Causal Forecasting: Generalization Bounds for Autoregressive Models

Abstract

Despite the increasing relevance of forecasting methods, causal implications of these algorithms remain largely unexplored. This is concerning considering that, even under simplifying assumptions such as causal sufficiency, the statistical risk of a model can differ significantly from its *causal risk*. Here, we study the problem of *causal generalization*—generalizing from the observational to interventional distributions—in forecasting. Our goal is to find answers to the question: How does the efficacy of an autoregressive (VAR) model in predicting statistical associations compare with its ability to predict under interventions? To this end, we introduce the framework of *causal learning theory* for forecasting. Using this framework, we obtain a characterization of the difference between statistical and causal risks, which helps identify sources of divergence between them. Under causal sufficiency, the problem of causal generalization amounts to learning under covariate shifts albeit with additional structure (restriction to interventional distributions). This structure allows us to obtain uniform convergence bounds on causal generalizability for the class of VAR models. To the best of our knowledge, this is the first work that provides theoretical guarantees for causal generalization in the time-series setting.

1 INTRODUCTION

Forecasting algorithms are increasingly relevant in a variety of applications including meteorology, climatology, economics, and business. While traditional economic modelling relies on relatively simple time series models (Brockwell et al. 1991), e.g., autoregressive models, or methods like co-integration, modern business planning heavily uses neural networks for forecasting (Faloutsos et al. 2018; Januschowski et al. 2020; Salinas et al. 2020). Despite the advancements of forecast quality, causal implications are not yet well understood. There has been notable progress in 'explainable' forecasts in the sense of feature relevance (Lundberg et al. 2017; Molnar 2019; Janzing et al. 2020; Wang et al. 2020). Furthermore, specialized models (Hatt et al. 2021; Bica et al. 2020; Lim et al. 2018) have shown remarkable success for causal inference in forecasting.

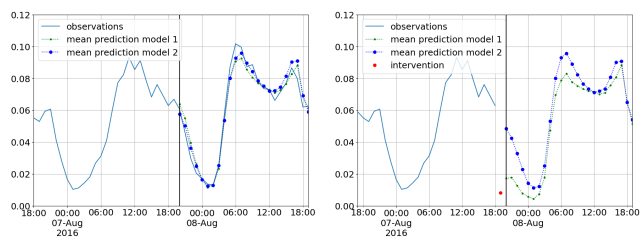

Figure 1: An example time series with predictions of two DeepAR models (top) under an intervention in red (bottom).

Here, we are interested in the fundamental question: what is the relation between the statistical predictability of a forecasting model and its causal generalizability — ability to predict under interventions. In practice, while forecasting models tend to agree on their statistical predictions, they can differ substantially on their causal predictions. This is also demonstrated in our experiments on real data in Section 5. We show an example in Figure 1 on the traffic dataset from Dua et al. 2017 . While we do not know the ground-truth, we see that two models disagree when faced with an intervention more than on the in-distribution forecasting. Since at most one of them can be right, we conclude that at least the other one makes a notable forecasting error under the intervention.

In this work, we argue that even for very simple models and even under simplifying assumptions such as causal sufficiency and absence of contemporaneous influence, causal interpretation of forecasting models is non-trivial. To appreciate the challenges, consider a simple example of a process with strongly correlated observations where $x_t \approx x_{t-1}$, and

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

hence $x_t \approx x_{t-2}$. These observations can be explained either by a causal model with a strong influence of x_{t-1} on x_t or a causal model with a strong influence from x_{t-2} on x_t. The difference between the models gets apparent when an intervention randomizes x_{t-1} and x_{t-2} independently. Then, predictions become hard, particularly when x_{t-1} and x_{t-2} are set to significantly different values. While both models are similar in their statistical predictions, they differ substantially in their *causal predictions*. This example already shows that, even in a simple setting, causal and statistical predictability can differ significantly. The question of causal generalization is thus practically relevant and non-trivial and begs for a better theoretical understanding.

Specifically, we consider the simple class of vector autoregressive models (VAR) and ask the question

How does the efficacy of an autoregressive model in predicting statistical associations compare with its ability to predict under interventions?

These models are widely applied in domains ranging from econometrics (Lütkepohl 2009; Grabowski et al. 2020) and finance (Zivot et al. 2006) to neuroscience (Valdés-Sosa et al. 2005).

Connection to Covariate Shift. The problem of causal generalization is closely related to the problem of covariate shift. To see this, we first ignore the time series setting and consider the scenario where a variable Y should be predicted from a variable X, which is known not to be an effect of Y. If there is no common cause of X and Y, that is, we assume causal sufficiency (Spirtes et al. 1993), the statistical relation between X and Y is entirely due to the influence of X on Y. Therefore, the observational and interventional conditionals coincide ($P_{Y|x=x^*} = P_{Y|do(x=x^*)}$ in Pearl's language (Pearl 2009)) and the true parameters would be optimal both from a statistical and causal perspective. However, due to *estimation bias*, a prediction model learned using finite samples from P_x may perform poorly when randomized interventions draw x-values from a different distribution $\tilde{P}_X$, which is the usual covariate shift scenario (Sugiyama et al. 2012). In our setting, X and Y are represented by the past and the present values of a (possibly multivariate) time series, respectively. Accordingly, we focus on interventional distributions that are natural for this setting: independent interventions at different time points and components of the multivariate process. Hence, we have additional structure in comparison with the standard covariate shift problem. We are not aware of any theoretical work on covariate shift in the time-series setting. Nevertheless, we describe the connections to learning theory in the standard covariate shift setting and other related work in Section 6.

Our Contributions. Our central goal in this work is to develop a formal and thorough understanding of causal generalization for the class of VAR models.

a. To this end, we introduce a framework of causal learning theory for forecasting to analyze when forecasting models can generalize from the *observational* to the *interventional distributions* (Section 2). This is closely related to the setting of learning under domain adaptation.

b. Using this framework, we provide a thorough characterization of the difference in the statistical and *causal* risks (Section 3). Such a characterization allows us to identify the sources of divergence between the two quantities. Our results show that the strength of correlation of the underlying process plays a key role in determining causal generalizability. They also highlight that already for simple models, causal and statistical errors can even diverge.

c. Further, we provide finite-sample, uniform convergence bounds on causal generalization for the class of VAR models (Section 3). Our simulations demonstrate that our bounds indeed capture the key drivers of causal generalization. To the best of our knowledge, this is the first work that provides theoretical guarantees for causal generalization of any kind in the time-series setting.

d. As a by-product of our analysis, we provide an explicit characterization of the powers of a companion matrix (see Section 2) using symmetric Schur polynomials (Macdonald 1998) of its eigenvalues (Lemma 2) which, to the best of our knowledge, has not been noted in the literature. This result could be of independent interest in theoretical endeavors that build upon companion matrices which, for instance, are ubiquitous in stochastic processes and in Linear-Time-Invariant dynamical systems (Davison 1976; Melnyk et al. 2016).

e. We conduct experiments with a variety of deep neural networks on real data. Our experiments approach causal risks in this setting and explore its relationship to uncertainty.

2 CAUSAL LEARNING THEORY FOR FORECASTING

In this section, we introduce a framework to formally evaluate the quality of a forecasting model with respect to prediction and the validity of its causal implications. We refer to this framework as causal learning theory for forecasting. First, we introduce some relevant notation.

Notation. For any stochastic process $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, we use $\mathbf{x}_{t-\omega}^n = \{x_{t-\omega-n+1}, \cdots, x_{t-\omega-1}, x_{t-\omega}\}$ to denote the *set* of $x_{t-\omega}$ and the $n-1$ variables in the past of $x_{t-\omega}$. We distinguish this from y_t^n which denotes the *vector* $\left(x_t, x_{t-1}, \cdots, x_{t-n+1}\right)^T \in \mathbb{R}^{nd}$. When it is clear from context, to reduce cumbersome notation, we simply use y_t. For any random variable x, $\mathbb{E}[x]$ denotes its expectation. For any matrix A, we use $A_{i:}$ and $A_{:j}$ to denote the ith row and jth column of A respectively. We use A_{1k}^j to denote the $(1,k)$th element of A^j. For any vector x_t at time t, we use $x_{t,i}$ to denote the ith element of x_t. We use

$\lambda_{\max}(A), \lambda_{\min}(A), \kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ to denote the maximum and minimum eigenvalues and the condition number of A respectively. $\mathbb{I}_p$ denotes the identity matrix of size p, $\mathbb{N}, \mathbb{Z}$ denote the set of natural numbers and integers respectively and $[n]$ denotes the set $\{1, 2, \cdots n\}$.

To evaluate the statistical and causal efficacy of an estimator we introduce the notions of statistical and *causal* forecast risks. To define statistical forecast risk, we consider the setting of $\omega-$step forecasting where the goal is to predict x_t from observations $\mathbf{x}_{t-\omega}^n$ drawn from a stochastic process $\{x_t\}_{t\in\mathbb{Z}}$ for some $\omega \in \mathbb{N}$. To define the causal forecast risk, we consider interventions on $x_{t-\omega,i}$ for some $i \in [d]$.[1]

Definition 2.1 (**Statistical forecast error**). The statistical forecast error of an estimator $\hat{f}$ in the prediction of a target variable x_t from $\mathbf{x}_{t-\omega}^n$, drawn from the *observational distribution*, can be defined as

$$\mathcal{S}_\omega = \mathbb{E}_{\mathbb{P}(x_t, \mathbf{x}_{t-\omega}^n)}\left[\left(x_t - \hat{f}(x_{t-\omega}^n)\right)^2\right]. \tag{1}$$

The empirical counterpart ($\hat{\mathcal{S}}_\omega$), is defined naturally by replacing the expectation by the empirical mean.

