the series by neglecting higher terms, that is by choosing some $L \in \mathbb{N}$,

$$T^{\phi,n}_k(f(X)) \overset{a.s.}{=} \sum_{\ell=0}^{L} d_\ell P^n_{\psi,\phi}(\ell,k) + (r_L)_k \tag{9}$$

with $(r_L)_k$ being the error term given by $(r_L)_k = \sum_{\ell=L+1}^{\infty} d_\ell P^n_{\psi,\phi}(\ell,k)$. Hence, the coefficients d_ℓ can be estimated using an approximation by a linear regression problem with outliers.

3.2 Nonlinear DecoR

We summarize the observations and results obtained so far in the following algorithm, extending DecoR for solving the confounding problem for nonlinear causal effects.

Algorithm 3 Nonlinear DecoR

Require: $X^n \in \mathbb{R}^{n\times d}$, $Y^n \in \mathbb{R}^n$, orthonormal bases ψ and ϕ, robust linear regression algorithm $\mathcal{A}$

$$P^n_{\psi,\phi} \leftarrow \begin{pmatrix} \frac{1}{n}\langle \psi_1(X), \phi_1\rangle & \dots & \frac{1}{n}\langle \psi_1(X), \phi_\ell\rangle \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\langle \psi_n(X), \phi_1\rangle & \dots & \frac{1}{n}\langle \psi_n(X), \phi_\ell\rangle \end{pmatrix}$$

$$Y^n_\phi \leftarrow \left(T^{\phi,n}_1(Y), \dots, T^{\phi,n}_n(Y)\right)^\top$$

$\hat{d}^{\psi,n}_{DecoR} \leftarrow \mathcal{A}(P^n_{\psi,\phi}, Y^n_\phi)$

return $\hat{d}^{\psi,n}_{DecoR}$

This version differs from DecoR for linear effects as it requires a different transformation of the original sample. In this thesis, we focus on DecoR executed with the robust regression algorithm Torrent, Algorithm 1, but we stress that DecoR can be run with any robust linear regression algorithm $\mathcal{A}$. In particular, with a regularized version of Torrent as discussed later in Chapter 4.

3.3 Asymptotics of Short Time Intervals

In the following, we show that DecoR yields a consistent estimator when we choose Torrent (Algorithm 1) for the linear robust regression algorithm $\mathcal{A}$. For this section, we assume that we are in Setting 1 and Assumption 1 is satisfied. Moreover, we assume that Assumption 2 holds with $P^n_{\psi,\phi}$ in place of X^n_ϕ.

We are going to make use of the approximation in Equation (9). Observe that we are in the setting of Corollary 2.4 with an error $\tilde{\epsilon} = \epsilon + r_{L(n)}$ where ϵ has mean zero. Fortunately, Corollary 2.4 extends nicely to this case, as shown in the more general version Corollary 4.3. To obtain a consistent estimator, choose the number of coefficients $L(n)$ to be estimated such that $\lim_{n\to\infty} L(n) = \infty$ and $\lim_{n\to\infty} \frac{L(n)}{n} = 0$. This leads to the error bound given in the following theorem.

Theorem 3.2 (Nonlinear DecoR)**:** *Let $g : \mathbb{N} \to \mathbb{R}$ be a function such that $\lim_{n\to\infty} g(n) = 0$ and let $\{\psi_\ell\}_{\ell\in\mathbb{N}}$ be an ordered, orthonormal, uniformly bounded basis of $L^2([a,b]^d)$. Assume that $\forall t \in [0,T]$, $X_t \in [a,b]$ and for $d_\ell = \langle f, \psi_\ell\rangle$ we have $\sum_{\ell\geq L} |d_\ell| \in \mathcal{O}(g(L))$. Let c_n be a known sequence of natural numbers such that $|G_n| \leq c_n$ and assume that Assumption 2 is satisfied for $P^n_{\psi,\phi}$ (in place of X^n_ϕ). Suppose that for all $\delta > 0$ there exists $\overline{n} \in \mathbb{N}$ such that for all $n \geq \overline{n}$ it holds that*

$$\mathbb{P}\left[\max_{S\subseteq\{1,\dots,n\}\ s.t.\ |S|=n-c_n} \frac{\left\|(P^n_{\psi,\phi})^\top_{V(S)}\right\|_2}{\sqrt{\lambda_{min}\left((P^n_{\psi,\phi})^\top_S (P^n_{\psi,\phi})_S\right)}} < \frac{1}{\sqrt{2}}\right] \geq 1-\delta. \tag{10}$$

If nonlinear DecoR is run with Torrent and the sequence $a_n = n - c_n$ of threshold parameters, then

$$\left\|\hat{d}^{\psi,n}_{DecoR} - d\right\|_2 \in \mathcal{O}_{\mathbb{P}}\left(\sigma_\eta\left(\sqrt{\frac{c_n \log(n/c_n)}{n}} + \sqrt{\frac{L(n)}{n}}\right) + \sqrt{n} g(L(n))\right).$$

The proof of Theorem 3.2 can be found in Appendix B.2. The first term is the error caused by the confounding that also appears in the linear setting, the second is due to the increase in degrees of freedom with the sample size n, and the third term comes from the bias introduced by the approximation of the function. The eigenvalue condition Equation (10) is hard to check analytically but numerical computations in Appendix C.2 show that it tends to hold for a large number of examples. Furthermore, since ψ is an orthonormal basis, this does not only yield convergence of the coefficients, but also implies by Parseval's equality (see Proposition A.3), that

$$\left\|\hat{f}^{\psi,n}_{DecoR} - f\right\|_{L^2} = \left\|\hat{d}^{\psi,n}_{DecoR} - d\right\|_2.$$

Hence Theorem 2.5 implies that $\hat{f}^{\psi,n}_{DecoR}$ converges to f in $L^2([a,b]^d)$.

The next lemma shows that if we assume that f is twice continuously differentiable and we choose the Fourier basis, the conditions of Proposition 3.1 and Theorem 3.2 are fulfilled.

Lemma 3.3: *Assume that $f \in C^2([a,b])$ and let $\{\psi_n\}_{n\in\mathbb{Z}}$ be the Fourier basis of $L^2([a,b])$ given by $\psi_n(x) = e^{-\frac{2\pi i n x}{b-a}}$. Then $\sum_{n\in\mathbb{Z}} |d_n| < \infty$ where $d_n = \langle f, \psi_n\rangle$ and*

$$\sum_{m\in\mathbb{Z},\,|m|\geq n} |d_m| \in \mathcal{O}\left(n^{-3/2}\right).$$

The proof of Lemma 3.3 can be found in Appendix B.2. Lemma 3.3 together with Theorem 3.2 leads directly to the following corollary.

Corollary 3.4: *Under the assumptions of Theorem 3.2 and Lemma 3.3, choosing $L(n) \in \Theta(n^{1/2})$ yields*

$$\left\|\hat{f}^{\psi,n}_{DecoR} - f\right\|_2 \in \mathcal{O}_{\mathbb{P}}\left(\sigma_\eta\left(\sqrt{\frac{c_n \log(n/c_n)}{n}} + \frac{1}{n^{1/4}}\right) + \frac{1}{n^{1/4}}\right).$$

With higher degrees of smoothness of f, for example, if f is three times continuously differentiable, even faster convergence can be achieved due to faster decay of the coefficients d_ℓ.

Nonetheless, while Theorem 3.2 holds for multivariate stochastic processes, Corollary 3.4 fails in that regard since Lemma 3.3 does not carry over. The higher the dimension d of the stochastic process X_t is, the harder it is to achieve good convergence rates for the sum $\sum_{\ell \geq L} |d_\ell|$. However, this problem can be solved by assuming that f is more than twice continuously differentiable or by using an additive model. To this end, let $d \in \mathbb{N}$ and assume that $f : \mathbb{R}^d \to \mathbb{R}$ is given by

$$f(x_t) = b_0 + \sum_{i=1}^{d} b_i f_i \left(x_t^i\right),$$

with $f_i : \mathbb{R} \to \mathbb{R}$ twice continuously differentiable and $b_i \in \mathbb{R}$ for $i \in \{1, \ldots, d\}$. This allows us to use the same expansion as before to approximate every coordinate f_i separately. Thus, Corollary 3.4 still holds, as we can simply replace $L(n)$ with $dL(n)$.

3.4 Asymptotics of Long Time Horizons

In many applications, the setting differs regarding the fact that one does not obtain more measurements over a fixed time interval, but the time span over which one has measurements grows. Let $\Delta \in \mathbb{R}_{>0}$ be the time interval in which one observes X_t and Y_t, thus for $n \in \mathbb{N}$ at time $T_n := n\Delta$ we have observed $X_n = \left(X_0, X_\Delta, \ldots, X_{(n-1)\Delta}\right)$ and $Y_n = \left(Y_0, Y_\Delta, \ldots, Y_{(n-1)\Delta}\right)$. We extend the notion of sparsity as follows.

Definition 3.5 ($(\phi, \{G_n\})$-sparsity)**:** *Let $\{G_n\}_{n\in\mathbb{N}}$ be a sequence of sets $G_n \subseteq \{1, \ldots, n\}$ and ϕ an orthonormal basis of $L^2([0,1])$. We say U is $(\phi, \{G_n\})$-sparse if for all $k \leq n$ and $k \notin G_n$ almost surely*

$$\left\langle U_{t/T_n}, \phi_k \right\rangle = 0.$$

Note that in this setting, for $n \leq m$ the elements of G_n do not need to be in G_m. This restriction is lifted because we define sparsity over the rescaled version $U(t/T_n)$ and therefore the frequency k of the original, not scaled U corresponds to different frequencies ϕ_ℓ when we consider time intervals of different lengths T_n. With these adjustments, we are back in the setting of Theorem 3.2 and obtain the same convergence rate.

Corollary 3.6: *Assume that Assumption 2 holds for P^n_{t/T_n} and Y^n_{t/T_n} (in place of X^n_t and Y^n_t) and U is a $(\phi, \{G_n\})$-sparse process. Let c_n be a sequence of natural numbers such that $|G_n| \leq c_n$ and assume that the remaining assumptions of Theorem 2.5 hold. If DecoR is run with Torrent and the sequence $a_n = n - c_n$ of threshold parameters, then*

$$\left\| \hat{d}^{\psi,n}_{DecoR} - d \right\|_2 \in \mathcal{O}_{\mathbb{P}} \left(\sigma_\eta \left(\sqrt{\frac{c_n \log(n/c_n)}{n}} + \sqrt{\frac{L(n)}{n}} \right) + \sqrt{n} g(L(n)) \right).$$

The statistics is identical in the two settings, the main difference lies in the interpretation when considering applications to data. The intuition behind the two settings can be summarized as follows:

longer observational time horizon	$\longrightarrow$	can remove confounding in lower frequencies, sparsity in the low frequency domain needed
shorter time intervals	$\longrightarrow$	can remove confounding in higher frequencies, sparsity in the high frequency domain needed.

Chapter 4: Regularized Torrent with Applications to DecoR

In this chapter, we introduce a regularized version of Torrent. The main motivation behind regularized Torrent is to reduce the influence of the choice of the number of basis functions $L(n)$ by using a smoothness penalty on the function f. Yet, this comes with the challenge of choosing a regularization parameter λ in the presence of outliers. In addition, we discuss the construction of confidence intervals.

4.1 Regularized Torrent

We introduce a regularized version of Torrent from Bhatia et al. (2015). Consider the linear adversarial outlier problem of Setting 2 and let $Q \in \mathbb{R}^{d\times d}$ be a positive semi-definite matrix. We include regularization by replacing the ordinary least squares estimator with the Ridge estimator given by $\hat{\beta} = \arg\min_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_Q^2$ where $\|\beta\|_Q^2 = \beta^\top Q\beta$ and $\lambda \geq 0$ (see Hastie et al. (2017) for an introduction). If Q has full rank and $\lambda > 0$, the solution is always unique and given by $\hat{\beta} = (X^\top X + \lambda Q)^{-1}X^\top Y$.

Algorithm 4 Regularized Torrent

Require: $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^n$, $a \in \{1,\dots,n\}$, $\lambda > 0$, $Q \succeq 0$

```
S_0 ← {1,...,n}, e ← Y, err ← ∞, t ← 0
while ‖e_{S_t}‖_2 + λ‖β̂^t‖_Q < err do
    t ← t + 1
    err ← ‖e_{S_{t-1}}‖_2 + λ‖β̂^{t-1}‖_Q
    β̂^t_Tor ← (X^⊤_{S_{t-1}} X_{S_{t-1}} + λQ)^{-1} X^⊤_{S_{t-1}} Y_{S_{t-1}}
    e ← |Y − X β̂^t_Tor|
    S_t ← HT(e, a)
end while
return β̂^{n,a}_Tor := β̂^t_Tor
```

Recall that the hard holding is given by $HT(e,a) := \{s_1,\dots,s_a\}$ with the sequence $\{s_i\}_{1\leq i\leq a}$ the permutation such that the residuals e_i are ordered increasingly, i.e. $e_{s_1} \leq e_{s_2} \leq \cdots \leq e_{s_a}$. The main changes are the replacement of the ordinary least square estimator by the Ridge estimator and the use of ℓ_2-error plus the norm of β as stopping criteria. The second adjustment is important to guarantee that the results that hold for Torrent carry over to the regularized version. For example, the following lemma guarantees that, like the non-regularized Torrent, the regularized Torrent will terminate after finitely many steps.

Lemma 4.1: *For all $n \in \mathbb{N}$, there exists a constant $N \in \mathbb{N}$ such that Algorithm 4 terminates in less than N steps.*

Similarly, the next theorem gives a non-asymptotic bound on the error in the 2-norm of the estimated coefficients $\hat{\beta}^{n,a}_{Tor}$ obtained by running the regularized version of Torrent.