For causal questions, we want to investigate the behavior of a model under interventions. Here, we consider atomic interventions. Using Pearl's do notation (Pearl 2009), an atomic intervention $do(x = x^*)$ refers to *setting* the variable x to some value x^*.

Definition 2.2 (**Causal errors**). The interventional forecast error of $\hat{f}$ in predicting the *effect of an intervention* $do(x_{t-\omega,i} = x_{t-\omega,i}^*)$, on target variable x_t is defined as

$$\mathcal{G}_{do_{\omega,i}} = \mathbb{E}_{\mathbb{P}_{do_{\omega,i}}(x_t, \mathbf{x}_{t-\omega}^n)}\left[\left(x_t - \hat{f}(x_{t-\omega}^n)\right)^2\right], \tag{2}$$

where $do_{\omega,i}$ is shorthand for $do(x_{t-\omega,i} = x_{t-\omega,i}^*)$ and $\mathbb{P}_{do_{\omega,i}}$ denotes the distribution induced by the intervention $do(x_{t-\omega,i} = x_{t-\omega,i}^*)$. To isolate from the dependence on specific values that the intervened variables are set to, we present our results via the notion of *average causal error*. It is defined as the expected interventional error for interventions drawn from the marginal distribution of $x_{t-\omega,i}$ since it provides a natural scale at which the statistical and causal errors can be compared.

$$\mathcal{G}_{\omega,i} = \mathbb{E}_{x_{t-\omega,i}^* \sim \mathbb{P}(x_{t-\omega,i})}\left[\mathcal{G}_{do_{\omega,i}}\right]. \tag{3}$$

Statistical and Causal Learning Theory. Consider the standard framework of statistical learning in time-series prediction. For any stochastic process $\{x_t\}_{t\in\mathbb{Z}}$ taking values in $\mathcal{X}$, given a loss function $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, the goal of statistical learning is to learn a function $f_{\mathcal{S}}^*$ that achieves the optimal statistical risk $\mathcal{S}^\omega(f)$: Since the true process is unknown, the empirical average ($\widehat{\mathcal{S}}^\omega$) of generalization risk is used to estimate $\mathcal{S}^\omega$. Statistical generalization bounds of the form: $\mathcal{S}^\omega(f) < \widehat{\mathcal{S}}^\omega(f) + \mathcal{C}(\mathcal{F}, n)$ are then used to provide guarantees on the uniform deviation of empirical risk from expected risk given sufficiently many samples and when the "complexity" of the function class is small.

Analogously, the goal of *causal learning* is to find a function $f_{\mathcal{G}}^*$ that achieves the optimal *causal* risk $\mathcal{G}^\omega(f)$ In contrast to statistical learning, the empirical averages of the causal error cannot be utilized to estimate $\mathcal{G}_\omega$ since we often do not have access to data from the interventional distributions. Instead, we are only provided with data from the observational/statistical distribution of the stochastic process and the goal of causal learning theory is to understand, to what extent is it possible to provide *causal generalization* guarantees of the form: $\mathcal{G}^\omega(f) < \widehat{\mathcal{S}}^\omega(f) + \mathcal{C}(\mathcal{F}, n)$.

To summarize, we ask: Can the predictors in $\mathcal{F}$ generalize from the *empirical observational distribution* to the *true interventional distribution* assuming that we control the complexity of $\mathcal{F}$ and that we observe sufficiently many samples drawn from the observational distribution? One cannot address this question in a very general setting and would need model assumptions to make any meaningful statements. To this end, we now formally introduce our problem setup and some preliminaries. We provide additional relevant background in the Appendix A.

Statistical and Causal Models. We assume that the stochastic process $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$ follows a weakly stationary vector autoregressive model(VAR(p)) of order p for some $p, d \in \mathbb{N}$ which is defined as

$$x_t = A_1 x_{t-1} + A_2 x_{t-2} + \cdots A_P x_{t-p} + \epsilon_t, \tag{4}$$

where $x_t \in \mathbb{R}^d$ is a vector-valued time-series, for all $i \in [p]$, $A_i \in \mathbb{R}^{d\times d}$ are the coefficients of the VAR model, and $\epsilon_t \in \mathbb{R}^d$ denotes the noise vector such that $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t \epsilon_{t+h}^T] = \Sigma_\epsilon$ if $h = 0$ and 0 otherwise. For some $\sigma_\epsilon^2 > 0$, we simply set $\Sigma_\epsilon = \sigma_\epsilon^2 \mathbb{I}$ for enhanced readability. Our results can be easily generalized to arbitrary covariance matrices by means of the spectral properties $(\lambda_{\min}, \lambda_{\max})$ of Σ_ϵ. The autocovariance matrix of $\{x_t\}_{t\in\mathbb{Z}}$ plays a central role in our results and analysis. For any $n \in \mathbb{N}$, we use Σ_n to denote the autocovariance matrix of size n defined as $\mathbb{E}[(y_t^n - \mathbb{E}[y_t^n])(y_t^n - \mathbb{E}[y_t^n])^T]$. It is convenient to rewrite a VAR model of order p in Equation (4) as a VAR(1) model, $y_t = Ay_{t-1} + e_t$, where $y_t \in \mathbb{R}^{dp}, e_t \in \mathbb{R}^{dp}$ are defined as $y_t = \left(x_t, x_{t-i}, \cdots, x_{t-p+1}\right)^T$, $e_t = \left(\epsilon_t, 0, \cdots, 0\right)^T$, and $A \in \mathbb{R}^{dp\times dp}$ is a *(multi) companion matrix* defined as:

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix}. \tag{5}$$

[1]The results for simultaneous interventions are qualitatively similar to those of interventions on single variables, and for ease of exposition, we present our discussion in the latter case.

The eigenvalues of the multi-companion matrix A fully characterize the stability and stationarity of the VAR process. For a VAR(p) process to be weakly stationary, that is for the mean and the covariance of the process to not change over time, the eigenvalues of A, which satisfy

$$\det|\mathbb{I}_d\lambda^p - A_1\lambda^{p-1} - A_2\lambda^{p-2} - \cdots - A_p| = 0, \quad (6)$$

are constrained to not lie on the unit circle. If the magnitude of all the eigenvalues are $|\lambda_i| < 1$, then the process is stable, that is, its values do not diverge (Lütkepohl 2013).

Causal Models. Under the assumptions of causal sufficiency and absence of contemporaneous influences, a causal interpretation of the VAR model in (4) as structural equations naturally yields the corresponding causal model. We consider the family of all VAR models as our function class $\mathcal{F}$ of statistical and causal estimators.

3 CAUSAL GENERALIZATION FOR VAR

In this section, we present causal generalization bounds for the family of VAR models under atomic interventions. We first provide an overview of our results in the more general case of VAR(p) models and later provide a thorough interpretation of the results, often by deriving simplified versions of the results for AR(p) models. We begin by providing an exact characterization of the difference in statistical and causal errors in terms of the model and estimated parameters and the autocovariance matrix of the underlying process.

Lemma 1 (Difference in Causal and Statistical errors (VAR)). *Consider a vector-valued time series $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, following a VAR(q) process parameterized by $\{A_1, A_2, \cdots A_q\}$. Let $\nu = \max\{p, q\}$. For any VAR(p) model f with parameters $\{\hat{A}_1, \hat{A}_2, \cdots \hat{A}_p\}$,*

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| = 2\left|(A^\omega_{ii} - \hat{A}^\omega_{ii})\sum_{k\neq i}^{d\nu}(A^\omega_{ik} - \hat{A}^\omega_{ik})\Sigma^\nu_{ik}\right|,$$

where Σ^ν denotes the autocovariance matrix of x_t of size ν, A is a multi-companion matrix of the form described in (5) with the first d rows populated by $\{A'_1, A'_2, \cdots A'_\nu\}$, with A'_l defined as A_l for all $l \leq p$ and as $\mathbf{0}_{d\times d}$ for all $l > p$. $\hat{A}$ is analogously defined.

Building on Lemma 1, we establish that the condition number of the autocovariance matrix of the underlying process controls causal generalizability from the *observational* to *interventional distributions*.

Proposition 1 (Stability Controls Causal Generalization (VAR)). *Let $\{x_t\}_{t\in\mathbb{Z}}$ follow a VAR(q) process for some $q \in \mathbb{N}$. For any VAR(p) model,*

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| \leq (2\kappa(\Sigma^\nu) - 1)(\mathcal{S}_\omega - \sigma^2_\epsilon), \quad (7)$$

where $\kappa(\Sigma^\nu)$ denotes the condition number of the autocovariance matrix Σ^ν. Further, one can construct processes where equality holds upto a small constant factor.

The result states that the difference in expected causal and statistical errors is controlled by the *condition number* of the autocovaraince matrix of size $\max\{p, q\}$. It also states that without incorporating additional information, one cannot obtain a much tighter bound which is also verified by our experiments in Section 4. The condition number of the autocovariance matrix can get arbitrarily large as the process gets closer to the boundary of the stability domain. This result therefore shows that even for very simple classes of forecasting models, causal interpretations can get challenging. We later provide a detailed interpretation of this result and provide an explicit bound on $\kappa(\Sigma_\nu)$ in terms of the stability parameter for AR(p) models (Corollary 2).

Proposition 1 allows us to employ generalization bounds for time-series (Yu 1994; Meir 2000; Mohri et al. 2009; McDonald et al. 2017) to derive finite-sample *causal generalization bounds* for VAR models. In particular, we utilize Rademacher complexity bounds for generalization in time-series under mixing conditions (Mohri et al. 2009) to derive Theorem 1.

Theorem 1 (Finite sample bounds for VAR(p) models). *Let $\mathcal{F}$ denote the family of all VAR models of dimension d and order p. For any $n > \max\{p, q\} \in \mathbb{N}$, let $\mu, m > 0$ be integers such that $2\mu m = n$ and $\delta > 2(\mu-1)\rho^m$ for a fixed constant $0 < \rho < 1$ determined by the underlying process. Let $\{x_1, x_2, \cdots x_n\} \in \mathbb{R}^d$ be a finite sample drawn from a VAR(q) process. Then, simultaneously for every $f \in \mathcal{F}$, under the square loss truncated at M, with probability at least $1 - \delta$,*

$$\mathcal{G}_{\omega,i} \leq \zeta\hat{\mathcal{S}}_\omega + \zeta\widehat{\mathfrak{R}}_\mu(\mathcal{F}) + 3\zeta M\sqrt{\frac{\log\frac{4}{\delta'}}{2\mu}} \quad (8)$$

where $\zeta = 2\kappa(\Sigma^\nu)$, $\delta' = \delta - 2(\mu-1)\rho^m$, and $\widehat{\mathfrak{R}}_\mu(\mathcal{F})$ denotes the empirical Rademacher complexity of $\mathcal{F}$.

Our causal generalization bound in Theorem 1 suggests that, given sufficiently many samples, the true causal error can be guaranteed to be close to empirical statistical error if our VAR models come from a class with a small Rademacher complexity, particularly when the process is associated with a small stability parameter.

We now focus on providing a detailed interpretation of our results. First, we take a minor detour to present a technical result (Lemma 2) which is useful both in deriving some of our main results as well as in interpreting them.

Lemma 2 (Expressing powers of a companion matrix using symmetric polynomials). *For a companion matrix A with distinct eigenvalues, for any $k \in [p]$, the $(1, k)$th element of A^j, can be expressed using Schur polynomials of the eigenvalues $\lambda = \{\lambda_1, \lambda_2, \cdots \lambda_p\}$ of A, that is, $A^j_{1,k} = S_{j,k}(\lambda)$, where $S_{j,k}(\lambda)$ refers to the Schur polynomial indexed by $K = \{j, 1, \cdots k-1 \text{ times} \cdots, 1, 0, \cdots, 0\}$.*

Lemma 2 shows that the coefficients of the powers of a companion matrix can be fully characterized using symmetric Schur polynomials of its eigenvalues. A good overview of these polynomials can be found in Chaugule et al. (2019). An advantage of expressing the coefficients using symmetric Schur polynomials is that these polynomials have been a subject of extensive research in combinatorics and an equivalence between several alternate definitions has been established. To name a few, Cauchy's bialternant expression, (Cauchy 1815; Jacobi 1841), the combinatorial formula (Macdonald 1998) or Jacobi–Trudi identity (Jacobi 1841) are all equivalent ways to define Schur polynomials. It is therefore possible and often beneficial to choose the definition that yields the most useful notion for the context. We utilize this connection to interpret our results. First, for easier interpretation, we simplify Lemma 1 to the following result for scalar AR models.

Corollary 1 (**Difference in Causal and Statistical errors (AR)**). *Let $\{x_t\}_{t\in\mathbb{Z}}$ follow an AR(q) process. Then, for any AR(p) model with parameters $\hat{A}$,*

$$|\mathcal{G}_{do_\omega} - \mathcal{S}_\omega| = 2\left|(A^\omega_{11} - \hat{A}^\omega_{11})\sum_{k=2}^{\nu}(A^\omega_{1k} - \hat{A}^\omega_{1k})\gamma_{k-1}\right|, \quad (9)$$

where, for any $k \in \mathbb{N}$, γ_k denotes the autocovariance of $\{x_t\}_{t\in\mathbb{Z}}$ with lag k. A and $\widehat{A}$ are the corresponding companion matrices of the model and estimated parameters as defined in Lemma 1.

Lemma 1 identifies factors that control causal generalizability. We now describe them.

Correlations control causal generalizability. Recall our motivating example of the two highly correlated time-series where the casual and statistical errors diverge. Intuitively, one would therefore expect that large correlations among time series potentially induce large differences between observational and interventional distributions. The quantitative dependence of causal generalizability on the correlation structure of the process is, however, less obvious. Lemma 1 confirms the intuition and shows that correlations between the intervened time-series $x_{t-\omega,i}$ across both the components and time instances in $\mathbf{x}_{t-\omega}$ control generalizability from observational to the interventional distributions.

High-dimensional and higher-order processes can hurt generalization. For high-dimensional processes it is not unlikely to have strong correlations across components, which may obscure causal relations in the same way as strong correlations across time does for univariate processes. Lemma 1 also supports this intuition and shows that strong correlations across components as well as time instances play a role. With increasing order or dimension of the processes, larger orders of covariances across time and dimensions could entail poor causal generalizability.

Dependence on ω. The dependence of the error on ω arises through the elements of the matrix power A^k. A simple computation shows that, even for an AR(2) model, the dependence of these coefficients on the model parameters is asymmetric and highly intricate. However, using the Weyl formulation of Schur polynomials, we have that for any AR(p) model, the coefficients A^ω_{1k} can be expressed as $A^\omega_{1k} = (-1)^{k+1}\frac{\sum_{i=1}^{p}\lambda_i^{p+\omega-1}e_k(\lambda_i)}{\det\left|\left\{\lambda_k^{p-k'}\right\}_{k,k'\in[p]}\right|}$, where $e_k(\lambda_i)$ refers to the elementary symmetric polynomial of order k and with variables $\{\lambda_1, \cdots \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_p\}$. While this is not the most interpretable definition per se, the dependence of the coefficients on ω is easily understood and it is easy to verify that if the underlying model as well as the estimated model are both stable ($|\lambda| < 1$), the coefficients and hence the difference in errors exponentially decays with interventions arbitrarily in the past of the target variable and if either of the process is not stable ($|\lambda| > 1$), the difference can indeed diverge.

Proposition 1 allows us to obtain a high-level perspective on causal generalizability. It states that the condition number of the autocovariance matrix controls causal generalizability. Both the maximum and the minimum eigenvalue of the autocovariance matrix (and hence the condition number) can be used as a measure of stability and hence determine the strength of correlation of the underlying process (Basu et al. 2015; Melnyk et al. 2016). As the process gets closer to the boundary of stability domain, the autocovariance matrix gets singular and hence the condition number of the autocovariance matrix can get arbitrarily large. Proposition 1, therefore, can be interpreted as if the underlying process gets closer to the boundary of the stability domain the causal and statistical errors can diverge.

For intuition, let us revisit our motivating example from the introduction with strongly correlated observations in an $AR(p)$ process. Let, without loss of generality $p = q$. Introducing the vectors $a := (a_1, a_2, \ldots, a_p)$ and $\hat{a} := (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p)$ and the covariance matrix $\Sigma_p = \Sigma_{\max\{p,q\}}$. Then the quotient between causal and statistical error for predicting one time step ahead i.e. ($\omega = 1$) reads:

$$\frac{\mathcal{G}_{do_\omega}}{\mathcal{S}_\omega} = \frac{(\hat{a}-a)^T(\hat{a}-a) + \sigma_\epsilon^2}{(\hat{a}-a)^T\Sigma_p(\hat{a}-a) + \sigma_\epsilon^2}, \quad (10)$$

Where we have assumed X_t to have unit variance without loss of generality. The quotient is maximized if $(\hat{a} - a)$ is a multiple of the eigenvector to the smallest eigenvalue of Σ_p. This aligns with the intuition that causal loss diverges when the auto-covariance matrix gets singular. Moreover, we see that the vector $(\hat{a} - a)$ can be large with little observable effect when it mainly consists of eigenvectors with small eigenvalues of Σ_p. In the extreme case, if the minimum eigenvalue of the autocovariance matrix is 0, it is possible to arbitrarily deviate from the true model parameters along the direction of the corresponding eigenvector which can significantly affect the causal error without affecting the statistical

error at all. For an $AR(2)$ process, for instance, we obtain $\Sigma_p = \begin{pmatrix} 1 & a_1/(1-a_2) \\ a_1/(1-a_2) & 1 \end{pmatrix}$, which becomes singular for $a_1 = \pm(1-a_2)$ which indeed is the boundary of the stability domain (see for example, Lütkepohl (2009)). This is the limit in Section 1 where $X_t = \pm X_{t-1}$. The eigenvector for eigenvalue 0 reads $(1, \mp 1)$. Accordingly, the quotient (10) diverges when $\hat{a}$ differs from a by $(1, \mp 1)$.

This further highlights that even for simple classes of forecasting models and with simplifying assumptions such as causal sufficiency, causal risks may even diverge from statistical risks. To show this formally, by means of Lemma 2, we can derive an explicit upper bound on the condition number of the autocovariance matrix $\kappa(\Sigma_{\max\{p,q\}})$ for AR(p) models and arrive at Corollary 2.

Corollary 2 (**Stability Controls Causal Generalization (AR)**). *Consider an AR(q) process, such that eigenvalues of its companion matrix satisfy $|\lambda| < \delta < 1$. For any AR(q) model f,*

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| \leq K_p \mathcal{S}_\omega(f) \nu (1+\delta)^{2\nu}/(1-\delta^2), \qquad (11)$$

where K_p is some finite constant that depends on the order p of the underlying process.