Theorem 4.2: *Assume Setting 2 and that there exists a known $c_n \in \mathbb{N}$ such that $|G_n| \leq c_n$. Let S_t be the estimated subset in the final iteration of the regularized Torrent executed*

on the data X, Y using the threshold parameter $a_n = n - c_n$. Furthermore, assume that

$$\eta := \max_{S\subseteq\{1,\dots,n\} \text{ s.t. } |S|=a_n} \frac{\|X_{V(S)}\|_2}{\sqrt{\lambda_{min}\left((X_S^\top X_S + \lambda Q\right)}} < \frac{1}{\sqrt{2}}. \tag{11}$$

Then,

$$\left\|\hat{\beta}^t_{Tor} - \beta\right\|_2 \leq \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1}\left(X_{S_t}^\top \epsilon_{S_t} - \lambda Q\beta\right)\right\|_2 + \frac{\sqrt{2}\,\|X_{V_t}\|_2 \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1}\left(X_{S_t}^\top \epsilon_{S_t} - \lambda Q\beta\right)\right\|_2 + \sqrt{2}\left\|\epsilon_{V(S_t)}\right\|_2}{\sqrt{\lambda_{min}\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)}(1-\sqrt{2}\eta)}.$$

Note that we can always fulfill Equation (11) by choosing λ large enough, in contrast to the condition of the standard Torrent. However, this comes with the price of a weaker bound on the error. Observe that we can recover Theorem 2.3 by setting $\lambda = 0$. For $\lambda \neq 0$ we get two additional terms containing $\lambda Q\beta$ and an λQ in the denominator of the second summand. We extend the definition of γ_{a_n} by introducing

$$\gamma^\lambda_{a_n}(X, Q) := \min_{S\subseteq\{1,\dots,n\} \text{ s.t. } |S|=a_n} \sqrt{\lambda_{\min}\left(X_S^\top X_S + \lambda Q\right)}.$$

This allows us to derive an upper bound for the error in a sub-Gaussian noise setting with a bias r. We include some bias since it can occur when using DecoR due to the approximation of functions by a finite linear combination of the first basis functions.

Corollary 4.3: *Assume the setting of Theorem 4.2 with noise $r_i + \epsilon_i$ where ϵ_i are i.i.d. zero-mean sub-Gaussian noise with variance proxy σ^2, a bias $r_i \in \mathbb{R}$ and let $c_n < n/2$. Then, there exists a constant $K > 0$ such that for all $\delta > 0$ with probability at least $1 - \delta$*

$$\left\|\hat{\beta}^{n,a_n}_{Tor} - \beta\right\|_2 \leq \frac{1}{\gamma^\lambda_{a_n}}\left(1 + \frac{\sqrt{2}}{1-\sqrt{2}\eta}\right)\left(\sigma\left(\sqrt{d} + \sqrt{2c_n K \log(2en/c_n\delta)}\right) + \|r_{S_t}\|_2\right) + \frac{2\sigma\sqrt{c_n}\left(1 + \sqrt{K\log(2en/c_n\delta)}\right) + \|r_{V_t}\|_2}{\gamma^\lambda_{a_n}(1-\sqrt{2}\eta)} + \lambda\frac{\|Q\beta\|_2}{\gamma^\lambda_{a_n}}\left(1 + \frac{1}{1-\sqrt{2}\eta}\right).$$

Note that the bound depends on the dimension d of X and therefore does not give a suitable bound in high-dimensional settings. High-dimensional robust regression with outliers is challenging and usually makes some sparsity assumptions on β or uses projections onto lower-dimensional subspaces; see Filzmoser and Nordhausen (2021) for a survey. However, the Ridge penalty will allow us to incorporate smoothness assumptions when using it with DecoR, yielding a vanishing error with increasing sample size n in a simulation study in Section 5.3.

4.2 Ridge-DecoR and Smoothing

In the following, we show that DecoR yields a consistent estimator when we choose the regularized version of Torrent for the robust regression algorithm $\mathcal{A}$; see Algorithm 4. To this end, suppose that we are in Setting 1. The regularized version of Torrent builds on

the Ridge regression and thus we need to choose a regularization parameter $\lambda \geq 0$ and a positive semi-definite matrix $Q \in \mathbb{R}^{L \times L}$ for the penalty given by $\lambda \|d\|_Q = \lambda \cdot d^\top Q d$ which is added to the square loss. The choice of Q is discussed at the end of this section. Assume that Assumption 1 is satisfied and that the following assumption holds.

Assumption 2′ : *Assume that $Q_n \in \mathbb{R}^{L_n \times L_n}$ is a sequence of positive semi-definite matrices and $\lambda_n \in \mathbb{R}_{\geq 0}$ a sequence of non-negative real numbers. We assume a modified version of Assumption 2 where we replace Equation (6) by*

$$\lambda_{min}\left((P^n_{\psi,\phi})^\top_{S''}(P^n_{\psi,\phi})_{S''} + \lambda_n Q_n\right) \geq c'. \tag{12}$$

This differs from Assumption 2 in that we assume a weaker eigenvalue condition since we are adding a positive semi-definite matrix. In particular, Equation (12) can always be satisfied by choosing λ_n large enough, although this comes with a weaker bound on the error.

Theorem 4.4 (Ridge-DecoR)**:** *Let $\{\psi_k\}_{k\in\mathbb{N}}$ be a basis of $L^2([a,b]^d)$ such that for $d_\ell = \langle f, \psi_\ell\rangle$ it holds that $\sum_{\ell \geq L} |d_\ell| \in \mathcal{O}(g(L))$. Let $Q_n \in \mathbb{R}^{L_n \times L_n}$ be a sequence of regularization matrices such that for all $n \in \mathbb{N}$, $\|d_{\{1,\dots,L_n\}}\|_{Q_n} < C$ for some $C > 0$. Let c_n be a known sequence of natural numbers such that $|G_n| \leq c_n$ and assume that Assumption 2′ is satisfied. Suppose that for all $\delta > 0$ there exists $\overline{n} \in \mathbb{N}$ such that for all $n \geq \overline{n}$ it holds that*

$$\mathbb{P}\left[\max_{S\subseteq\{1,\dots,n\}\ s.t.\ |S|=n-c_n} \frac{\left\|(P^n_{\psi,\phi})^\top_{V(S)}\right\|_2}{\sqrt{\lambda_{min}\left((P^n_{\psi,\phi})^\top_S (P^n_{\psi,\phi})_S + \lambda_n Q_n\right)}} < \frac{1}{\sqrt{2}}\right] \geq 1-\delta. \tag{13}$$

If DecoR is run with regularized Torrent and the sequence $a_n = n - c_n$ of threshold parameters, then

$$\left\|\hat{d}^{\psi,n}_{DecoR} - d\right\|_2 \in \mathcal{O}_{\mathbb{P}}\left(\sigma_\eta\left(\sqrt{\frac{c_n \log(n/c_n)}{n}} + \sqrt{\frac{L(n)}{n}}\right) + \sqrt{n} g(L(n)) + \lambda_n\right).$$

The proof of Theorem 4.4 can be found in Appendix B.3. Let $Q \in \mathbb{R}^{\infty\times\infty}$ be such that $\|d\|_Q \leq C$ for some $C > 0$. Choosing $Q_n = Q_{\{1,\dots L_n\}^2}$ and the regularization parameters such that $\lim_{n\to\infty} \lambda_n = 0$ yields consistency in the same cases as discussed in the nonlinear extension in Theorem 3.2. Similarly to when using the not regularized Torrent, if we assume that f is continuously differentiable on a one-dimensional domain and we choose the Fourier basis, the conditions of Proposition 3.1 and Theorem 4.4 are fulfilled, and we obtain:

Corollary 4.5: *Under the assumptions of Theorem 4.4 and Lemma 3.3, choosing $L(n) \in \Theta(n^{1/2})$ yields*

$$\left\|\hat{f}^{\psi,n}_{DecoR} - f\right\|_2 \in \mathcal{O}_{\mathbb{P}}\left(\sigma_\eta\left(\sqrt{\frac{c_n \log(n/c_n)}{n}} + \frac{1}{n^{1/4}}\right) + \frac{1}{n^{1/4}} + \lambda_n\right).$$

The optimal sequence λ^{opt}_n yielding the best convergence rate depends on the matrix $P^n_{\psi,\phi}$ and is given by

$$\lambda^{opt}_n = \min\left\{\lambda > 0 : \mathbb{P}\left[\max_{\substack{S\subseteq\{1,\dots,n\}\\ \text{s.t. } |S|=n-c_n}} \frac{\sqrt{2}\left\|(P^n_{\psi,\phi})^\top_{V(S)}\right\|_2}{\sqrt{\lambda_{\min}\left((P^n_{\psi,\phi})^\top_S (P^n_{\psi,\phi})_S + \lambda Q_n\right)}} < 1\right] \geq 1-\delta\right\}$$

This is not practical since we need to compute the eigenvalues over a large number of sets and we don't know the set of outliers G_n needed to determine $V(S) = S\Delta G_n^c$. Moreover, the assumption is too strong, as only a few of the sets in $\mathcal{U}_n$ are visited by Torrent. Thus, in practice, we suggest choosing λ using a bootstrap method as proposed in Section 4.3.

Summarized, the theory extends well to the regularized Torrent and weakens some of the assumptions, but does not improve the guarantees for the convergence rates. However, using regularization allows us to take into account the smoothness of the underlying truth f, making the choice of L less important as a simulation study in Section 5.3 shows. To avoid overfitting it is common to use a smoothness penalty (e.g. see Hastie et al. (2017)) of the form

$$\lambda \int_a^b f''(x)^2 dx.$$

If $\{\psi_k\}_{k\in\mathbb{N}}$ is the Fourier or cosine basis, the smoothness penalty transforms nicely since for every one-dimensional function $f \in C^2([a,b])$ $f''(x) = \sum_{\ell=0}^{\infty} -\ell^2 d_\ell \psi_\ell(x)$ holds. Thus, using the orthogonality, see Proposition A.3, we obtain

$$\lambda \int_a^b \left(f''(x)\right)^2 dx = \lambda \sum_{\ell=0}^{\infty} \ell^4 d_\ell.$$

When running the regularized Torrent, this penalty corresponds to using for $Q_L \in \mathbb{R}^{L\times L}$ the positive semi-definite diagonal matrix

$$Q_{L+1} = \begin{pmatrix} 0 & & & \\ & 1^4 & & \\ & & \ddots & \\ & & & L^4 \end{pmatrix}.$$

The smoothness penalty can improve the L^2-error and reduce the influence of the choice of L, the number of coefficients. This is shown by a simulation study in Chapter 5, take a look at Figure 5.5 for an example.

4.3 Fine Tuning with Outliers

In practice, we need a rule to choose the regularization parameter λ. Usually, regularization parameters are chosen by performing some kind of cross-validation over a set $\Lambda = \{\lambda_1, \dots, \lambda_m\}$ of $m \in \mathbb{N}$ regularization parameters. However, in the presence of outliers, this is not straightforward and imposes two challenges. First, one needs to find an appropriate loss function which reduces the influence of the outliers, and second, the partition should be done in a way such that the outliers are distributed equally among the sets.

We outline the method of out-of-bootstrap generalization error estimation; see Tibshirani and Efron (1993) for a comprehensive introduction. Consider the transformed sample $Z_{\psi,\phi}^n = (P_{\psi,\phi}^n, Y_\phi^n)$. We generate $B \in \mathbb{N}$ bootstrap samples. One bootstrap sample is generated by sampling n times with replacement from the original data Z, denoted by $\mathcal{L}^{*(b)} = \{Z_1^*, \dots, Z_n^*\}$ for $b = 1, \dots, B$. For each of the bootstrap samples $\mathcal{L}^{*(b)}$, we compute the regularized Torrent estimator $\hat{f}_\lambda^{*(b)}$ for all $\lambda \in \Lambda$. To estimate the generalization error on the transformed sample, we use the out-of-bootstrap sample $\mathcal{L}_{out}^{*(b)} = \bigcup_{i=1}^n \left\{Z_i : Z_i \notin \mathcal{L}^{*(b)}\right\}$. To overcome the problem that the out-of-bootstrap sample

contains outliers which we don't want to use for testing, we try different loss functions ρ. Denote by $e^{*(b)} \coloneqq |Y_\phi^n - \hat{f}_\lambda^{*(b)}(P_{\psi,\phi}^n)|_{\{i\,:\,Z_i \in \mathcal{L}_{out}^{*(b)}\}}$ the absolute values of the residuals and by $a^{*(b)} \coloneqq \left\lfloor \frac{a_n}{n} \cdot |\mathcal{L}_{out}^{*(b)}| \right\rfloor$ the expected number of outliers in the out-of-bootstrap sample.

Clipping: We clip the residuals at the value of the $\left|\mathcal{L}_{out}^{*(b)}\right| - a^{*(b)}$ largest residual and compute the average. This yields the loss function

$$\rho_{clip}\left(e^{*(b)}; a^{*(b)}\right) = \frac{1}{|\mathcal{L}_{out}^{*(b)}|} \left\| \min\left(e, \max\{e_{HT(e,a^{*(b)})}\}\right) \right\|_1,$$

where the minimum is taken elementwise and the maximum is taken over all entries.

Omitting: Take only the smallest $a^{*(b)}$ residuals and take the average

$$\rho_{omit}\left(e^{*(b)}; a^{*(b)}\right) = \frac{1}{a^{*(b)}} \left\| e_{HT(e,a^{*(b)})} \right\|_1.$$

Median: Take the median of the residuals

$$\rho_{med}\left(e^{*(b)}\right) = \text{median}\left(e_1, \ldots, e_{|\mathcal{L}_{out}^{*(b)}|}\right).$$

We decided to use the ℓ_1-norm here because in the simulation studies in Chapter 5 we are looking at the L^1-loss. Alternatively, one can consider the corresponding loss functions using the ℓ_2-norm. The first two loss functions depend on the parameter a_n chosen for running Torrent, whereas taking the median has the advantage of being independent of a_n. We summarize the proposed out-of-bootstrap generalization error estimation with following algorithm.

Algorithm 5 Out-of-bootstrap estimation of generalization error

Require: $P_{\psi,\phi}^n \in \mathbb{R}^{n \times L_n}$, $Y_\phi^n \in \mathbb{R}^n$, $a \in \{1, \ldots, n\}$, $\Lambda = \{\lambda_1, \ldots \lambda_m\}$, B, loss-function ρ

```
Draw B bootstrap samples L^{*(b)}
for λ ∈ Λ do
    for b = 1, ..., B do
        f̂_λ^{*(b)} ← Tor-reg(λ; a, L^{*(b)})
        e_λ^{*(b)} ← |Y_φ^n − f̂_λ^{*(b)}(P_{ψ,φ}^n)|_{{i : Z_i ∈ L_out^{*(b)}}}
    end for
    err_λ ← (1/B) Σ_{b=1}^{B} ρ(e_λ^{*(b)}; a^{*(b)})
end for
λ_min ← arg min_{λ∈Λ} err_λ
return f̂_DecoR ≔ Tor-reg(P_{ψ,φ}^n, Y_φ^n, λ_min)
```

Using bootstrapping and therefore drawing many samples aims to reduce the influence of a possibly bad bootstrap sample not preserving the fraction of outliers, i.e. having almost only or nearly no outliers in the bootstrap sample. This is an advantage over a cross-validation scheme that relies on only one partition of the data.