The bound in Corollary 2 is elegant due to its simplicity and generality. However, the cost of generality of the bound that relies only on the stability parameter is clearly that it cannot explain the variations in behavior exhibited by individual processes with the same stability parameter. For instance, consider an AR(2) model with parameters a_1 and a_2 with $a_2 \approx 0$ so that it is essentially an AR(1) model. Then, it is easy to verify that $\lambda_2 \approx 0$. The combinatorial definition of the Schur polynomials (Macdonald 1998) allows us to express the coefficients as follows: $A_{11}^\omega = \sum_{i=0}^{\omega} \lambda_1^{\omega-i}\lambda_2^i, \quad A_{12}^\omega = \sum_{i=1}^{\omega-1} \lambda_1^{\omega-i}\lambda_2^i$. Combining this with Corollary 1, it is easy to see that if the estimated model is also close to AR(1), then the coefficients A_{12}^ω and $\widehat{A}_{12}^\omega$ and hence the difference in statistical and causal errors is close to 0. The bound in (11) which relies on the stability parameter does not capture this. For tighter bounds that utilize additional information about the spectrum of the companion matrix, we can exploit the connection to Schur polynomials to arrive at the following bound.

$$|\mathcal{G}_{\omega,i} - \mathcal{S}_\omega| \leq K_{p,q} \max\left\{\delta, \widehat{\delta}\right\}^\omega \sum_{k=2}^{\nu} \left(S_{\omega k}^{\lambda} - S_{\omega k}^{\hat{\lambda}}\right)\gamma_{k-1},$$

where $K_{p,q}$ is a constant that depends on p, q, δ and $\widehat{\delta}$ are the stability parameters of the true and estimated processes respectively and λ and $\widehat{\lambda}$ denote the set of eigenvalues of A and $\widehat{A}$ respectively.

4 SIMULATIONS

To verify the practical behavior of causal and statistical risks, we provide some simple simulations to study the errors of different estimators under AR processes. For each presented plot, we draw parameters for 10,000 stationary $AR(p)$ processes. We draw the coefficients of each process independently and uniformly from $[-2, 2]$ and reject sets of parameters that yield a non-stationary process. For each process, we draw a training sample with 100 timesteps and a test sample with 1000 timesteps. For all figures in the main paper we set $\omega = 1$. To estimate the coefficients we use Ordinary Least Squares (OLS). In Appendix E we provide additional plots with hidden confounder, as well as varying order, sample size, ω and other estimators: Ridge, Lasso, and Elastic Net regressors. OLS minimizes the empirical statistical error, that is, $\sum_{y_i, \hat{y}_i}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ denotes the model prediction with estimated parameters $\hat{a}$.

In line with our theoretical results, we find that even for simple scalar AR processes of small orders, the causal error of the estimators is often several times larger than the statistical error (see Figure 2). In Figure 3 we sorted the randomly drawn datasets by their autocorrelation (measured by the condition number κ of the autocorrelation matrix) and split the sorted list into buckets of 500 dataset. For each we calculated the maximum, mean and 90% quantile of the difference in causal and statistical error for the OLS and Ridge estimators. The plots corresponding to the other estimators are provided in Appendix E We can see that upto constant factors, our theoretical finite sample causal generalization bound matches the difference in causal and statistical risks observed empirically.

5 EXPERIMENTS ON REAL DATA

Data. We conduct experiments on the m4 hourly dataset (Makridakis et al. 2018) that includes timeseries from a diverse set of sources. The datasets has a hourly frequency and a prediction length of 48. To create an interventional distribution without a generative model, for each time series we replace the last time step prior to the evaluation window by sampling at random either from all time-series at that time step (referred to as *across-ts*) or from previous values of the same time series (referred to as *within-ts*).

Models. We include three popular deep neural network architectures in our evaluation. DeepAR consists of an RNN that takes the previous time steps as inputs and predicts the parameters of an auto-regressive model (Gasthaus et al. 2019). Wavenet is a hierarchical CNN developed for speech-to-text (Oord et al. 2016). Transformer is an attention-based deep neural network widely applied to NLP tasks including translation (Vaswani et al. 2017).

Metrics. For the observational distribution, we compute

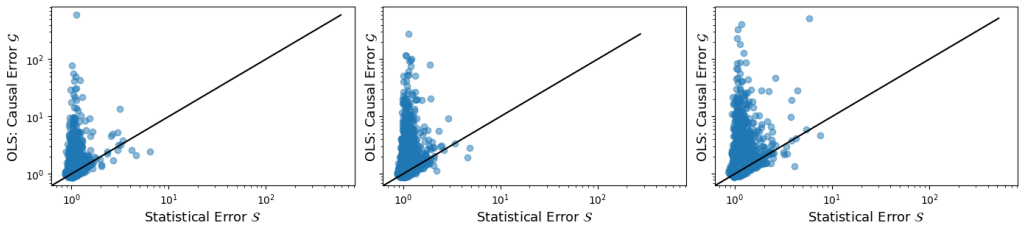

Figure 2: The causal error $\mathcal{G}$ versus the statistical error $\mathcal{S}$ for AR(p) processes with $p = 3, 5, 7$.

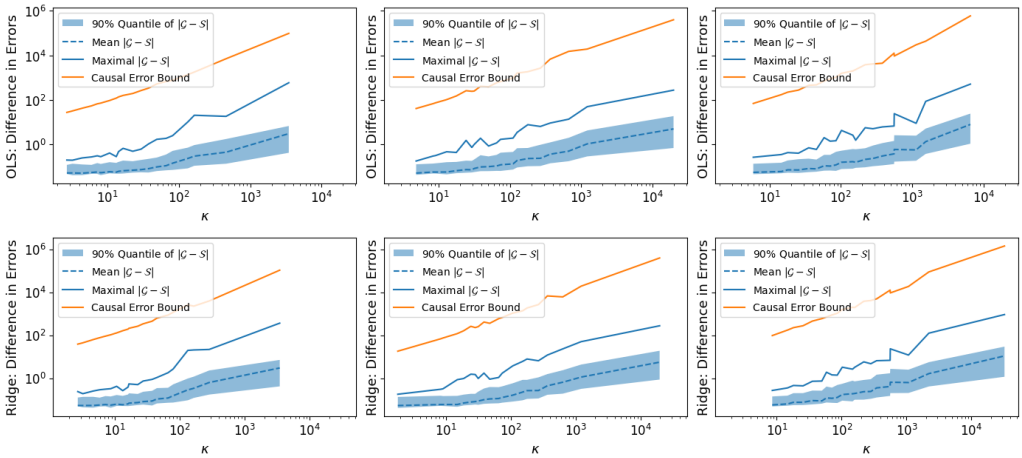

Figure 3: The maximal difference between statistical error $\mathcal{S}$ and causal error $\mathcal{G}$ as well as an estimate for the generalization bound in Theorem 1 for increasing condition number κ for process orders $p = 3, 5, 7$ (from left to right). The maximum is taken over 500 datasets with the closest κ. Our theoretical bounds (orange) closely match the empirical evaluations up to constant factors (blue).

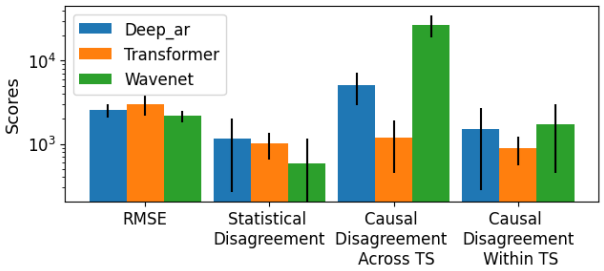

Figure 4: Results of the evaluation of three different deep neural network architectures on the m4hourly dataset. The RMSE is computed comparing prediction on the observational data against the ground truth and comparing the predictions across two models of the same architecture on the observational and interventional distributions. The results are averaged over 5 runs of training and evaluation and include standard deviation.

the root-mean-square error (RMSE) comparing average prediction for each time point with the ground truth in the evaluation set. For the interventional distribution we are lacking ground truth. Therefore, we train two separate models and compute their disagreement. This disagreement is a measure of uncertainty introduced by the randomness in the training and evaluation procedure. Here, however, we use it to approach the causal risk, that we cannot compute directly. If the disagreement is high on the interventional distribution at least one of the models must have a high causal risk. Specifically, we compute the RMSE across the two models' mean forecasts. For comparison, we also included this disagreement measure for examples from the observational distribution. Finally, to explore the relationship between causal forecasting error and uncertainty, we also compute the width of the 80% prediction interval for both the observational and interventional distribution.

The experiments were conducted using GluonTS (Alexandrov et al. 2019) with default hyperparameters on instances with 4 virtual CPUs and a 2.9 GHz processor.

Limitations. The dataset and models have clear shortcomings. Likely, the dataset is not causally sufficient. Also, we did not tune the models. Moreover, we are lacking samples from the marginal distribution for the interventions and groundtruth on what happens under these interventions. Nevertheless, we hope to get a sense for how popular deep learning networks can behave on real data for relevant prediction tasks under interventions.

Results. Figure 4 shows the results of the metrics when we evaluate the models on the datsets for both observation

Model	observ.	across-ts interv.	within-ts interv.
DeepAR	940.0 ± 126.2	1329.2 ± 187.5	953.1 ± 124.2
wavenet	1253.9 ± 96.6	3444.7 ± 649.4	1612.7 ± 257.7
transformer	1259.3 ± 139.3	1355.1 ± 129.6	1255.7 ± 139.3

Table 1: 80% prediction width for observational and interventional forecasts. Averaged over 5 runs with std.

and interventional distributions. We see that the causal disagreement between two models of the same architecture and hyper-parameters can be much higher than their disagreement on the observational distribution. While there are only smaller differences in the statistical risk between the model architectures, their causal disagreement differs more. Overall, the the causal disagreement can be high, which implies high causal risk, but it varies across datasets and model architectures. Wavenet's disagreement is an order of magnitude larger when sampling interventions from other time-series. For transformer models their interventional disagreement is close to the observational one.

Uncertainty. When we compare the width of the 80% interval of predictions in Table 1 we see that this uncertainty measure is higher for the the interventional distribution compared to the observational one. Moreover, directionally it relates to the causal disagreement across models. Unlike the disagreement that requires a second model to be trained, this uncertainty measure is readily available from the predicted forecasts.

6 RELATED WORK

Our work intersects with domain adaption, RL, and treatment effect estimation, reviewed separately below.

Domain Adaptation. The literature that is perhaps most relevant to our context is that of learning theory for domain adaptation, in particular, for covariate shift. Theoretical analysis of domain adaptation when labelled samples from the source distribution and unlabelled samples from the target distribution are generated i.i.d was initiated by Ben-David et al. (2007), who provided VC bounds for binary classification under covariate shifts based on a *discrepancy measure* $d_{\mathcal{F}}$ between source and target distributions that depends on the hypothesis class $\mathcal{F}$ and is estimable from finite samples. Mansour et al. (2009) extended the work to the context of regression in the i.i.d setting by adapting the discrepancy measure for more general loss functions and by providing tighter, data-dependent Rademacher bounds. Despite the i.i.d assumption that is necessary to derive their finite-sample bounds, the results in Mansour et al. (2009) are perhaps the most relevant to our setting. We can utilize one of the main results from Mansour et al. (2009, Theorem 8) which does not rely on the i.i.d assumption to arrive at the following population-level bound for our setting: $|\mathcal{G}_{\omega,i}(f, f^*) - \mathcal{S}_\omega(f, f^*)| \leq \sup_{f,f' \in \mathcal{F}} |\mathcal{G}_{\omega,i}(f, f') - \mathcal{S}_\omega(f, f')|$. These bounds are clearly non-informative in our context since they do not incorporate structural knowledge of the class of interventional distributions.

Estimation of Treatment Effects. A related problem is that of estimating treatment effects in the potential outcomes framework (Hill et al. 2006; Shi et al. 2019), where the goal is to estimate the effects of binary-valued treatments from observational data under a multivariate confounding model. Our setting is more general in that variables in the multivariate process can take a continuum of interventions and play a multiplicity of roles — each variable plays the role of treatment, confounder, and the target variable. Of particular relevance is the work of Shalit et al. (2017) and Johansson et al. (2020), who prove generalization error bounds on estimating individual-level treatment effects in terms of standard generalization error and a distance measure between the treated and control distributions. This result is similar to domain adaptation bounds in Ben-David et al. (2007) and Mansour et al. (2009) and may be interpreted as causal learning theory in the sense of our paper.

Reinforcement Learning. The ratio of observational versus interventional densities in our setting play a similar role as the state density ratio in off-policy evaluation in reinforcement learning(RL) (Bennett et al. 2021). In RL, however, the clear separation between the state of actions and the state space acted on admits techniques that we do not see for our problem, e.g., deconfounding (Hatt et al. 2021), or learning representations of the history that are independent of the actions (Bica et al. 2020), which overcomes the problem of high inverse probability weightings (Lim et al. 2018).

7 DISCUSSION AND CONCLUSION

Our work highlights that even for very simple models and even under simplifying assumptions such as causal sufficiency, causal and statistical errors can diverge. It emphasizes the need for providing guarantees for causal generalization in a similar vein as providing guarantees for statistical learning. To this end, we initiate a first analysis in this direction by introducing a framework for causal learning theory for forecasting and providing conditions under which one can guarantee generalization in the causal sense for the class of VAR models. We hope that this work inspires more theoretical work that allows certifying the validity of the causally interpreting forecasting models.

Our theoretical as well as empirical results challenge the causal interpretation of forecasting models used in practice which are typically far more complex. Our experiments show that causal disagreement can be high for some models which implies a high causal risk. This cautions against the use of statistical deep learning models for causal forecasting. The difference we observe in causal disagreement across models motivates further development of specific model architectures suitable for causal forecasting. For existing models, the uncertainty measure considering the width of the prediction interval can be an indicator for causal risk.

REFERENCES

Alexandrov, Alexander, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang (2019). *GluonTS: Probabilistic Time Series Models in Python*. arXiv: 1906.05264 [cs.LG].

Basu, Sumanta and George Michailidis (2015). "Regularized estimation in sparse high-dimensional time series models". In: *The Annals of Statistics* 43.4, pp. 1535–1567.

Ben-David, Shai, John Blitzer, Koby Crammer, Fernando Pereira, et al. (2007). "Analysis of representations for domain adaptation". In: *Advances in neural information processing systems* 19, p. 137.

Bennett, Andrew, Nathan Kallus, Lihong Li, and Ali Mousavi (2021). "Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1999–2007.

Bica, Ioana, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar (2020). "Estimating counterfactual treatment outcomes over time through adversarially balanced representations". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=BJg866NFvB.

Brand, Louis (1964). "The companion matrix and its properties". In: *The American Mathematical Monthly* 71.6, pp. 629–634.

Brockwell, Peter J, Richard A Davis, and Stephen E Fienberg (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

Cauchy, Augustin Louis (1815). "M 'e memory on functions which can obtain only two equal values é and of opposite signs as a result of the transpositions op 'e r é es between the variables that they contain". In: *Journal de l'Ecole polytechnique* 10.17, pp. 29–112.

Chaugule, Prasad, Mrinal Kumar, Nutan Limaye, Chandra Kanta Mohapatra, Adrian She, and Srikanth Srinivasan (2019). "Schur Polynomials do not have small formulas if the Determinant doesn't!" In: *arXiv preprint arXiv:1911.12520*.

Davison, Edward (1976). "The robust control of a servomechanism problem for linear time-invariant multivariable systems". In: *IEEE transactions on Automatic Control* 21.1, pp. 25–34.

De Marchi, Stefano (2001). "Polynomials arising in factoring generalized Vandermonde determinants: an algorithm for computing their coefficients". In: *Mathematical and computer modelling* 34.3-4, pp. 271–281.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Faloutsos, Christos, Jan Gasthaus, Tim Januschowski, and Yuyang Wang (2018). "Forecasting Big Time Series: Old and New". In: 11.12. ISSN: 2150-8097. DOI: 10.14778/3229863.3229878. URL: https://doi.org/10.14778/3229863.3229878.

Fisk, Steve (2005). "A very short proof of Cauchy's interlace theorem for eigenvalues of Hermitian matrices". In: *arXiv preprint math/0502408*.

Gasthaus, Jan, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski (Apr. 2019). "Probabilistic Forecasting with Spline Quantile Function RNNs". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1901–1910. URL: https://proceedings.mlr.press/v89/gasthaus19a.html.

Grabowski, Wojciech and Aleksander Welfe (2020). "The Tobit cointegrated vector autoregressive model: An application to the currency market". In: *Economic Modelling* 89, pp. 88–100.

Hatt, Tobias and Stefan Feuerriegel (2021). *Sequential Deconfounding for Causal Inference with Unobserved Confounders*. arXiv: 2104.09323 [stat.ME].

Hill, Jennifer and Jerome P Reiter (2006). "Interval estimation for treatment effects using propensity score matching". In: *Statistics in medicine* 25.13, pp. 2230–2256.

Horn, Roger A and Charles R Johnson (2012). *Matrix analysis*. Cambridge university press.

Jacobi, Carl Gustav Jakob (1841). "De functionibus alternantibus earumque divisione per productum e differentiis elementorum conflatum." In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1841.22, pp. 360–371.

Januschowski, Tim, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot (2020). "Criteria for classifying forecasting methods". In: *International Journal of Forecasting* 36.1. M4 Competition, pp. 167–177. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.05.008. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301529.

Janzing, D., L. Minorics, and P. Bloebaum (Aug. 2020). "Feature relevance quantification in explainable AI: A causal problem". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, pp. 2907–2916.

Johansson, Fredrik D, Uri Shalit, Nathan Kallus, and David Sontag (2020). "Generalization bounds and representation learning for estimation of potential outcomes and causal effects". In: *arXiv preprint arXiv:2001.07426*.

Lim, Bryan, Ahmed Alaa, and Mihaela van der Schaar (2018). "Forecasting Treatment Responses over Time Using Recurrent Marginal Structural Networks". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 7494–7504.

Lundberg, S. and S. Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 4765–4774.

Lütkepohl, Helmut (2009). "Econometric analysis with vector autoregressive models". In: *Handbook of computational econometrics*, pp. 281–319.

– (2013). "Vector autoregressive models". In: *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing.

Macdonald, Ian Grant (1998). *Symmetric functions and Hall polynomials*. Oxford university press.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilis Assimakopoulos (June 2018). "The M4 Competition: Results, findings, conclusion and way forward". In: *International Journal of Forecasting* 34. DOI: 10.1016/j.ijforecast.2018.06.001.

Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh (2009). "Domain adaptation: Learning bounds and algorithms". In: *arXiv preprint arXiv:0902.3430*.

McDonald, Daniel J, Cosma Rohilla Shalizi, and Mark Schervish (2017). "Nonparametric risk bounds for time-series forecasting". In: *The Journal of Machine Learning Research* 18.1, pp. 1044–1083.

Meir, Ron (2000). "Nonparametric time series prediction through adaptive model selection". In: *Machine learning* 39.1, pp. 5–34.

Melnyk, Igor and Arindam Banerjee (2016). "Estimating structured vector autoregressive models". In: *International Conference on Machine Learning*. PMLR, pp. 830–839.

El-Mikkawy, Moawwad EA (2003). "Explicit inverse of a generalized Vandermonde matrix". In: *Applied mathematics and computation* 146.2-3, pp. 643–651.

Mohri, Mehryar and Afshin Rostamizadeh (2009). "Rademacher Complexity Bounds for Non-I.I.D. Processes". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc.

Mokkadem, Abdelkader (1988). "Mixing properties of ARMA processes". In: *Stochastic processes and their applications* 29.2, pp. 309–315.

Molnar, C. (2019). *Interpretable Machine Learning*. Molnar, C. URL: https://christophm.github.io/interpretable-ml-book/.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv: 1609.03499 [cs.SD].