We provide an example of a regularization path for the three loss functions discussed above. For this, we choose the function $f(x) = 6\sin(2\pi x)$ on $[0, 1]$ and the noise to be given by centered Gaussian variables with variance $\sigma_\eta^2 = 4$ (for the complete configuration we refer

to Section 5.1). We choose a sample size of $n = 2^8$ and use the first $L = 30$ basis functions for the approximation. The plot below shows the regularization paths of the estimated generalization error using $B = 500$ bootstrap samples of the transformed sample, where the interval presents one standard error. The λ_{min} minimizes the estimated loss and $\lambda_{1-S.E.}$ is the regularization parameter chosen by the one-standard rule; see chapter in 7.10 Hastie et al. (2017). We also plot the L^1-loss of the corresponding $\hat{f}_{DecoR}$, scaled by some constant $c \in \mathbb{R}$ for better visibility.

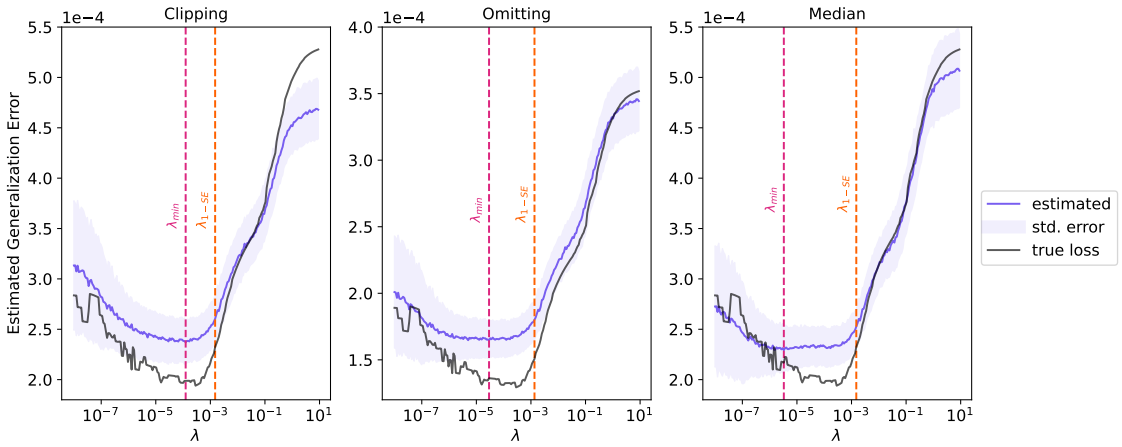


Figure 4.1: Examples of regularization paths of the estimated generalization error on the transformed sample with one standard error.

Note the similarity of the three paths. Looking at L^1-loss of the estimated $\hat{f}_{DecoR}(\lambda)$, all three paths tend to resemble its shape with a nearly constant part for $\lambda \in [10^{-6}, 10^{-3}]$. Due to the challenges in estimating the generalization error, having a wide range of λ's producing good results gives us some leeway when it comes to choosing a good regularization parameter λ. However, one possible underlying problem could be the fact that the true loss suffers from local fluctuations, i.e. the many changes from being decreasing to increasing and vice versa. We refer to Section 5.3 for a simulation study investigating the performance of the different methods.

4.4 Confidence Intervals

Due to the confounding and therefore the presence of the outliers after our sample transformation, it is hard if not impossible to compute the distribution of our estimator obtained by DecoR. Unfortunately, the distribution is needed to obtain a confidence interval. Nevertheless, we are going to suggest an approximation of a confidence interval: Assume that the noise η_t is independent and identically normal distributed with mean 0. To estimate the variance, we compute the residuals for **all** our observations of the transformed sample

$$r = Y_\phi^n - P_{\psi,\phi}^n \hat{d}_{DecoR}^{\psi,n}$$

and use them to get an estimation of the variance

$$\hat{\sigma}^2 = \frac{1}{n-L} \sum_{i=1}^{n} r_i^2.$$

Only using the observations in S_t, the estimated inliers of DecoR, would lead to an underestimation, since DecoR minimizes the residuals. Although using all observations is likely to give an overestimation of the variance, the hope lies in a compensation of the bias from which our estimator $\hat{f}_{DecoR}$ suffers. Suppose we want to estimate f at the points $x_1, \ldots, x_m$, then the estimation obtained by DecoR can be written as a linear estimator

$$\hat{f}_{DecoR}(x_1, \ldots, x_m) = HY_{S_t},$$

with the hat matrix H given by

$$H = \Psi_L(x_1, \ldots, x_m) \left((P^n_{\psi,\phi})^\top_{S_t} (P^n_{\psi,\phi})_{S_t} \right)^{-1} (P^n_{\psi,\phi})^\top_{S_t}.$$

Here, $\Psi_L(x_1, \ldots, x_m) \in \mathbb{R}^{m \times L}$ is the matrix with entries $(\psi_\ell(x_i))_{i,\ell}$. Since the noise is identically, independent normally distributed it follows that

$$\hat{f}_{DecoR}(x_i) \sim \mathcal{N}\left(\mathbb{E}[\hat{f}_{DecoR}(x_i)], \sigma^2 \left(HH^\top \right)_{ii} \right)$$

Using our estimator $\hat{\sigma}^2$ of the variance from above, $\hat{f}_{DecoR}(x_i)$ is approximately t-distributed with $df = a_n - L$ degrees of freedom. Thus, we get the approximated "confidence interval"

$$\left[\hat{f}_{DecoR}(x_i) - q_t \left(\frac{1-\alpha}{2} \right) \widehat{s.e.}_i,\ \hat{f}_{DecoR}(x_i) + q_t \left(\frac{1-\alpha}{2} \right) \widehat{s.e.}_i \right]$$

with $\widehat{s.e.}_i = \hat{\sigma} \sqrt{(HH^\top)_{ii}}$ for $i = 1, \ldots, m$, following the standard procedure for constructing a confidence interval for the linear model (Faraway, 2015, Chapter 4.1). In Section 5.4 we perform a small simulation study which indicates that this approximation of the confidence intervals leads to a larger coverage than $1 - \alpha$. Thus, the proposed approximation appears to provide an upper bound.

Chapter 5: Simulations

This chapter studies the performance of DecoR on synthetic data. After a case study, we investigate the convergence of the nonlinear extensions, the performance of the regularized version, and the coverage of confidence intervals obtained by using different techniques.

5.1 A Visual Example

In this section, we present a visual example to explain and provide some intuition for the mechanisms of DecoR. We consider the underlying truth to be given by the function $f(x) = 6\sin(2\pi x)$ on the interval $[0, 1]$. We take $X_t \overset{i.i.d.}{\sim} \mathcal{U}([0, 1])$ and let U_t be the projection of $60X_t - 30$ onto a randomly selected 25% of the discrete cosine basis functions. We run DecoR with Torrent without regularization and the threshold parameter $a_n = 0.7n$. As a benchmark, we fit a smoothing spline provided by the pyGAM package and use cross-validation to select the penalization parameter. In Figure 5.1 we show an example with a sample size of 2^8 using the first $L = 4$ coefficients of the cosine basis for the approximation with DecoR. We plot the two estimators together with the 0.95-confidence intervals, where the one for DecoR is approximated (see Section 4.4).

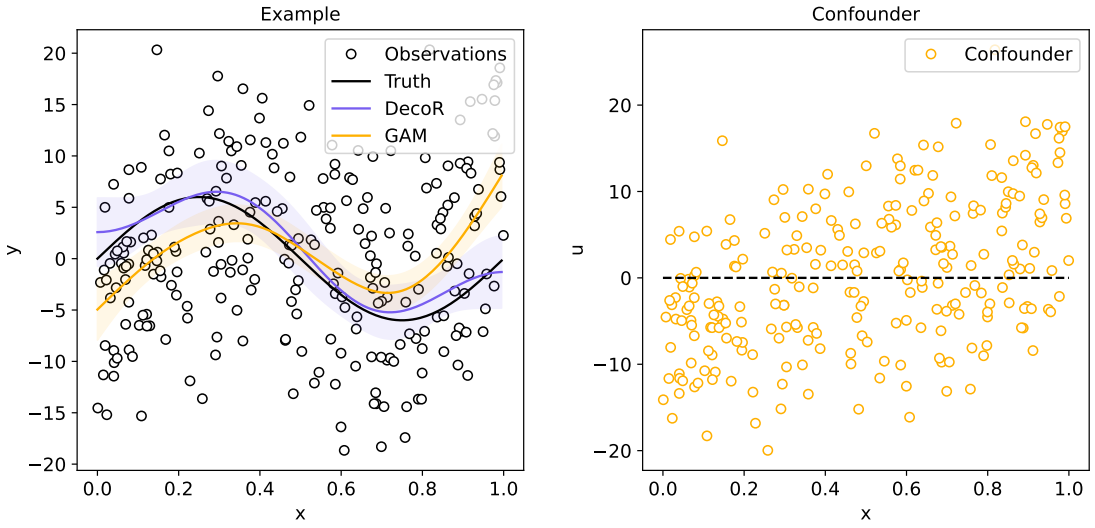


Figure 5.1: Example with underlying truth $f(x) = 6\sin(2\pi x)$. Standard nonparamteric regression methods like smoothing splines suffer from bias introduced by the confounder U_t, whereas DecoR is robust against sparse confounding in the frequency domain.

Observe that the smoothing spline tends to pick up the confounding where there is a local bias, particularly in the neighborhoods around 0 and 1. This contrasts DecoR which is less affected by the confounder plotted in the right figure of Figure 5.1.

To see how Torrent removes confounded frequencies, we show the transformed sample and the linear approximation fitted by DecoR in Figure 5.2. The purple observations are the outliers found by Torrent and the orange ones are the outliers that Torrent has missed. We see that Torrent detects large outliers relatively well, but has difficulties detecting outliers

that do not deviate much from the ground truth. However, these outliers do not have a large influence on the final fit.

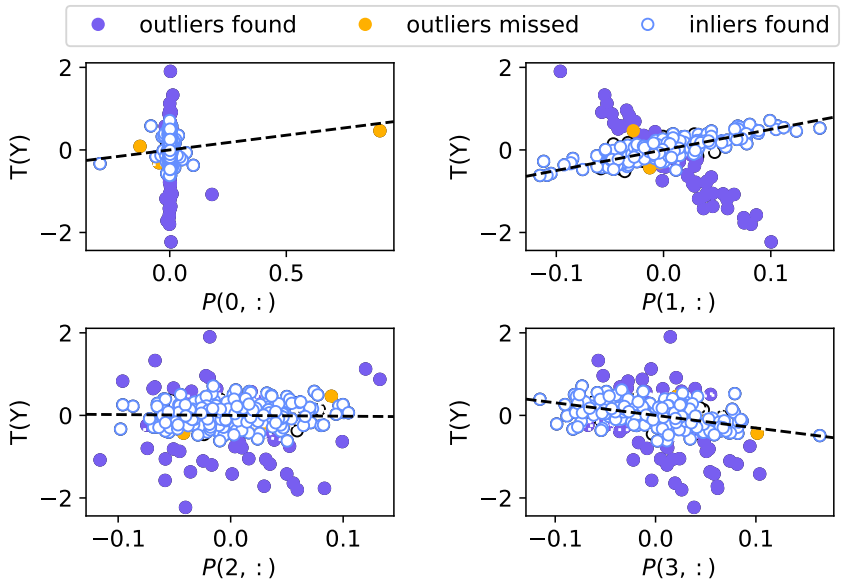


Figure 5.2: Outlier detection by Torrent: Large deviations have a good chance of being detected, whereas smaller outliers are more difficult to find.

After having studied an example, we analyze the consistency of the DecoR estimator next using a Monte Carlo simulation.

5.2 Analyzing Convergence

We perform a simulation study to investigate the L^1-consistency of DecoR on synthetic data. We look at two settings, $X_t \overset{i.i.d}{\sim} \mathcal{U}([0,1])$ uniformly distributed and X_t being an Ornstein-Uhlenbeck process reflected on the boundaries of the interval $[0,1]$. For U_t we take the projection on 25% of the cosine basis functions selected randomly, thus $|G_n| = 0.25n$. The noise $\eta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\eta^2)$ is normally distributed with variance $\sigma_\eta^2 \in \{0, 1, 4\}$. We run the nonlinear extension of DecoR with Torrent and again the threshold parameters ${}_n a = 0.7n$. For the underlying truth, we take the function $f(x) = 6\sin(2\pi x)$ which has no finite expansion w.r.t. orthonormal basis $\psi_\ell(x) = \cos(\ell\pi x)$, $\ell \in \mathbb{N}$. We chose the number of basis functions such that $L(n) \in \Theta(n^{1/2})$, the optimal rate for twice continuously differentiable functions as seen in Corollary 3.4. The accuracy is measured using the L^1-error given by $\int_0^1 |\hat{f}^n(x) - f(x)|dx$.

Figure 5.3 shows that the L^1-error vanishes for DecoR as the number of data points n increases, with the variance σ_η^2 of the noise influencing the speed of convergence. The larger the noise the slower the convergence. This is in contrast to smoothing splines (GAM) with the regularization parameter chosen by cross-validation for which the L^1-error does not converge to zero. Moreover, DecoR is consistent in more cases than Theorem 3.2 suggests as the fraction of confounded frequencies c_n/n does not converge to zero in our simulations.

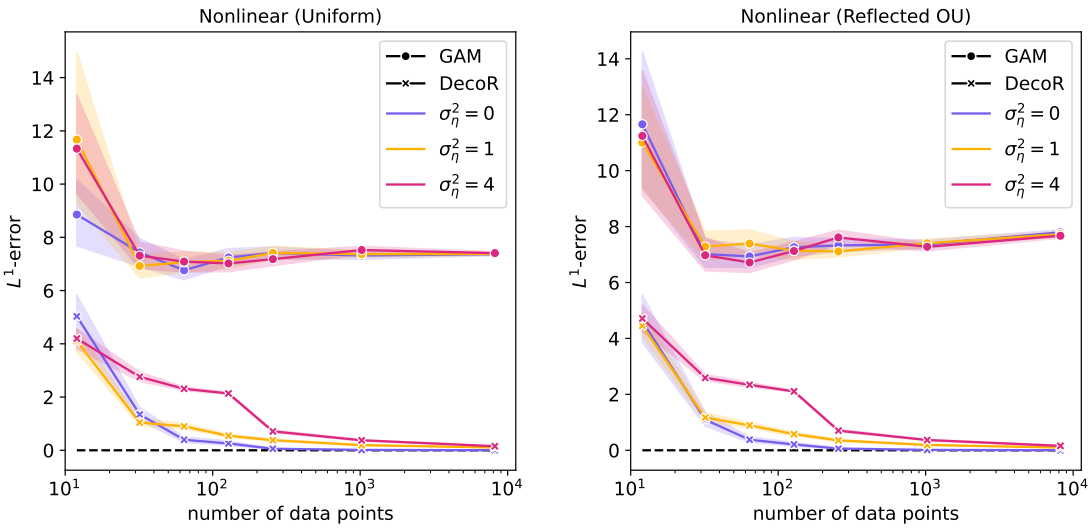


Figure 5.3: The L^1-error for DecoR vanishes as the sample size n increases for the nonlinear function $f(x) = 6\sin(2\pi x)$, whereas the smoothing spline tends to stay constant. The bigger the variance σ_η^2 of the noise, the slower the convergence.

Furthermore, we obtain similar results for the sigmoid function $f(x) = \frac{20}{1+\exp(-16x+6)} - 10$, indicating that the results are robust with respect to the choice of the true underlying function, see Appendix C.1.

5.3 Regularization and Fine-Tuning

In this section, we investigate the application of the regularized version of Torrent to DecoR using the smoothness penalty. We focus on understanding the effect of the number of basis functions L and the regularization strength λ on the estimation error. We choose the same configuration as in Section 5.2 but with X_t independent, identical uniformly distributed on $[0, 1]$. In Figure 5.5 we plot the L^1-estimation error for $L \in [1, 60]$ and $\lambda \in [10^{-8}, 10]$. The expected L^1-error was estimated using a Monte Carlo simulation with 200 repetitions. In Figure 5.5 we see that the estimation error yields an L-shaped valley. The magnitude of the effect of the regularization parameter depends on the sample size n and the variance σ_η^2 of the noise. If there is no noise and a large enough sample, no regularization is needed (bottom left). But as soon as some noise is added or the sample size is small, regularization can reduce the L^1-error. Choosing a suitable regularization parameter λ seems to be essential. For example, choosing the regularization $\lambda = 10^{-4}$ in the upper left situation ($n = 2^6$ and $\sigma_\eta^2 = 0$), we can choose any number L of base functions large enough, $L > 5$, to produce the same expected error.

Having seen that an appropriate λ can work for a wide range of L, we fix $L = 50$ and test the different fine-tuning methods presented in Section 4.3. That is, we use the out-of-bootstrap generalization error estimation method with three different loss functions based on the median, clipping, and omitting the largest residuals. For these loss functions, we test the performance when choosing λ_{min}, minimizing the estimated generalization loss, and $\lambda_{1-S.E.}$ using the one-standard error rule. The regularized methods are compared to DecoR using Torrent without any regularization, but $L(n) = \frac{1}{2}\sqrt{n}$. In Table 5.1 and

Table 5.2 we show the expected relative error given by

$$\text{rel. error} := \frac{\left\| \hat{f}_{reg} - f \right\|_{L^1}}{\left\| \hat{f}_{Tor} - f \right\|_{L^1}}.$$

method \ sample size n		32	64	128	256	1024	8192
L^1-error	Torrent	2.595	0.848	0.462	0.160	0.045	0.007
rel. error (regularized)	Clip	0.441	0.602	0.345	0.068	0.134	0.775
	Omit	0.549	0.678	0.327	0.061	0.135	0.775
	Median	0.415	0.687	0.331	0.072	0.134	0.775
	Clip 1-S.E.	0.365	0.660	0.291	0.076	0.132	0.756
	Omit 1-S.E.	0.486	0.643	0.339	0.154	0.132	0.757
	Median 1-S.E.	0.340	0.649	0.335	0.165	0.131	0.757

Table 5.1: Relative error when there is no noise $\sigma_\eta^2 = 0$ for using DecoR with the regularized Torrent ($L_n = 50$) compared to no regularization ($L_n \in \Theta(\sqrt{n})$). Regularization leads to an improvement when choosing λ with any of the tested bootstrap methods.

method \ sample size n		32	64	128	256	1024	8192
L^1-error	Torrent	2.672	1.070	0.658	0.342	0.208	0.105
rel. error (regularized)	Clip	0.518	0.792	0.979	1.020	0.985	0.838
	Omit	0.661	0.797	1.013	1.238	1.161	0.880
	Median	0.482	0.797	1.012	1.333	1.135	0.896
	Clip 1-S.E.	0.525	0.809	0.969	1.055	0.966	0.827
	Omit 1-S.E.	0.684	0.790	0.996	1.204	1.139	0.866
	Median 1-S.E.	0.499	0.789	0.994	1.293	1.133	0.883

Table 5.2: Relative error for $\sigma_\eta^2 = 1$. Using the one-standard error rule tends to lead better results than taking $\lambda_{\min}$ minimizing the estimated error.

If there is no noise, $\sigma_\eta^2 = 0$, choosing the regularization parameter λ with any of the six methods performs better than using no regularization and taking $L_n = \frac{1}{2}\sqrt{n}$. As soon as there is some noise, see Table 5.2, the bootstrapping methods do not always lead to better results. Nevertheless, in many cases there is an improvement, where clipping the residuals together with the one-standard deviation rule performs best.

5.4 Coverage of Confidence Intervals

We test the approximation of the confidence intervals discussed in Section 4.4. To this end, we choose the setting of Section 5.2 with X_t i.i.d. uniformly on $[0, 1]$, centered Gaussian noise with variance $\sigma_\eta^2 = 1$ and a sample size of 2^{10}. We draw 200 samples and test if the underlying true f lies in the "confidence interval", plotted as the estimated actual coverage in Figure 5.4. This is performed for $x \in \{0.1, 0.5, 0.9\}$ and the nominal coverage $\alpha \in [0.7, 0.99]$.

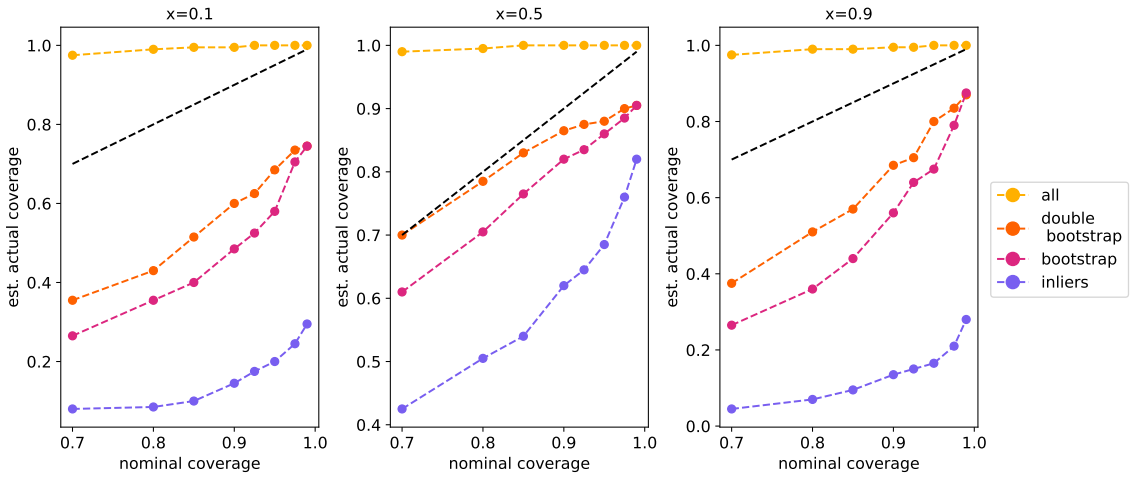
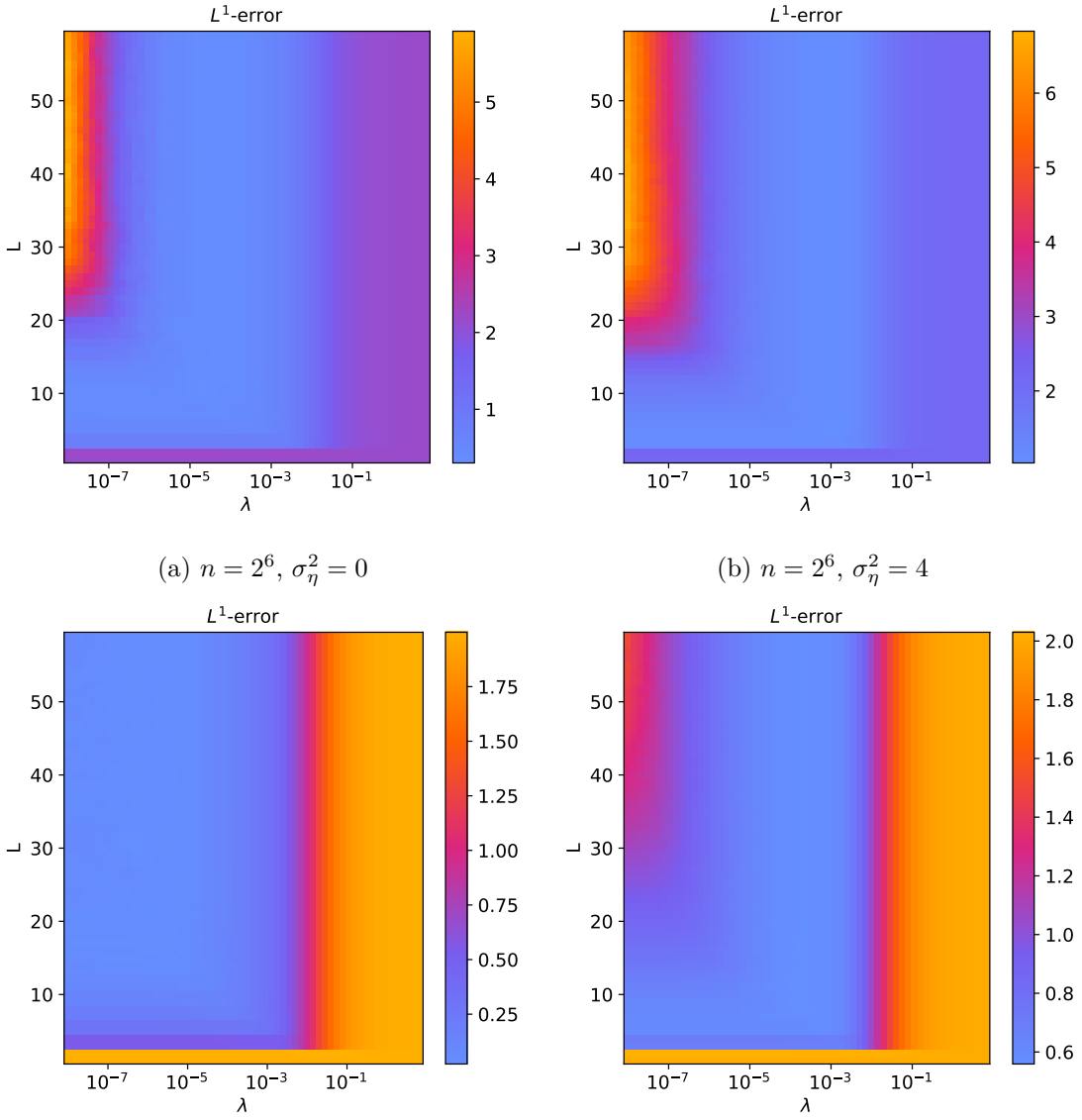


Figure 5.4: Actual coverage of the confidence intervals vs. the nominal coverage. Taking all the observations of the transformed sample to estimate the variance leads to a too large coverage, whereas taking only the estimated inliers of DecoR yields a coverage which is too small.

The simulation study shows the following. Taking all the observations of the transformed sample for estimating the variance leads to a confidence interval which is too large. Even for $\alpha = 0.7$, the actual coverage turns out to be around 1. This stands in contrast to taking only the inliers S_t estimated by DecoR which produces a coverage that is too low. The third option tested, bootstrapping, also leads to a too low coverage. Here, we used $B = 200$ random draws with replacements for the construction of the bootstrap confidence interval. The too low coverage can be corrected to some extent by using double bootstrapping, where for $x = 0.5$ we almost get the intended level of coverage. The performance of the double bootstrapping seems to depend on the underlying truth f: Take a look at f being the sigmoid function in Appendix C where double bootstrapping leads to almost the nominal coverage for all three points. In conclusion, including all the observations yields a conservative upper bound, and considering only the estimated inliers yields a lower bound, where the confidence interval with the initial coverage lies somewhere in between. Closest to the intended coverage comes double bootstrapping, but has the risk of having too low coverage, particularly at the end points.



(a) $n = 2^6$, $\sigma_\eta^2 = 0$

(b) $n = 2^6$, $\sigma_\eta^2 = 4$

(c) $n = 2^8$, $\sigma_\eta^2 = 0$

(d) $n = 2^8$, $\sigma_\eta^2 = 4$

Figure 5.5: Expected L^1-error for different regularization parameters λ and number of coefficients L using the smoothness penalty. Observe that regularization can improve the error in particular in settings with a small sample size and a large variance of the noise. Choosing λ appropriate reduces the influence of the choice of the number of basis functions L.

Chapter 6: Application to Environmental Epidemiology

We now apply the nonlinear extension of DecoR to a practical problem in environmental epidemiology, specifically we analyze the causal influence of ozone levels on health outcomes. This case study is inspired by Bhaskaran et al. (2013) and uses the same dataset to explore how daily mean ozone levels impact the number of deaths in London. Elevated ozone levels are known to have significant short-term effects on human health, particularly on the respiratory and cardiovascular system (Nuvolone et al. (2018)). We consider two time series: the daily mean ozone level X_t as covariate and the number of deaths Y_t in London as the target variable. Looking at the observations in Figure 6.1 reveals a pronounced seasonal periodicity in both variables, suggesting a potential confounding by seasonal factors.