Pearl, Judea (2009). *Causality*. Cambridge university press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference*. The MIT Press.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3, pp. 1181–1191. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301888.

Shalit, Uri, Fredrik D Johansson, and David Sontag (2017). "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Shi, Claudia, David M Blei, and Victor Veitch (2019). "Adapting neural networks for the estimation of treatment effects". In: *arXiv preprint arXiv:1906.02120*.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*. New York, NY: Springer-Verlag.

Sugiyama, Masashi and Motoaki Kawanabe (2012). *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press. ISBN: 0262017091.

Valdés-Sosa, Pedro A, Jose M Sánchez-Bornot, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García, and Erick Canales-Rodríguez (2005). "Estimating brain functional connectivity with sparse multivariate autoregression". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457, pp. 969–981.

Varga, Richard S (2010). *Geršgorin and his circles*. Vol. 36. Springer Science & Business Media.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, Jiaxuan, Jenna Wiens, and Scott Lundberg (2020). *Shapley Flow: A Graph-based Approach to Interpreting Model Predictions*. arxiv:2010.14592.

Yu, Bin (1994). "Rates of convergence for empirical processes of stationary mixing sequences". In: *The Annals of Probability*, pp. 94–116.

Zivot, Eric and Jiahui Wang (2006). "Vector autoregressive models for multivariate time series". In: *Modeling Financial Time Series with S-Plus®*, pp. 385–429.

A BACKGROUND

Notation. We recall the notation and some key definitions here for the reader's convenience. For any stochastic process $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, we use $\mathbf{x}_{t-\omega}^n = \{x_{t-\omega-n+1}, \cdots, x_{t-\omega-1}, x_{t-\omega}\}$ to denote the *set* of $x_{t-\omega}$ and the $n-1$ variables in the past of $x_{t-\omega}$. We distinguish this from y_t^n which denotes the *vector* $\left(x_t, x_{t-1}, \cdots, x_{t-n+1}\right)^T \in \mathbb{R}^{nd}$. When it is clear from context, to reduce cumbersome notation, we simply use y_t. For any random variable x, $\mathbb{E}[x]$ denotes its expectation. For any matrix A, we use $A_{i:}$ and $A_{:j}$ to denote the ith row and jth column of A respectively. We use A_{1k}^j to denote the $(1,k)$th element of A^j. For any vector x_t at time t, we use $x_{t,i}$ to denote the ith element of x_t. We use $\lambda_{\max}(A), \lambda_{\min}(A), \kappa(A)$ to denote the maximum and minimum eigenvalues and the condition number of A respectively, where $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$. $\mathbb{I}_p$ denotes the identity matrix of size p, $\mathbb{N}, \mathbb{Z}$ denote the set of natural numbers and integers respectively and $[n]$ denotes the set $\{1, 2, \cdots n\}$.

Definition A.1 (Vector Autoregressive Model). A vector autoregressive model (VAR(p)) of dimension d and order p is defined as

$$x_t = A_1 x_{t-1} + A_2 x_{t-2} + \cdots A_P x_{t-p} + \epsilon_t, \tag{12}$$

where $x_t \in \mathbb{R}^d$ is a vector-valued time-series, for all $i \in [p]$, $A_i \in \mathbb{R}^{d\times d}$ are the coefficients of the VAR model, and $\epsilon_t \in \mathbb{R}^d$ denotes the noise vector such that $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t \epsilon_{t+h}^T] = \Sigma_\epsilon$ if $h = 0$ and 0 otherwise. For some $\sigma_\epsilon^2 > 0$, we simply set $\Sigma_\epsilon = \sigma_\epsilon^2 \mathbb{I}$ for enhanced readability. Our results can be easily generalized to arbitrary covariance matrices by means of the spectral properties $(\lambda_{\min}, \lambda_{\max})$ of Σ_ϵ.

Definition A.2 (Weak Stationarity). A stochastic process $\{x_t\}_{t\in\mathbb{Z}}$ is weakly stationary if the mean and the covariance of the process does not change over time, that is, for all $t, \tau \in \mathbb{Z}$

$$\mathbb{E}[x_t] = \mathbb{E}[x_{t+\tau}], \;\; \mathbb{C}_x(t, t+\tau) = \mathbb{C}_x(0, \tau), \tag{13}$$

where $\mathbb{C}_x(t, t+\tau) = \mathbb{E}[(x_t - \mathbb{E}[x_t])(x_{t+\tau} - \mathbb{E}[x_{t+\tau}])]$ denotes the autocovariance function.

The autocovariance matrix of $\{x_t\}_{t\in\mathbb{Z}}$ plays a central role in our results and analysis. For any $n \in \mathbb{N}$, we use Σ_n to denote the autocovariance matrix of size n defined as $\mathbb{E}[(y_t^n - \mathbb{E}[y_t^n])(y_t^n - \mathbb{E}[y_t^n])^T]$.

It is often quite convenient to rewrite a VAR model of order p in Equation (12) as a VAR(1) model, $y_t = Ay_{t-1} + e_t$, where $y_t \in \mathbb{R}^{dp}, e_t \in \mathbb{R}^{dp}$ are defined as $y_t = \left(x_t, x_{t-i}, \cdots, x_{t-p+1}\right)^T$, $e_t = \left(\epsilon_t, 0, \cdots, 0\right)^T$, and $A \in \mathbb{R}^{dp\times dp}$ is a *(multi) companion matrix* defined as:

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix}. \tag{14}$$

The eigenvalues of the multi-companion matrix A fully characterize the stability and stationarity of the VAR process. For a VAR(p) process to be weakly stationary, the eigenvalues of A, which satisfy

$$\det|\mathbb{I}_d \lambda^p - A_1 \lambda^{p-1} - A_2 \lambda^{p-2} - \cdots - A_p| = 0, \tag{15}$$

are constrained to not lie on the unit circle. If the magnitude of the eigenvalues are $|\lambda_i| < 1$ for all $i \in [dp]$, then the underlying process is stable, that is, its values do not diverge (Lütkepohl 2013).

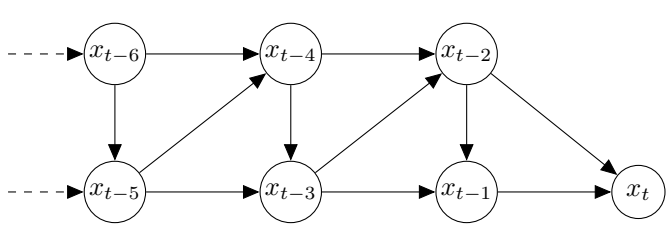

Figure 5: Causal DAG of an AR(2) model

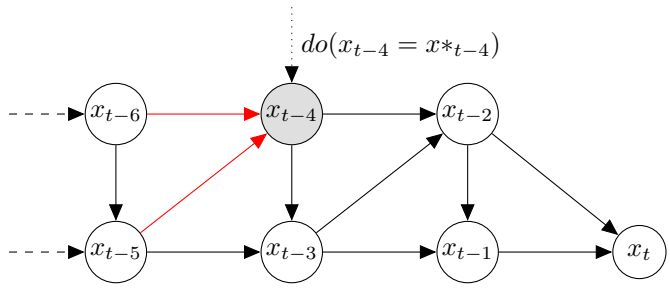

Figure 6: Graphical representation of the effect of an intervention $do(x_{t-4} = x*_{t-4})$ on an AR(2) model. Incoming edges into x_{t-4} are removed in the new DAG which are in red.

Definition A.3 (**Empirical Rademacher Complexity**). Given a finite sample $X = \{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^d$, the empirical Rademacher complexity of a hypothesis class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ is defined as:

$$\hat{\mathfrak{R}(\mathcal{F})} = \frac{2}{n}\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \sigma_i f(x_i)| \right],$$

where $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n)$ and for all $i \in [n]$, σ_i are independent random variables drawn from the Rademacher distribution, that is, a uniform distribution over $\{-1, +1\}$.

B PROOFS OF MAIN RESULTS

Lemma 3 (**Expressing powers of a companion matrix using symmetric polynomials**). *For a companion matrix A with distinct eigenvalues, for any $k \in [p]$, the $(1, k)$th element of A^j, can be expressed using Schur polynomials of the eigenvalues $\lambda = \{\lambda_1, \lambda_2, \cdots \lambda_p\}$ of A, that is, $A^j_{1,k} = S_{j,k}(\lambda)$, where $S_{j,k}(\lambda)$ refers to the Schur polynomial indexed by $K = \{j, 1, \cdots k-1$ times $\cdots, 1, 0, \cdots, 0\}$.*

Proof. Assuming that the eigenvalues of a companion matrix A are distinct, it can be diagonalized as $A = V\Lambda V^{-1}$, where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$ is the diagonal matrix of eigenvalues of A and V is a vandermonde matrix given by

$$V = \begin{pmatrix} \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_p^{p-2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1 & \lambda_2 & \cdots & \lambda_p \\ 1 & 1 & \cdots & 1 \end{pmatrix}, \tag{16}$$

where $\{\lambda_i\}_{i=1}^p$ denote the eigenvalues of A (Brand 1964). Let $e_k(\lambda_i)$ denote the elementary symmetric polynomial of order k with variables $\{\lambda_1, \lambda_2, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_p\}$. Also, for all $i \in [p]$, let

$$\alpha_i = \frac{1}{\prod_{j \neq i}(\lambda_i - \lambda_j)}. \tag{17}$$

The inverse of the Vandermonde matrix V can then be explicitly computed (El-Mikkawy 2003) to obtain

$$V^{-1} = \begin{pmatrix} \alpha_1 & -\alpha_1 e_1(\lambda_1) & \cdots & (-1)^{p-1}\alpha_1 e_{p-1}(\lambda_1) \\ \alpha_2 & -\alpha_2 e_1(\lambda_2) & \cdots & (-1)^{p-1}\alpha_2 e_{p-1}(\lambda_2) \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_p & -\alpha_p e_1(\lambda_p) & \cdots & (-1)^{p-1}\alpha_p e_{p-1}(\lambda_p) \end{pmatrix}, \tag{18}$$