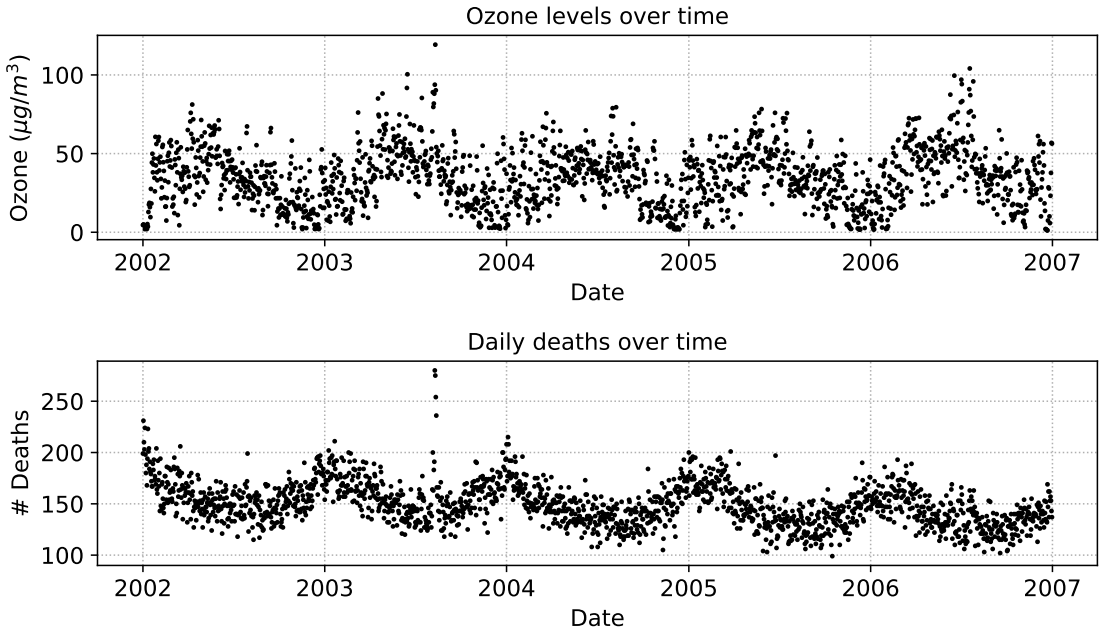
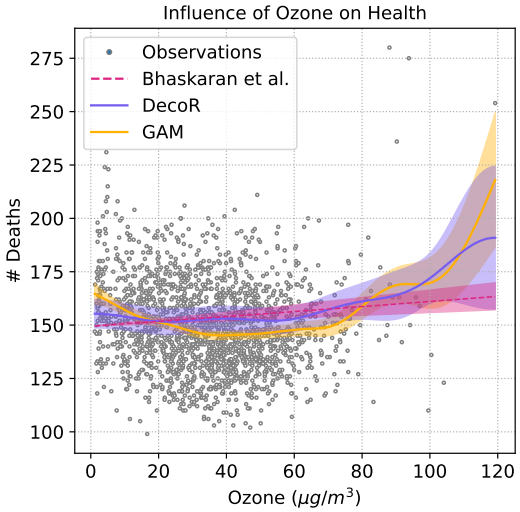


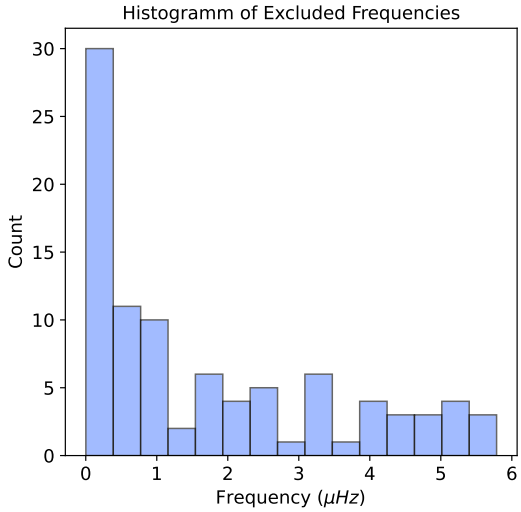
Figure 6.1: Daily mean ozone levels and number of deaths in London. Both time series exhibit a strong seasonal periodicity, indicating a seasonal confounding.

Following Bhaskaran et al. (2013) we allow a delay of one day between the level of ozone X_t and the health outcome Y_t. The function $f : X_{t-1} \to Y_t$ is approximated by the first 6 cosine basis functions plus an intercept. Since high levels of ozone often occur together with heatwaves which are non-periodic, we add the daily mean temperature as a second explanatory variable and model the two effects additively. There is no need to include standard confounding variables like age, sex, body mass index, smoking status or drinking since we only consider effects on the whole population and their distributions are very likely to not change over the considered time periods. Assuming that the seasonal confounding is sparse with respect to the discrete cosine basis, we run DecoR with $a = 0.95n$, that is, we allow for a confounding of 5% of the frequencies. To make the confounding visible, we

also fit a smoothing spline (GAM) to the data. As shown in the left panel of Figure 6.2a, the smoothing spline predicts a higher number of deaths for low ozone levels compared to intermediate levels—a counterintuitive result potentially driven by seasonal confounding, as it is implausible that low ozone levels directly cause higher mortality.



(6.2a) The smoothing spline clearly suffers from confounding, whereas DecoR is more plausible and is closer the estimator of Bhaskaran et al. (2013)



(6.2b) Torrent removes the lower frequencies which correspond to the expected seasonal confounding.

Comparing DecoR with the GAM estimator, we see that DecoR does not show this counterintuitive behavior, as it removes confounded low frequencies (see Figure 6.2b). Moreover, for low ozone values we get a nearly constant estimator, starting to increase in a nonlinear fashion around $60 \mu g/m^3$. Looking at the 95% confidence interval, this effect must be interpreted with caution. In contrast to Bhaskaran et al. (2013), who removed the confounding by fitting a spline and then explained the remaining residuals using Poisson regression, we do not need to assume knowledge of the confounded frequencies as DecoR removes them automatically and we do not need to assume an exponential relationship between ozone levels and the number of deaths.

It is important to emphasize that this experiment is conducted to demonstrate the applicability and effectiveness of DecoR in handling confounding within time series data. Our analysis does not intend to make definitive claims about the causal relationship between ozone levels and death rates. As statisticians, our primary focus lies on the methodological framework, and interpretations related to health outcomes should be undertaken in collaboration with medical and environmental health experts.

Chapter 7: Summary and Outlook

In this thesis, we provided an extension of DecoR for nonlinear causal effects and proved its consistency in L^2 under a sparsity assumption in the frequency domain and a smoothness assumption on the underlying causal effect. We showed that the asymptotics for longer time horizons coincide with the asymptotics for shorter time intervals. Moreover, we derived a regularized version of Torrent and showed that theoretical guarantees generalize. The regularized version can be used to incorporate a smoothness penalty in DecoR. We made several proposals for choosing an appropriate regularization parameter, and suggestions for the construction of confidence intervals, and tested them numerically. Using an example from environmental epidemiology, we demonstrated how the nonlinear extensions of DecoR can be applied in practice.

However, many questions remain open, offering promising directions for future research. While DecoR and its nonlinear extensions can be used with any robust regression algorithm, theoretical guarantees currently exist only when combined with (regularized) Torrent if the covariate is multidimensional. Thus, an interesting question is whether there are other algorithms that yield better convergence rates. Regarding the asymptotic results, another unsolved problem is the adaption of DecoR to nonregular time intervals, in particular, it is unclear how the projection onto the basis has to be discretized in order to preserve sparsity. Another potential avenue is to explore alternative methods to handle nonlinearity, for example kernel-based approaches. With the use of an unordered basis depending on the covariate it is no longer sensible to assume a decay of the coefficients, leading to the necessity of alternative suitable assumptions. An additional challenge is to obtain a better understanding of fine-tuning in the presence of adversarial outliers and develop an underlying theory. Furthermore, an unsolved problem remaining is the construction of confidence intervals with the intended coverage. Although there is not much literature on this matter, a reliable method could have broad applications.

Bibliography

N. Ahmed, T. Natarajan, and K. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

S. Bergner, T. Möller, D. Weiskopf, and D. Muraki. A spectral analysis of function composition and its implications for sampling in direct volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12:1353–60, 2006.

K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, and B. Armstrong. Time series regression studies in environmental epidemiology. *International Journal of Epidemiology*, 42 (4):1187–1195, 2013.

K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. *Advances in Neural Information Processing Systems*, 28, 2015.

P. Čížek and S. Sadıkoğlu. Robust nonparametric regression: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1492, 2020.

T. D'Orsi, G. Novikov, and D. Steurer. Consistent regression when oblivious outliers overwhelm. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 2297–2306. PMLR, 2021.

J. J. Faraway. *Linear Models with R.* Chapman and Hall, 2 edition, 2015.

G. Felder. *Mathematische Methoden der Physik I.* lecture notes, ETH Zürich, 2023.

P. Filzmoser and K. Nordhausen. Robust linear regression for high-dimensional data: An overview. *WIREs Computational Statistics*, 13(4):e1524, 2021.

L. Grafakos. *Classical Fourier Analysis.* Springer, 2008.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2017.

D. H.-Y. Leung. Cross-validation in nonparametric regression with outliers. *The Annals of Statistics*, 33(5):2291–2310, 2005.

M. D. Mahecha, M. Reichstein, N. Carvalhais, G. Lasslop, H. Lange, S. I. Seneviratne, R. Vargas, C. Ammann, M. A. Arain, A. Cescatti, I. A. Janssens, M. Migliavacca, L. Montagnani, and A. D. Richardson. Global convergence in the temperature sensitivity of respiration at ecosystem level. *Science*, 329(5993):838–840, 2010.

D. Nuvolone, D. Petri, and F. Voller. The effects of ozone on human health. *Environmental Science and Pollution Research*, 25:8074–8088, 2018.

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

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press, 2017.

W. Popiński. On least squares estimation of fourier coefficients and of the regression function. *Applicationes Mathematicae*, 22(1):91–102, 1993.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

F. Schur and J. Peters. Decor: Deconfounding time series with robust regression. *arXiv preprint*, 2024.

J. Schwartz, C. Spix, G. Touloumi, L. Bacharova, T. Barumamdzadeh, A. Le Tertre, T. Piekarksi, A. P. De Leon, A. Pönkä, G. Rossi, et al. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology & Community Health*, 50(Suppl 1):3–11, 1996.

Y. Shen and S. Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pages 5739–5748. PMLR, 2019.

S. Sippel, N. Meinshausen, A. Merrifield, F. Lehner, A. G. Pendergrass, E. Fischer, and R. Knutti. Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, 32(17):5677 – 5699, 2019.

A. S. Suggala, K. Bhatia, P. Ravikumar, and P. Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

R. J. Tibshirani and B. Efron. An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, 57(1):1–436, 1993.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

World Health Organization and others. WHO global air quality guidelines: particulate matter (pm2. 5 and pm10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, 2021.

Appendix A: Additional Details

A.1 Definitions and Standard Results

We present some additional definitions and well-known results that might be needed in this thesis.

Definition A.1 (Discrete cosine basis, Ahmed et al. (1974))**:** *Let $T \in \mathbb{R}_{>0}$ be fixed. Define for all $k \in \mathbb{N}$ the functions $\phi_k : [0, T] \to \mathbb{R}$ by*

$$t \mapsto \begin{cases} 1 & \text{if } t = 0 \\ \sqrt{2}\cos((k + 1/2)\pi t) & \text{else.} \end{cases}$$

We call $\phi = (\phi_k)_{k\in\mathbb{N}}$ the cosine basis.

The discrete cosine basis preserves the orthogonality and therefore satisfies Assumption 2. To approximate the underlying function $f \in L^2([0, 1])$ we use the following continuous version of the cosine basis since we want to obtain a continuous estimator.

Definition A.2: *The continuous cosine basis $\{\psi_\ell(x)\}_{\ell\in\mathbb{N}}$ that spans $L^2([0, 1])$ is given by $\psi_\ell(x) = \cos(\ell\pi x)$ for $x \in [0, 1]$. Note that this corresponds to the real part of the Fourier basis.*

The following proposition summarizes some of the main results of the theory of orthonormal systems, notably Parseval's identity.

Proposition A.3 (Parseval, e.g. see Felder (2023))**:** *Let H be a Hilbert space, and $\{\phi_j\}_{j\in\mathbb{N}}$ an orthonormal system. The following statements are equivalent:*

(i) $\{\phi_j\}_{j\in\mathbb{N}}$ *is complete*

(ii) $f = \sum_{j=1}^{\infty} \langle f, \phi_j\rangle \, \phi_j, \quad \forall f \in H$

(iii) $\|f\|^2 = \sum_{j=1}^{\infty} |\langle f, \phi_j\rangle|^2, \quad \forall f \in H.$

The next two definitions extend the big O-notation needed for the convergence analysis.

Definition A.4 (Knuth notation)**:** *Let $f, g : \mathbb{N} \to \mathbb{R}$ be real-valued functions. We say f is of the same order as g and write $f \in \Theta(g)$ if $f \in \mathcal{O}(g)$ and $g \in \mathcal{O}(f)$.*

Definition A.5 (Stochastic boundedness)**:** *Let $X_1, X_2, \ldots$ be a sequence of random variables and $g : \mathbb{N} \to \mathbb{R}$ a real-valued function. We say that*

$$X_n \in \mathcal{O}_{\mathbb{P}}(g(n))$$

if for all $\delta > 0$ there exists a real-valued function $f_\delta : \mathbb{N} \to \mathbb{R}$ such that

$$\text{for all } n \in \mathbb{N}: \qquad \mathbb{P}\left[X_n \leq f_\delta(n)\right] \geq 1 - \delta$$

and

$$f_\delta(n) \in \mathcal{O}(g(n)).$$

Next, we give one of many possible definitions of sub-Gaussian random variables. Intuitively, sub-Gaussian are random variables whose tails decay at least as fast as the tails of a Gaussian.

Definition A.6 (sub-Gaussian)**:** *Let X be a random variable. We call X sub-Gaussian if there exists a constant $K > 0$ such that*

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-\frac{t^2}{K^2}\right) \quad \textit{for all } t \geq 0.$$

Structural Causal Models

For the reader who is not familiar with causal inference or wants to refresh his memory, we present some of the fundamental definitions and results which are used in Section 2.2. For a more comprehensive introduction we refer to Peters et al. (2017). We begin by revisiting the definition of a structural causal model.

Definition A.7 (Structural causal models)**:** *A structural causal model (SCM) $\mathfrak{C} := (S, P_N)$ consists of a collection S of d (structural) assignments*

$$X_j := f_j(PA_j, N_j), \quad j = 1, \ldots, d, \tag{14}$$

where $PA_j \subseteq \{X_1, \ldots, X_d\} \setminus \{j\}$ are called parents of X_j, and a joint distribution $P_N = P_{N_1,\ldots,N_d}$ over the noise variables, which we require to be jointly independent. The graph $\mathcal{G}$ of an SCM is obtained by creating one vertex for each X_j and drawing directed edges from each parent in PA_j to X_j.

If the induce graph $\mathcal{G}$ is acyclic, a structural causal model defines a unique distribution over the variables $X = (X_1, \ldots, X_d)$.

Proposition A.8 (Uniqueness, Peters et al. (2017))**:** *Assume that the induced graph $\mathcal{G}$ is acyclic. An SCM $\mathfrak{C}$ defines a unique distribution over the variables $X_1, \ldots, X_d$: any $X_1, \ldots, Xd$, $N_1, \ldots, N_d$ satisfying $X_j = f_j(PA_j, N_j)$ almost surely, where $(N_1, \ldots, N_d)$ has the desired distribution, induce the same distribution over $X = (X_1, \ldots, X_d)$. We refer to it as the entailed distribution $P_X^{\mathfrak{C}}$ and sometimes write P_X.*

With the notion of structural causal models at hand, we can give the definition of an intervention and an introduction to the do calculus. Interventions are modifications of the original SCM, leading to new entailed distributions.

Definition A.9 (Intervention)**:** *Consider an SCM $\mathfrak{C} := (S, P_N)$ and its entailed distribution $P_X^{\mathfrak{C}}$. We replace one (or several) of the structural assignments to obtain a new SCM $\widetilde{\mathfrak{C}}$. Assume that we replace the assignment for X_k by*

$$X_k := \tilde{f}(\widetilde{PA}_k, \widetilde{N}_k).$$

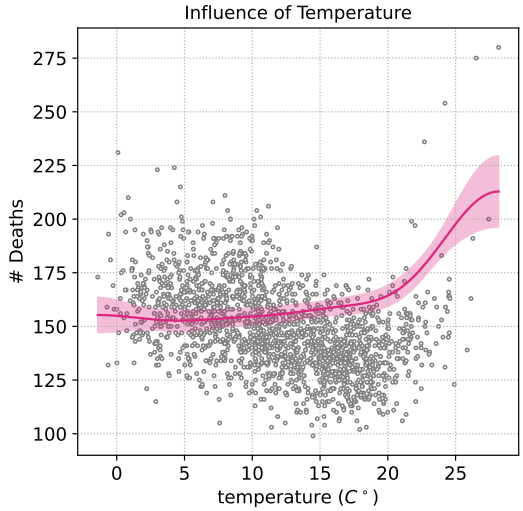
We then call the entailed distribution of the new SCM an intervention distribution and say that the variables whose structural assignment we have replaced have been intervened on. We denote the new distribution by

$$P_X^{\tilde{\mathfrak{C}}} =: P_X^{\mathfrak{C};do(X_k := \tilde{f}(\widetilde{PA}_k, \widetilde{N}_k))}.$$

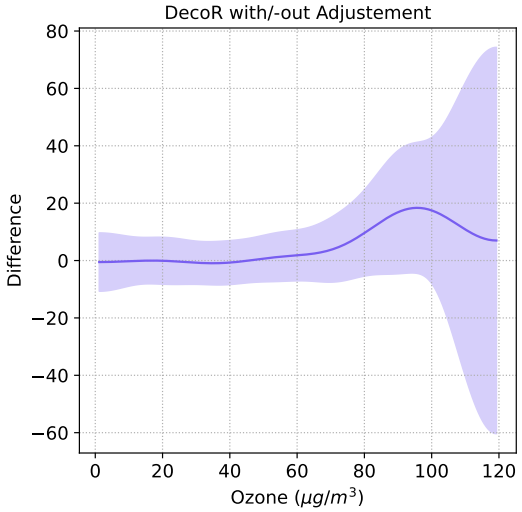
The set of noise variables in $\widetilde{\mathfrak{C}}$ now contains both, some "new" $\widetilde{N}$'s and some "old" N's, all of which are required to be jointly independent. When $\tilde{f}(\widetilde{PA}_k, \widetilde{N}_k)$ puts a point mass on a real value x, we simply write $P_X^{C;do(X_k := x)}$.

A.2 Environmental Epidemiology

We take a short look at the importance of adjustment with respect to temperature in the ozone example. Since heat waves do not need to occur periodically, DecoR can fail to remove all of the confounding effects if we do not include the temperature as a covariate.



(A.1a) Estimation of the influence of the temperature on the number of deaths.



(A.1b) Difference between before and after adjusting for the temperature of the estimation and the confidence intervals ($\Delta = \hat{y}_{\text{not adjusted}} - \hat{y}_{\text{adjusted}}$).

We see that there is a clear difference for high temperatures and, with this, for high ozone levels. Hence, some of the deaths that occur during times with high ozone levels are also related to heatwaves. Interesting in Figure A.1a is that there is no increase for low temperatures, whereas there clearly is a correlation. In Figure A.1b we plot the difference $\Delta = \hat{y}_{\text{not adjusted}} - \hat{y}_{\text{adjusted}}$ between the two estimators. Adjusting for temperature leads to a lower estimation of influence of ozone, in particular, for high ozone levels.

Appendix B: Proofs

B.1 Estimating Causal Effects in Time Series

Lemma 2.2, Theorem 2.3, Corollary 2.4 and Theorem 2.5 follow directly from the more general versions Lemma 4.1, Theorem 4.2, Corollary 4.5 and Theorem 4.4 by setting the regularization parameter $\lambda = 0$ and replacing $P^n_{\psi,\phi}$ with X^n_ϕ. Alternatively, the proofs can also be found in Schur and Peters (2024).

B.2 Nonlinear Extensions of DecoR

Proof of Proposition 3.1. Since for all $\ell \in \mathbb{N}$ the function ψ_ℓ is bounded $\psi_\ell \circ g$ is in $L^2([0,T])$ and since $\sum_{k=0}^{\infty} |d_k| < \infty$ this extends to $f \circ g$. This proves the first statement. For the second, recall that by the definition of an orthonormal basis, we can write the function f for almost all $x \in [a,b]^d$ as

$$f(x) = \sum_{\ell \in \mathbb{N}} d_\ell \psi_\ell(x).$$

Next we are going to consider the transformation of compositions, in our case of $f \circ g$ given by $t \mapsto f(g_t)$. Using the decomposition of f above yields

$$f(t) := f(g_t) = \sum_{\ell \in \mathbb{N}} d_\ell \psi_\ell(g_t)$$

for almost all $t \in [0,T]$. Thus projecting $f \circ g$ onto ϕ_k gives

$$\begin{aligned}
\langle f \circ g, \phi_k \rangle &= \left\langle \sum_{\ell \in \mathbb{N}} d_\ell \psi_\ell(g), \phi_k \right\rangle \\
&= \int_0^T \sum_{\ell=0}^{\infty} d_\ell \psi_\ell(g_t) \phi_k(t) dt \\
&= \sum_{\ell=0}^{\infty} d_\ell \int_0^T \psi_\ell(g_t) \phi_k(t) dt
\end{aligned}$$

where we can use the dominated convergence theorem in the third equality since $\sum_{\ell \in \mathbb{N}} |d_\ell| < \infty$ and the families $\{\psi_k\}_{k \in \mathbb{N}}$, $\{\phi_k\}_{k \in \mathbb{N}}$ are uniformly bounded. □

Proof of Theorem 3.2. The theorem follows directly from the more general Theorem 4.4 by setting for all $n \in \mathbb{N}$ the regularization parameters $\lambda_n = 0$. □

Proof of Lemma 3.3. We know from the theory of Fourier series (e.g. see Grafakos (2008)) that $d''_\ell := \langle f'', \psi_\ell \rangle = -\ell^2 d_\ell$ for all $\ell \in \mathbb{Z}$. Since the second derivative f'' is continuous on a compact interval, it is bounded and therefore

$$\int_a^b |f''(x)|^2 \, dx = \sum_{\ell=-\infty}^{\infty} |d''_\ell|^2 = \sum_{\ell=-\infty}^{\infty} \ell^4 |d_\ell|^2 < \infty,$$

where we used Parseval's inequality in the first equality. Together with the Cauchy-Schwarz inequality this yields

$$\sum_{|\ell|\geq L} |d_\ell| = \sum_{|\ell|\geq L} \frac{1}{\ell^2}\ell^2|d_\ell| \leq \left(\sum_{|\ell|\geq L} \frac{1}{\ell^4}\right)^{1/2} \cdot \left(\sum_{|\ell|\geq L} \ell^4 d_\ell^2\right)^{1/2} \leq \frac{C}{L^{3/2}},$$

for some constant $C \in \mathbb{R}$ depending on f''. □

Proof of Corollary 3.4. The corollary follows from Theorem 3.2 by plugging in the rate obtained in Lemma 3.3. □

Proof of Corollary 3.6. Since all the results for Torrent depend only on c_n and not on G_n, all of them still hold, and the statement follows directly from Theorem 3.2. □

B.3 Regularized Torrent with Applications to DecoR

Proof of Lemma 4.1. For the first part, note that the number of distinct sets S that Torrent can choose is bounded by $\binom{n}{a}$, thus Torrent has to terminate after a finite number of steps. Secondly, it holds for all t that

$$\begin{aligned}\left\|X_{S_{t+1}}\left(\beta - \hat{\beta}^{t+1}_{Tor}\right) + \epsilon_{S_{t+1}} + o_{S_{t+1}}\right\|_2 + \lambda\left\|\hat{\beta}^{t+1}_{Tor}\right\|_Q & \\ \leq \left\|X_{S_t}\left(\beta - \hat{\beta}^{t+1}_{Tor}\right) + \epsilon_{S_t} + o_{S_t}\right\|_2 + \lambda\left\|\hat{\beta}^{t+1}_{Tor}\right\|_Q & \\ \leq \left\|X_{S_t}\left(\beta - \hat{\beta}^{t}_{Tor}\right) + \epsilon_{S_t} + o_{S_t}\right\|_2 + \lambda\left\|\hat{\beta}^{t}_{Tor}\right\|_Q &\end{aligned} \tag{15}$$

where the first inequality holds because of the hard-thresholding step and the second one holds since $\hat{\beta}^{t+1}_{Tor}$ is a minimizer. □

Proof of Theorem 4.2. Assume that $X^\top_{S_t}X_{S_t} + \lambda Q$ is invertible, else the bound is void. Using the analytic expression for the solution of the Ridge regression we obtain

$$\begin{aligned}\left\|\hat{\beta}^t_{Tor} - \beta\right\|_2 &= \left\|\left(X^\top_{S_{t-1}}X_{S_{t-1}} + \lambda Q\right)^{-1}\left(X^\top_{S_{t-1}}(\epsilon_{S_{t-1}} + o_{S_{t-1}}) - \lambda Q\beta\right)\right\|_2 \\ &\leq \left\|\left(X^\top_{S_{t-1}}X_{S_{t-1}} + \lambda Q\right)^{-1}\left(X^\top_{S_{t-1}}\epsilon_{S_{t-1}} - \lambda Q\beta\right)\right\|_2 \\ &\quad + \left\|\left(X^\top_{S_{t-1}}X_{S_{t-1}} + \lambda Q\right)^{-1}X^\top_{S_{t-1}}\right\|_2 \left\|o_{S_{t-1}}\right\|_2.\end{aligned} \tag{16}$$

Next the hard thresholding step guarantees that

$$\begin{aligned}\left\|X_{S_t}\left(\beta - \hat{\beta}^t_{Tor}\right) + \epsilon_{S_t} + o_{S_t}\right\|_2^2 &= \left\|Y_{S_t} - X_{S_t}\hat{\beta}^t_{Tor}\right\|_2^2 \\ &\leq \left\|Y_{G_n^c} - X_{G_n^c}\hat{\beta}^t_{Tor}\right\|_2^2 \\ &= \left\|X_{G_n^c}\left(\beta - \hat{\beta}^t_{Tor}\right) + \epsilon_{G_n^c}\right\|_2^2\end{aligned} \tag{17}$$

Define $H_t := S_t \setminus G_n^c$ and $M_t := G_n^c \setminus S_t$, then it follows that

$$\left\|X_{H_t}\left(\beta - \hat{\beta}^t_{Tor}\right) + \epsilon_{H_t} + o_{H_t}\right\|_2 \leq \left\|X_{M_t}\left(\beta - \hat{\beta}^t_{Tor}\right) + \epsilon_{M_t}\right\|_2.$$

Observe that $V(S_t) = H_t \cup M_t = V_t$ and $\|o_{H_t}\|_2 = \|o_{S_t}\|_2$. Applying the triangle inequality leads to

$$\begin{aligned}\|o_{S_t}\|_2 \leq& \left\|X_{M_t}\left(\beta - \hat{\beta}^t_{Tor}\right)\right\|_2 + \left\|X_{H_t}\left(\beta - \hat{\beta}^t_{Tor}\right)\right\|_2 + \|\epsilon_{H_t}\|_2 + \|\epsilon_{M_t}\|_2 \\ \leq& \sqrt{2}\,\|X_{V_t}\|\left\|\beta - \hat{\beta}^t_{Tor}\right\|_2 + \sqrt{2}\,\|\epsilon_{V_t}\|_2 \\ \leq& \sqrt{2}\,\|X_{V_t}\|_2 \left(\left\|\left(X^\top_{S_{t-1}} X_{S_{t-1}} + \lambda Q\right)^{-1}\left(X^\top_{S_{t-1}}\epsilon_{S_{t-1}} - \lambda Q\beta\right)\right\|_2 \right. \\ &\left. + \left\|\left(X^\top_{S_{t-1}} X_{S_{t-1}} + \lambda Q\right)^{-1} X^\top_{S_{t-1}}\right\|_2 \|o_{S_{t-1}}\|_2 \right) + \sqrt{2}\,\|\epsilon_{V_t}\|_2\end{aligned}$$

where we used that $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$ for all $a, b \in \mathbb{R}^+_0$ and the submultiplicativity of the operator norm in the second inequality. The third inequality follows form Equation (16).

Next let t be the iteration where Torrent has converged. We show that $S_t = S_{t-1}$, hence the set S_t is stable. First, assume that $\hat{\beta}^t_{Tor} = \hat{\beta}^{t-1}_{Tor}$, then by the hard thresholding step it immediately follows that $S_t = S_{t-1}$. Therefore, assume that $\hat{\beta}^t_{Tor} \neq \hat{\beta}^{t-1}_{Tor}$. Note that $X^\top_{S_{t-1}} X_{S_{t-1}} + \lambda Q$ has to be invertible, else the assumptions on η are not satisfied. Thus $\hat{\beta}^t_{Tor}$ is the unique minimizer and Equation (15) has to be a strict inequality. But this implies that the algorithm did not stop, which is a contradiction to our assumption that S_t is the output of Torrent. Hence we can conclude that $S_t = S_{t-1}$.