Using the diagonalization of A, we can compute its power A^ω as

$$A^\omega = V\Lambda^\omega V^{-1} \tag{19}$$

and the coefficients A^{ω}_{1k} can be computed as

$$(-1)^{k-1}\sum_{i=1}^{p}\alpha_i\lambda_i^{p+\omega-1}e_{k-1}(\lambda_i)$$

Claim. $|A^{\omega}_{1k}|$ **is a Schur polynomial** $S_{\{\omega,1,1,\cdots\, k-1times\ 0,0,\cdots,0\}}$

For any $\{\mu_1,\mu_2,\cdots,\mu_p\}$, consider the generalized Vandermonde matrix

$$V_{\{\mu_1,\mu_2,\cdots,\mu_p\}} = \begin{pmatrix} \lambda_1^{p-1+\mu_1} & \lambda_2^{p-1+\mu_1} & \cdots & \lambda_p^{p-1+\mu_1} \\ \lambda_1^{p-2+\mu_2} & \lambda_2^{p-2+\mu_2} & \cdots & \lambda_p^{p-2+\mu_2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{1+\mu_{p-1}} & \lambda_2^{1+\mu_{p-1}} & \cdots & \lambda_p^{1+\mu_{p-1}} \\ \lambda_1^{\mu_p} & \lambda_2^{\mu_p} & \cdots & \lambda_p^{\mu_p} \end{pmatrix}. \tag{20}$$

The Bilaternant formulation defines Schur polynomial $S_{\{\mu_1,\mu_2,\cdots,\mu_p\}}$ as

$$S_{\{\mu_1,\mu_2,\cdots,\mu_p\}} = \frac{det|V_{\{\mu_1,\mu_2,\cdots,\mu_p\}}|}{det|V|}.$$

It can be shown that the determinant of the vandemonde matrix V can be given as

$$det|V| = \prod_{1\le i<j\le n}(\lambda_i-\lambda_j).$$

Proof of this statement can be found in most standard texts on Matrix analysis, for example, see Horn et al. (2012).

We can then rewrite A^{ω}_{1k} as

$$(-1)^{k-1}\sum_{i=1}^{p}\alpha_i\lambda_i^{p+\omega-1}e_{k-1}(\lambda_i) = det|V|(-1)^{k-1}\sum_{i=1}^{p}\lambda_i^{p+\omega-1}e_{k-1}(\lambda_i)(-1)^{i+1}\prod_{\substack{1\le l<l'\le p\\ l,l'\neq i}}(\lambda_l-\lambda_{l'}).$$

Observe that the term $\prod_{\substack{1\le l<l'\le p\\ l,l'\neq i}}(\lambda_l-\lambda_{l'})$ is the Vandermonde determinent of the principal submatrix of V obtained by deleting ith row and column. Using Theorem 3.2 in De Marchi (2001), which provides a way to compute the determinant of a generalized Vandermonde matrix, we can show that $e_{k-1}(\lambda_i)\prod_{\substack{1\le l<l'\le p\\ l,l'\neq i}}(\lambda_l-\lambda_{l'})$ is the determinant of the corresponding principal submatrix of $V_{\{\mu_1,\mu_2,\cdots,\mu_p\}}$. Using the Laplace expansion for a determinent along with first row of $V_{\{\mu_1,\mu_2,\cdots,\mu_p\}}$, we have

$$\sum_{i=1}^{p}\lambda_i^{p+\omega-1}e_{k-1}(\lambda_i)(-1)^{i+1}\prod_{\substack{1\le l<l'\le p\\ l,l'\neq i}}(\lambda_l-\lambda_{l'}) = det|V_{\{\mu_1,\mu_2,\cdots,\mu_p\}}|$$

and

$$|A^{\omega}_{1k}| = \sum_{i=1}^{p}\alpha_i\lambda_i^{p+\omega-1}e_{k-1}(\lambda_i) = \frac{det|V_{\{\mu_1,\mu_2,\cdots,\mu_p\}}|}{det|V|} = S_{\{\mu_1,\mu_2,\cdots,\mu_p\}}.$$

□

Lemma 4 (**Form of Interventional Autocovariance matrix**). *Consider a vector-valued time series $\{x_t\}_{t\in\mathbb{Z}}\in\mathbb{R}^d$, following a VAR(q) process with autocovariance matrix of size $nd\times nd$ denoted by Σ_n. Consider simultaneous atomic interventions*

on components $\{l_1, l_2, \cdots, l_r\} \subset [d]$ of $x_{t-\omega}$, that is, consider the intervention $do(x_{t-\omega,l_1} = x*_{t-\omega,l_1}, \cdots, x_{t-\omega,l_r} = x*_{t-\omega,l_r})$. Then, the autocovariance matrix of size $nd \times nd$ (Γ'_n) of the corresponding joint interventional distribution, denoted by $\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})$ is given by

$$\Gamma'_n(i,j) = \begin{cases} 0 & \text{if } i \neq j, i = l_m,\ j = l_m\ \forall m \in [r] \\ x*^2_{t-\omega,l_m} & \text{if } i = j = l_m\ \forall m \in [r] \\ \Sigma_n(i,j) & \text{otherwise} \end{cases}. \tag{21}$$

Moreover, let

$$\Gamma_n = \mathbb{E}_{\{x*_{t-\omega,l_m}\}_{m\in[r]} \sim \prod_{m\in[r]} \mathbb{P}(x_{t-\omega,l_m})} \Gamma'_n.$$

Then,

$$\Gamma_n(i,j) = \begin{cases} 0 & \text{if } i \neq j, i = l_m,\ j = l_m\ \forall m \in [r] \\ \Sigma_n(i,j) & \text{otherwise} \end{cases}. \tag{22}$$

The autocovariance matrix of the interventional distribution under simultaneous interventions on consecutive time-steps can be analogously obtained.

Proof of Lemma 4.. Note that due to time ordering and since instantaneous effects are not modelled by a VAR model, there is no directed path from any of the variables $x_{t-\omega,l_1}, x_{t-\omega,l_2}, \cdots, x_{t-\omega,l_r}$ to $\mathbf{x}^n_{t-\omega-1}$ as well as to variables in $\{x_{t-\omega,1}, x_{t-\omega,2}, \cdots, x_{t-\omega,d}\} / x_{t-\omega,l_1}, x_{t-\omega,l_2}, \cdots, x_{t-\omega,l_r}$. Peters et al. (2017, Proposition 6.14) provides graphical criterion for determining the existence of a total causal effect from a variable x to a variable y under interventions on x. Absence of a directed path from x to y implies there is no total causal effect from x to y and from Proposition 6.12 of Peters et al. (2017), we know that $x \perp\!\!\!\perp y$ under the corresponding interventional distribution. As a consequence of these Propositions, we have our desired result.

□

Lemma 5 (**Difference in Causal and Statistical error (VAR(p))**). *Consider a vector-valued time series $\{x_t\}_{t\in\mathbb{Z}} \in \mathbb{R}^d$, following a VAR(q) process with model parameters $\{A_1, A_2, \cdots A_q\}$. Assuming $n > \max\{p, q\}$, for any VAR(p) model f with parameters $\{\widehat{A}_1, \widehat{A}_2, \cdots \widehat{A}_p\}$,*

$$|\mathcal{G}_{do_\omega}(f) - \mathcal{S}(f)| = \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T (\Gamma - \Sigma)(A^\omega_{i:} - \widehat{A}^\omega_{i:}), \tag{23}$$

Proof of Lemma 5. Let A denote the multi-companion matrix corresponding to the true VAR(q) process with model parameters $\{A_1, A_2, \cdots, A_q\}$ of the form described in (14) with the first d rows populated by $\{A'_1, A'_2, \cdots A'_{\max\{p,q\}}\}$, where A'_l is defined as A_l for all $l \leq q$ and as $\mathbf{0}_{d\times d}$ for all $l > q$. Define $\widehat{A}^{(\max\{p,q\})}$ analogously as the multi-companion matrix corresponding to parameters $\left\{\widehat{A}_1, \widehat{A}_2, \cdots, \widehat{A}_p\right\}$ of the estimated VAR(p) model f obtained independently from some statistical estimation procedure $\mathcal{E}$.