To obtain a bound for the outliers, recall that the 2-norm of a matrix $A \in \mathbb{R}^{d\times n}$ is equal to largest singular value, in other words $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)} = \sqrt{\lambda_{\max}(AA^\top)} = (\lambda_{\min}((AA^\top)^{-1}))^{-1/2}$. Together with the superadditivity of eigenvalues for symmetric matrices this implies that

$$\begin{aligned}\left\|\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1} X^\top_{S_t}\right\|_2 =& \sqrt{\lambda_{\max}\left(\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1} X^\top_s X_s \left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1}\right)} \\ \leq& \sqrt{\lambda_{\max}\left(\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1} (X^\top_s X_s + \lambda Q) \left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1}\right)} \\ =& \sqrt{\lambda_{\max}\left(\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1}\right)} = \left(\lambda_{\min}\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)\right)^{-1/2}.\end{aligned}$$

Thus, we obtain the bound

$$\|X_{V_t}\|\left\|\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1} X^\top_{S_t}\right\|_2 \leq \max_{S\subseteq\{1,\dots,n\} \text{ s.t. } |S|=a_n} \frac{\|X_{V(S)}\|_2}{\sqrt{\lambda_{\min}\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)}} = \eta < \frac{1}{\sqrt{2}}$$

with high probability for all n large enough. Because $S_t = S_{t-1}$, the norm of the outliers can be bounded as follows

$$\|o_{S_t}\|_2 \leq \sqrt{2}\eta\|o_{S_t}\|_2 + \sqrt{2}\,\|X_{V_t}\|_2\left\|\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1}\left(X^\top_{S_t}\epsilon_{S_t} - \lambda Q\beta\right)\right\|_2 + \sqrt{2}\|\epsilon_{V_t}\|_2$$

and therefore

$$\|o_{S_t}\|_2 \leq \frac{\sqrt{2}\,\|X_{V_t}\|_2\left\|\left(X^\top_{S_t} X_{S_t} + \lambda Q\right)^{-1}\left(X^\top_{S_t}\epsilon_{S_t} - \lambda Q\beta\right)\right\|_2 + \sqrt{2}\|\epsilon_{V_t}\|_2}{1 - \sqrt{2}\eta}.$$

Substituting this into Equation (16) gives

$$\left\|\hat{\beta}^t_{Tor} - \beta\right\|_2 \leq \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1}\left(X_{S_t}^\top \epsilon_{S_t} - \lambda Q\beta\right)\right\|_2 + \frac{\sqrt{2}\left\|X_{V_t}\right\|_2 \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1}\left(X_{S_t}^\top \epsilon_{S_t} - \lambda Q\beta\right)\right\|_2 + \sqrt{2}\|\epsilon_{V_t}\|_2}{\sqrt{\lambda_{\min}\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)}(1-\sqrt{2}\eta)}.$$

□

Before showing Corollary 4.3, we are going to state a concentration inequality for sub-Gaussian random variables.

Lemma B.1 (Vershynin (2018))**:** *Let μ be a sub-Gaussian distribution with zero mean and unit variance. There exists a constant $K > 0$ such that for all $n \in \mathbb{N}$ the following holds: Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. distributed with respect to μ and define $\epsilon^n := (\epsilon_1, \ldots, \epsilon_n)$. Then for all $A \in \mathbb{R}^{m\times n}$ and $t \geq 0$ we have that*

$$\mathbb{P}\left(\left|\|A\epsilon^n\|_2 - \|A\|_F\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{K\|A\|_2^2}\right).$$

Proof of Corollary 4.3. By construction it holds that $|S_t|, |G_n^c| \geq n - c_n$ implying that $|(G_n^c \cap S_t)| \geq n - 2c_n$. Thus

$$|V(S_t)| \leq 2c_n$$

From Lemma B.1 we know that there exists a $K > 0$ such that for all $t \geq 0$ we have

$$\mathbb{P}\left[\forall S \in \mathcal{U}_n, \left|\|\epsilon_{V(S)}\|_2 - \sigma\sqrt{|V(S)|}\right| \geq t\right] \leq 2|\mathcal{U}_n|\exp\left(-\frac{t^2}{K\sigma^2}\right)$$

and in particular

$$\mathbb{P}\left[\forall S \in \mathcal{U}_n, \left|\|\epsilon_{V(S)}\|_2 - \sigma\sqrt{|V(S)|}\right| \leq \sigma\sqrt{K\log(2|\mathcal{U}_n|/\delta)}\right] \geq 1-\delta.$$

Thus we can bound the error term for all $t \in \mathbb{N}$ by

$$\left\|\epsilon_{V(S_t)}\right\|_2 \leq \sigma\left(\sqrt{2c_n} + \sqrt{K\log(2|\mathcal{U}_n|/\delta)}\right) \leq \sigma\sqrt{2c_n}\left(1 + \sqrt{K\log(2en/c_n\delta)}\right)$$

where we used that $|\mathcal{U}_n| = |\{S \subseteq \{1,\ldots,n\} \text{ s.t. } |S| = a_n\}| = \binom{n}{a_n} = \binom{n}{c_n} \leq (en/c_n)^{c_n}$. Using a similar argument and $\left\|\left(\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} X_{S_t}^\top\right\|_F \leq \sqrt{d}\left\|\left(\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} X_{S_t}^\top\right\|_2$ we get that for all $t \in \mathbb{N}$ with probability at least $1-\delta$

$$\left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} X_{S_t}^\top \epsilon_{S_t}\right\|_2 \leq \sigma\left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} X_{S_t}^\top\right\|_2\left(\sqrt{d} + \sqrt{2c_nK\log(2en/c_n\delta)}\right).$$

By the fact that

$$\left\|\left(\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1}\right\|_2 = \left(\lambda_{\min}\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)\right)^{-1/2}$$

and Theorem 4.2 we obtain

$$\begin{aligned}
\left\|\hat{\beta}^t_{Tor} - \beta\right\|_2 \leq& \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1}\left(X_{S_t}^\top(\epsilon_{S_t} + r_{S_t}) - \lambda Q\beta\right)\right\|_2 \\
&+ \frac{\sqrt{2}\left\|X_{V_t}\right\|_2 \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1}\left(X_{S_t}^\top(\epsilon_{S_t} + r_{S_t}) - \lambda Q\beta\right)\right\|_2 + \sqrt{2}\|\epsilon_{V_t} + r_{V_t}\|_2}{\sqrt{\lambda_{\min}\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)}(1-\sqrt{2}\eta)} \\
=& \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} X_{S_t}^\top(\epsilon_{S_t} + r_{S_t})\right\|_2 \\
&+ \frac{\sqrt{2}\left\|X_{V_t}\right\|_2 \left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} X_{S_t}^\top(\epsilon_{S_t} + r_{S_t})\right\|_2 + \sqrt{2}\|\epsilon_{V_t} + r_{V_t}\|_2}{\sqrt{\lambda_{\min}\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)}(1-\sqrt{2}\eta)} \\
&+ \lambda\left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} Q\beta\right\|_2 \left(1 + \frac{\sqrt{2}\left\|X_{V_t}\right\|_2}{\sqrt{\lambda_{\min}\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)}(1-\sqrt{2}\eta)}\right) \\
\leq& \frac{1}{\gamma^\lambda_{a_n}}\left(1 + \frac{\sqrt{2}}{1-\sqrt{2}\eta}\right)\left(\sigma\left(\sqrt{d} + \sqrt{2c_nK\log(2en/c_n\delta)}\right) + \|r_{S_t}\|_2\right) \\
&+ \frac{2\sigma\sqrt{c_n}\left(1 + \sqrt{K\log(2en/c_n\delta)}\right) + \|r_{V_t}\|_2}{\gamma^\lambda_{a_n}(1-\sqrt{2}\eta)} + \lambda\frac{\|Q\beta\|_2}{\gamma^\lambda_{a_n}}\left(1 + \frac{1}{1-\sqrt{2}\eta}\right),
\end{aligned}$$

where we used that

$$\left\|\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)^{-1} X_{S_t}^\top\right\|_2 \leq \left(\lambda_{\min}\left(X_{S_t}^\top X_{S_t} + \lambda Q\right)\right)^{-1/2},$$

see the proof of Theorem 4.2 for a derivation. □

To prove Theorem 4.4, we need the following lemma.

Lemma B.2: *Let c_n be a sequence such that $|G_n| \leq c_n$ for all n, let $\mathcal{U}_n = \{S \subseteq \{1,\dots,n\} \mid |S| = n - c_n\}$ and let Assumption 2 be satisfied. Then*

(i) $|\mathcal{U}_n| \leq \left(\frac{en}{c_n}\right)^{c_n}$.

(ii) for all $n \in \mathbb{N}$ the components of η^n_ϕ are i.i.d. centered Gaussian random variables with variance $\overline{\sigma}^2 = \sigma^2_\eta/n$,

(iii) for all $\delta > 0$ there exists $c' > 0$ and $\overline{n} \in \mathbb{N}$ such that for all $n > \overline{n}$ it holds that

$$\mathbb{P}\left[\min_{S\in\mathcal{U}_n}\sqrt{\lambda_{min}\left((P^n_{\psi,\phi})^\top_{S\setminus G_n}(P^n_{\psi,\phi})^\top_{S\setminus G_n} + \lambda_n Q_n\right)} \geq c'\right] \geq 1-\delta.$$

.

Proof. (Lemma B.2)

(i) It follows by Sterling's inequality directly that

$$|\mathcal{U}_n| = \binom{n}{c_n} \leq \left(\frac{en}{c_n}\right)^{c_n}.$$

(ii) Note that $\left\{T_k^{\phi,n}(\eta)\right\}_{k\leq n}$ are jointly Gaussian distributed with mean zero since they are linear combinations of independent Gaussian variables. From Assumption 2 it follows that for all $k, \ell \in \{1, \dots, n\}$

$$\mathbb{E}\left[T_k^{\phi,n}(\eta)T_\ell^{\phi,n}(\eta)\right] = \frac{1}{n^2}\sum_{i,j=1}^{n}\phi_k(Ti/n)\phi_\ell(Tj/n)\mathbb{E}\left[\eta_{iT/n}\eta_{jT/n}\right] = \frac{\sigma_\eta^2}{n}\mathbb{1}_{\{k=\ell\}},$$

thus the transformed noise is i.i.d. centered Gaussian with variance $\overline{\sigma}^2 = \sigma_\eta^2/n$.

(iii) Observe that for all $S \in \mathcal{U}_n$ we have $|S| = n - c_n$ by construction and $|G_n| \leq c_n$. Thus for all $S \in \mathcal{U}_n$ it holds that $|S \setminus G_n| \geq n - 2c_n$. Next consider $\overline{n}$ and S_n' from Assumption 2 and an arbitrary $n \geq \overline{n}$. Since $|S \setminus G_n| \geq n - c_n$ we have $|(S \setminus G_n) \cap S_n'| \geq L(n)$. Thus we can chose a set $S'' \subseteq (S \setminus G_n) \cap S_n'$ with $|S''| = L(n)$. Hence by assumption Assumption 2 we have with probability at least $1 - \delta$

$$\min_{S\in\mathcal{U}_n}\sqrt{\lambda_{\min}\left((P_{\psi,\phi}^n)_{S''}^\top(P_{\psi,\phi}^n)_{S''} + \lambda_n Q_n\right)} \geq c_n'.$$

Since the minimal eigenvalue is superadditive for positive semi-definite matrices, we have with probability at least $1 - \delta$

$$\min_{S\in\mathcal{U}_n}\sqrt{\lambda_{\min}\left((P_{\psi,\phi}^n)_{S\setminus G_n}^\top(P_{\psi,\phi}^n)_{S\setminus G_n} + \lambda_n Q_n\right)} \geq c_n'.$$

□

Proof of Theorem 4.4. Using the triangle inequality we can bound

$$\left\|\hat{d}_{DecoR}^{\psi,n} - d\right\|_2 \leq \left\|\hat{d}_{DecoR}^{\psi,n} - d_{\{1,\dots,L\}}\right\|_2 + \left\|d_{\{L+1,L+2,\dots\}}\right\|_2. \tag{18}$$

We are going to bound the two summands on the right side separately. For the first summand, Equation (13) implies that for all n large enough we have with high probability

$$\eta \coloneqq \max_{S\subseteq\{1,\dots,n\}\text{ s.t. }|S|=n-c_n}\frac{\left\|\left((P_{\psi,\phi}^n)^\top\right)_{V(S)}\right\|_2}{\sqrt{\lambda_{\min}\left((P_{\psi,\phi}^n)_S(P_{\psi,\phi}^n)_S^\top + \lambda_n Q_n\right)}} < \frac{1}{\sqrt{2}}.$$

Moreover we know by Lemma B.2 and the superaddivity of the eigenvalues for positive semi-definite matrices, that for n large enough for all $S \in \mathcal{U}_n$ the expression $\sqrt{\lambda_{\min}\left((P_{\psi,\phi}^n)_S(P_{\psi,\phi}^n)_S^\top + \lambda_n Q_n\right)}$ is lower bounded by some constant. Thus the assumptions of Corollary 4.3 are fulfilled and we can bound the first summand by

$$\begin{aligned}\left\|\hat{d}_{DecoR}^{\psi,a_n} - d_{\{1,\dots,L\}}\right\|_2 \leq& \frac{1}{\gamma_{a_n}^\lambda(P_{\psi,\phi}^n)}\left(1 + \frac{\sqrt{2}}{1-\sqrt{2}\eta}\right)\left(\sigma_\eta\left(\sqrt{L(n)} + \sqrt{2c_nK\log(2en/c_n\delta)}\right) + \|r_{S_t}^n\|_2\right)\\ &+ \frac{2\sigma_\eta\sqrt{c_n}\left(1 + \sqrt{K\log(2en/c_n\delta)}\right) + \|r_{V_t}^n\|_2}{\gamma_{a_n}^\lambda(P_{\psi,\phi}^n)(1-\sqrt{2}\eta)}\\ &+ \lambda_n\frac{\|Q_n d_{L(n)}\|_2}{\gamma_{a_n}^\lambda}\left(1 + \frac{1}{1-\sqrt{2}\eta}\right).\end{aligned} \tag{19}$$

By assumption, there exists a $C_1 > 0$ such that

$$\sum_{\ell \geq L(n)} |d_\ell| \leq C_1 g(L(n)),$$

and recall from Equation (9) that the bias introduce by the approximation is given by

$$(r_L^n)_k = \sum_{\ell \geq L} d_\ell P^n_{\psi,\phi}(\ell, k).$$

Because the bases ψ and ϕ are uniformly bounded it follows by definition that there exists some $C_2 > 0$ such that $max_{k,\ell} P^n_{\phi,\psi}(k,\ell) \leq C_2$. Thus it holds that

$$\begin{aligned}\|r_L^n\|_2 &= \left(\sum_{k=1}^n (r_L^n)_k^2\right)^{1/2} = \left(\sum_{k=1}^n \left(\sum_{\ell \geq L} d_\ell P^n_{\psi,\phi}(\ell,k)\right)^2\right)^{1/2} \\ &\leq \left(\sum_{k=1}^n \left(\sum_{\ell \geq L} \left|d_\ell P^n_{\psi,\phi}(\ell,k)\right|\right)^2\right)^{1/2} \leq C_2\sqrt{n}\sum_{\ell \geq L}|d_\ell| \leq C_1 C_2 \sqrt{n} g(L(n)).\end{aligned}$$

Using this bound on the bias term r in Equation (19) together with the fact $\sigma = \sigma_\eta/\sqrt{n}$ from Lemma B.2 and the assumption $\sup_n \|Q_n d_{L(n)}\|_2 < \infty$ yields

$$\left\|\hat{d}^{\psi,n}_{DecoR} - d_{1,\dots,L}\right\|_2 \in \mathcal{O}\left(\sigma_\eta\left(\sqrt{\frac{c_n \log(n/c_n)}{n}} + \sqrt{\frac{L(n)}{n}}\right) + \sqrt{n} g(L(n)) + \lambda_n\right).$$

For the second summand in Equation (18) it follows from Hölder's inequality that

$$\left\|d_{\{L+1,L+2,\dots\}}\right\|_2 \leq \left\|d_{\{L+1,L+2,\dots\}}\right\|_1 \in \mathcal{O}(g(L(n)).$$

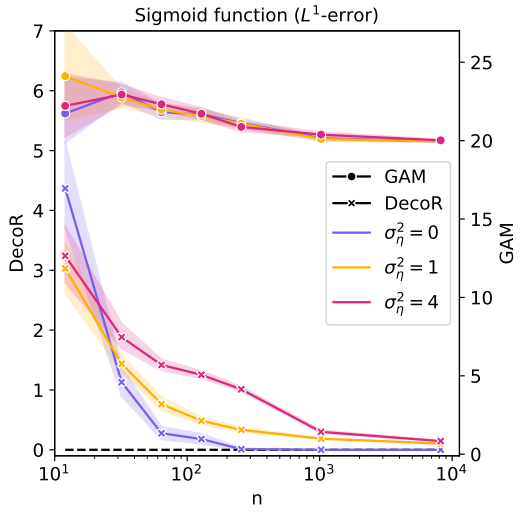
Since for all $n \in \mathbb{N}$ $g(L(n)) \leq \sqrt{n} g(L(n))$ the statement follows. □

Appendix C: Supplementary Numerical Results

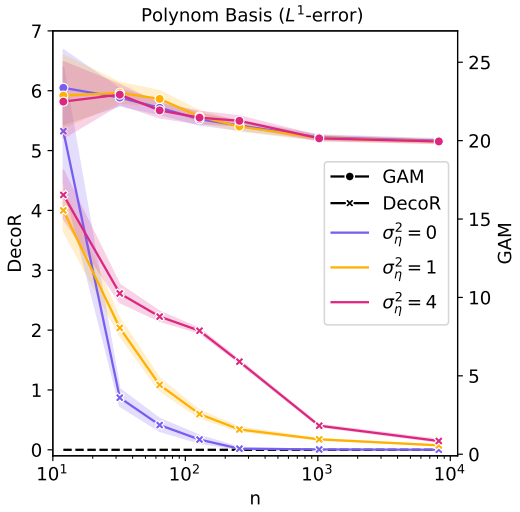
Here we present some additional numerical examples, indicating that the nonlinear extensions of DecoR yield similar results across a wide range of settings.

C.1 Sigmoid Function and Polynom Basis

Consider the sigmoid function $f(x) = \frac{20}{1+\exp(-16x+6)} - 10$. We investigate two settings. In Figure C.1a we choose $X_t \overset{i.i.d.}{\sim} \mathcal{U}([0,1])$ and ψ is taken to be the cosine basis. In Figure C.1b we choose X_t to be a reflected Ornstein-Uhlenbeck process and choose $\psi_\ell = x^\ell$ the monomial basis. Note that the monomial basis is neither orthogonal nor normalized. In both settings, the noise is i.i.d. normally distributed with mean 0 and variance $\sigma_\eta^2 \in \{0, 1, 4\}$. We have an outlier fraction of 0.25 and run DecoR with Torrent and the threshold parameters $a_n = 0.7n$.



(C.1a) L^1-error vanishes for DecoR with the underlying truth f being the sigmoid function and X_t i.i.d. uniformly distributed.



(C.1b) DecoR is robust and yields a consistent estimator when the monomial basis is used for the approximation.

In both settings, the L^1-error of DecoR vanishes with increasing sample size. Even when using the monomial basis, which is not orthonormal, we still obtain a consistent estimator, indicating robustness to some model violations.

In addition, we also test the performance of the proposed confidence intervals like we did in Section 4.4 for the sine function. We take X_t i.i.d. uniformly on $[0, 1]$ and Gaussian noise with variance $\sigma^2 = 1$. Both analytic confidence intervals, relying on the t-distribution and looking at DecoR as a linear estimator, yield systematically too high and too low coverage for taking all and only the estimated inliers to estimate the variance, respectively. Again, using double bootstrapping leads to the actual coverage closest to the nominal coverage.

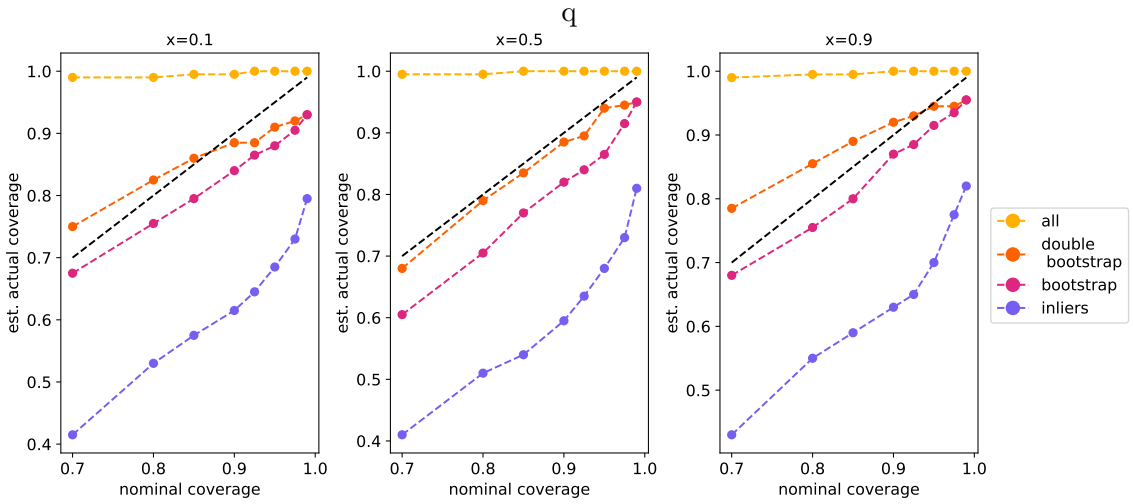


Figure C.2: Performance of the different confidence intervals for the sigmoid function. Double bootstrapping leads to a coverage closest to the intended coverage, similar to the case of f being the sine function.

C.2 Eigenvalue Condition

We choose as underlying truth f the sine function of Section 5.2 and adopt all parameters. Looking at the proof of Theorem 3.2, we see that it is sufficient for the eigenvalue condition Equation (10) to hold for the estimated set of inliers S_t by DecoR. We draw 10^4 samples and compute the fraction

$$\frac{\left\|(P^n_{\psi,\phi})^\top_{V(S_t)}\right\|_2}{\sqrt{\lambda_{\min}\left((P^n_{\psi,\phi})^\top_{S_t}(P^n_{\psi,\phi})_{S_t}\right)}}.$$

For the eigenvalue condition to hold, this fraction must be smaller than $1/\sqrt{2}$. Looking at Figure C.3, we observe that the variance makes it harder for the eigenvalue condition to be satisfied. However, for around 70% of the cases, the condition is satisfied. It is important to note that the eigenvalue condition is sufficient for the theorem but not necessary; meaning it is too strong.

C.3 Regularized Torrent

We provide in Table C.1 the relative error for DecoR run with regularized Torrent when the variance of the noise is $\sigma_\eta = 4$, complementing Table 5.1 and Table 5.2. Observe that clipping the residuals and using the one-standard error rule performs best, leading to an L^1-error that is at most 2.8% greater than when using no regularization. In some of the settings, the expected L^1-error becomes worse when choosing the regularization parameter with any of the six rules. Interestingly, this is mainly the case for medium sample sizes $n \in \{2^7, 2^8\}$.

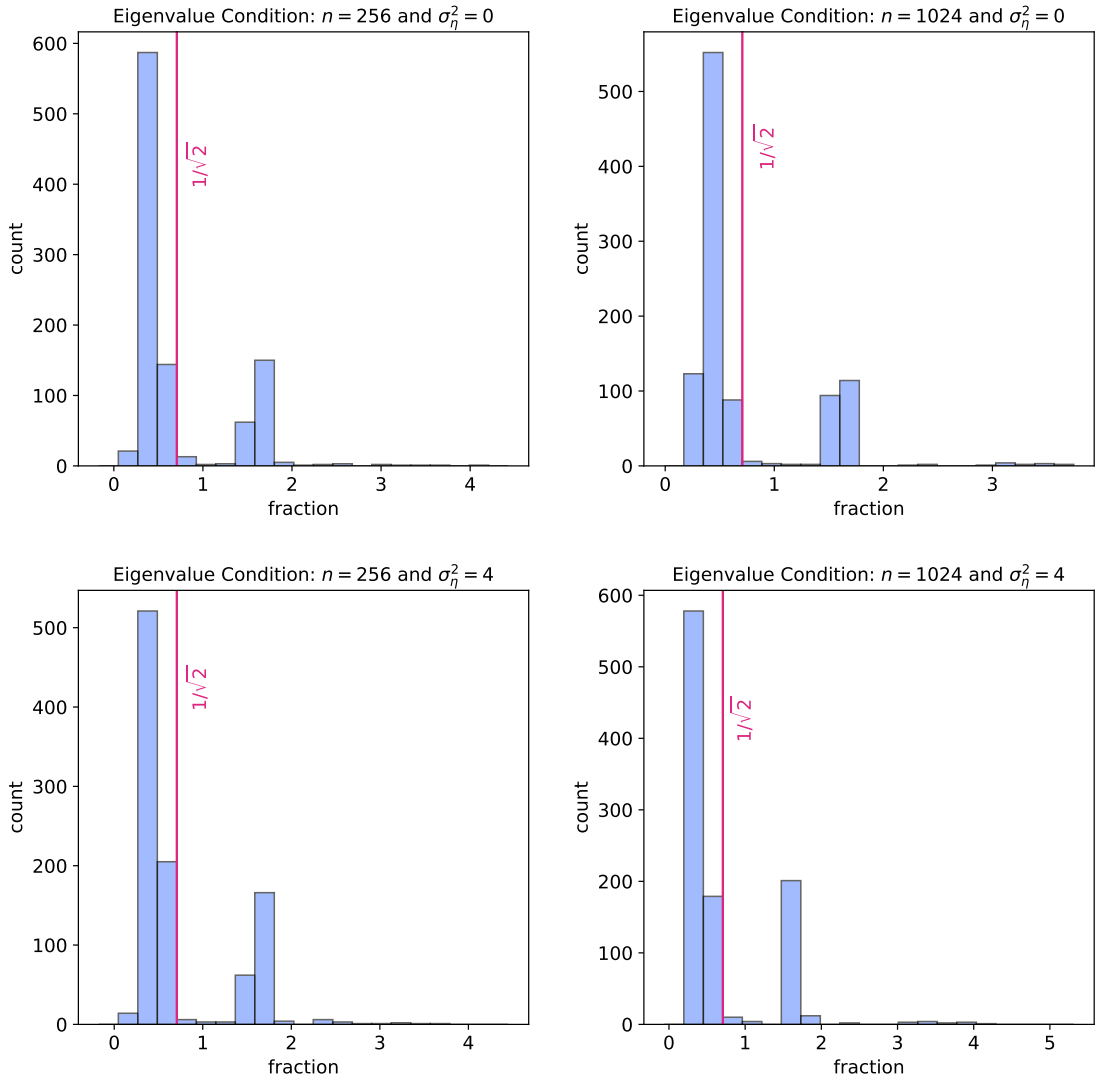


Figure C.3: Satisfiability of the eigenvalue condition given in Equation (10). It appears to be the case that the bigger the variance, the less often the eigenvalue condition holds.

method \ sample size n		32	64	128	256	1024	8192
L^1-error	Torrent	2.607	1.106	0.852	0.636	0.371	0.207
rel. error (regularized)	Clip	0.688	0.981	1.037	1.041	0.926	0.724
	Omit	1.227	1.179	1.191	1.289	1.212	0.800
	Median	0.938	1.115	1.123	1.422	1.256	0.790
	Clip 1-S.E.	0.689	0.975	1.028	1.031	0.907	0.711
	Omit 1-S.E.	1.201	1.154	1.164	1.271	1.176	0.783
	Median 1-S.E.	0.918	1.099	1.187	1.383	1.225	0.777

Table C.1: Relative error compared to no regularization for $\sigma_\eta^2 = 4$.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. In consultation with the supervisor, one of the following three options must be selected:

○ I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

○ I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used and cited generative artificial intelligence technologies[2].

◉ I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used generative artificial intelligence technologies[3]. In consultation with the supervisor, I did not cite them.

Title of paper or thesis:

Inference of Nonlinear Causal Effects in Time Series in the Presence of Confounding

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Blieske	Pio

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guide.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zürich, 17.03.2025	Pio Blieske

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] E.g. ChatGPT, DALL E 2, Google Bard
[2] E.g. ChatGPT, DALL E 2, Google Bard
[3] E.g. ChatGPT, DALL E 2, Google Bard