Using (12) recursively, we can write

$$y_t^{(\max\{p,q\})} = A^\omega y_{t-\omega}^{(\max\{p,q\})} + A^\omega e_{t-\omega+1}^{(\max\{p,q\})} + A^{\omega-1} e_{t-\omega+2}^{(\max\{p,q\})} + \cdots + A e_{t-1}^{(\max\{p,q\})} + e_t^{(\max\{p,q\})} \tag{24}$$

To reduce cumbersome notation, we let $\zeta_t = A^\omega e_{t-\omega+1} + A^{\omega-1} e_{t-\omega+2} + \cdots + A e_{t-1} + e_t \in \mathbb{R}^{dp}$ and write

$$Y_t = A^\omega y_{t-\omega} + \zeta_t. \tag{25}$$

Let $\hat{x}_t$ denote the prediction of the target variable x_t corresponding to the estimated model f. Then, Statistical error $\mathcal{O}_\omega$ defined with respect to the squared norm can be computed as follows:

$$= \mathbb{E}_{\mathbb{P}(\curvearrowleft^n_{t-\omega}, x_t)}(\|x_t - \hat{x}_t\|^2) \tag{26}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}(\mathbf{x}^n_{t-\omega})\mathbb{P}(x_t|\mathbf{x}^n_{t-\omega})}\left[x_{t,i} - \hat{x}_{t,i}\right]^2 \tag{27}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}(\mathbf{x}^n_{t-\omega})\mathbb{P}(x_t|\mathbf{x}^n_{t-\omega})}\left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 \tag{28}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}(\mathbf{x}^n_{t-\omega})}\left[\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x^2_{t,i}] + (\widehat{A}^{\omega}_{i:})^T y_{t-\omega})^2 - 2\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x_{t,i}](\widehat{A}^{\omega}_{i:})^T y_{t-\omega}\right]^2 \tag{29}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}(\mathbf{x}^n_{t-\omega})}\left[((A^{\omega}_{i:})^T y_{t-\omega} + \zeta_t)^2 + (\widehat{A}^{\omega}_{i:})^T y_{t-\omega})^2 - 2((A^{\omega}_{i:})^T y_{t-\omega})((\widehat{A}^{\omega}_{i:})^T y_{t-\omega})\right]^2 \tag{30}$$

$$= \sum_{i=1}^{d} \Big((A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:})^T \mathbb{E}(y_{t-\omega} y^T_{t-\omega})(A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:}) + \mathbb{E}(\zeta^2_{t,i})\Big) \qquad (\mathbb{E}(x_{t-i}\epsilon^T_t) = 0,\ \forall i \in \mathbb{N}\ ) \tag{31}$$

$$= \sum_{i=1}^{d} (A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:})^T \Sigma_{\max\{p,q\}}(A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:}) + \mathbb{E}(\zeta^2_{t,i}) \tag{32}$$

where (29) holds since the DAG corresponding to the VAR(q) process shows that ($x_{t,i}$ is d-separated from $x_{t-\omega-j,l}$ by $\left\{x_{t-\omega,\cdots,x_{t-\omega-q+1}}\right\}$ for all $j > q, l \in [d]$ and therefore ($x_{t,i} \perp\!\!\!\perp x_{t-\omega-j,l}|x_{t-\omega,\cdots,x_{t-\omega-q+1}} \forall j > q, l \in [d]$).

$$\mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega}, x_t)}(\|x_t - \hat{x}_t\|^2) \tag{33}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^n_{t-\omega})}\left[x_{t,i} - \hat{x}_{t,i}\right]^2 \tag{34}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^n_{t-\omega})}\left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 \tag{35}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^q_{t-\omega})}\left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 \tag{36}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}(X_t|\mathbf{x}^q_{t-\omega})}\left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 \tag{37}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^q_{t-\omega})}\left[\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x^2_{t,i}] + (\widehat{A}^{\omega}_{i:})^T y_{t-\omega})^2 - 2\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x_{t,i}](\widehat{A}^{\omega}_{i:})^T y_{t-\omega}\right]^2 \tag{38}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})}\left[((A^{\omega}_{i:})^T y_{t-\omega} + \zeta_t)^2 + (\widehat{A}^{\omega}_{i:})^T y_{t-\omega})^2 - 2((A^{\omega}_{i:})^T y_{t-\omega})((\widehat{A}^{\omega}_{i:})^T y_{t-\omega})\right]^2 \tag{39}$$

$$= \sum_{i=1}^{d} \Big((A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:})^T \mathbb{E}_{do_\omega}(y_{t-\omega} y^T_{t-\omega})(A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:}) + \mathbb{E}(\zeta^2_{t,i})\Big) \qquad (\mathbb{E}(x_{t-i}\epsilon^T_t) = 0,\ \forall i \in \mathbb{N}\ ) \tag{40}$$

$$= \sum_{i=1}^{d} (A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:})^T \Gamma'_{\max\{p,q\}}(A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:}) + \mathbb{E}(\zeta^2_{t,i}) \tag{41}$$

To see why Equation (37) holds, note that the structural equations that specify the dependence of x_t on $\mathbf{x}^q_{t-\omega}$ remain unchanged under interventions on $x_{t-\omega}$ and therefore the conditional distributions remain unchanged under these interventions.

Therefore,

$$\mathbb{E}_{x*_{t-\omega}\sim\mathbb{P}(x_{t-\omega})}\mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega}, x_t)}(\|x_t - \hat{x}_t\|^2) = \sum_{i=1}^{d} (A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:})^T \Gamma_{\max\{p,q\}}(A^{\omega}_{i:} - \widehat{A}^{\omega}_{i:}) + \mathbb{E}(\zeta^2_{t,i}),$$

where Γ can be obtained using Lemma 4.

□

Corollary 3 (**Difference in Causal and Statistical errors (AR)**). *Let $\{x_t\}$ follow an AR(q) process. Then, for any AR(p) model f with parameters $\{\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p\}$,*

$$|\mathcal{G}_{do_\omega}(f) - \mathcal{S}_\omega(f)| = 2\left|(A^\omega_{1,1} - \widehat{A}^\omega_{1,1}) \sum_{k=2}^{\max\{p,q\}} (A^\omega_{1,k} - \widehat{A}^\omega_{1,k})\gamma_{k-1}\right|, \tag{42}$$

where, for any $k \in \mathbb{N}$, γ_k denotes the autocovariance of $\{x_t\}$ with lag k. A and $\widehat{A}$ are the corresponding companion matrices of the model and estimated parameters as defined in Lemma 5.

Proof of Corollary 3. Corollary 1 directly follows from Lemmas 4 and 5. □

Proposition 2 (**Stability Controls Causal Generalization (VAR)**). *Consider a VAR(q) process. Assuming $n > \max\{p, q\}$, for any VAR(p) model f,*

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq 2\kappa(\Sigma_{\max\{p,q\}})(\mathcal{S}_\omega(f) - \sigma^2_\epsilon), \tag{43}$$

where $\kappa(\Sigma_{\max\{p,q\}})$ denotes the condition number of the autocovariance matrix $\Sigma_{\max\{p,q\}}$.

Proof. From Lemma 5, it remains to prove that

$$|\sum_{j=1}^{d}(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:})| \leq (2\kappa(\Sigma) - 1)(\mathcal{S}_\omega(f) - \sigma^2_\epsilon).$$

First, we show that

$$|(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:})| \leq (2\lambda_{\max}(\Sigma)) \left\|A^\omega_{j:} - \widehat{A}^\omega_{j:}\right\|^2. \tag{44}$$

Case 1. $(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:}) \geq 0.$

$$|(A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:})| = (A^\omega_{j:} - \widehat{A}^\omega_{j:})^T(\Gamma - \Sigma)(A^\omega_{j:} - \widehat{A}^\omega_{j:}), \tag{45}$$

$$\leq (\lambda_{\max}(\Gamma) - \lambda_{min}(\Sigma)) \left\|A^\omega_{j:} - \widehat{A}^\omega_{j:}\right\|^2. \tag{46}$$

where (46) holds by an application of Rayleigh's principle. We still need to show that $\lambda_{\max}(\Gamma) \leq 2\lambda_{\max}(\Sigma)$.

Without loss of generality, assume that $i = 1$, that is the component of $x_{t-\omega}$ that is intervened upon is indexed by 1. Note that, this merely simplifies notation and the following steps also hold simultaneous interventions on multiple components and consecutive time instances without any additional steps.

Representing Σ and Γ in block matrix form, we have

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}. \tag{47}$$

From Lemma 4, we have

$$\Gamma_{11} \in \mathbb{R}^{1\times 1} = \sigma^2 = \mathbb{E}(X_t^2),\ \Gamma_{12}^T = \Gamma_{21} \in \mathbb{R}^{1\times d\max\{p,q\}-1} = 0, \text{ and } \Gamma_{22} = \Sigma_{22}.$$

We can write Γ as follows:

$$\Gamma = \Gamma'_1 + \Gamma'_2, \tag{48}$$

where

$$\Gamma'_1 = \begin{pmatrix} \sigma^2 & \mathbf{0}_{1\times d\max\{p,q\}-1} \\ \mathbf{0}_{d\max\{p,q\}-1\times 1} & \mathbf{0}_{d\max\{p,q\}-1\times d\max\{p,q\}-1} \end{pmatrix}, \tag{49}$$

and

$$\Gamma_2' = \begin{pmatrix} \mathbf{0}_{1\times 1} & \mathbf{0}_{1\times d\max\{p,q\}-1} \\ \mathbf{0}_{d\max\{p,q\}-1\times 1} & \Sigma_{22} \end{pmatrix}. \tag{50}$$

Since Γ_1' and Γ_2' are Hermitian matrices, $\lambda_{\max}(\Gamma) \leq \lambda_{\max}(\Gamma_1') + \lambda_{\max}(\Gamma_2')$.

Observe that Γ_2 is a principal sub-matrix of Γ obtained by deleting the first row and column, by Cauchy's interlacing theorem (Fisk 2005), we have

$$\lambda_{\max}(\Gamma_2') \leq \lambda_{\max}(\Sigma). \tag{51}$$

Note that, when we intervene simultaneously on multiple components and time instances, instead of setting the first row to 0, the covariance matrix of the corresponding interventional distribution Γ can be obtained by deleting the off-diagonal elements of the corresponding rows and columns. It remains to show that $\sigma^2 \leq \lambda_{\max}(\Sigma)$. Note that

$$\lambda_{\max}(\Gamma_2') = \sigma^2 = \Sigma_{11} = e_1^T \Sigma e_1 \leq \lambda_{\max}(\Sigma), \tag{52}$$

where e_i denotes the ith standard basis vector. Combining (51) and (52) we have

$$\lambda_{\max}(\Gamma) \leq 2\lambda_{\max}(\Sigma) \tag{53}$$

and

$$|(A_{j:}^\omega - \widehat{A}_{j:}^\omega)^T(\Gamma - \Sigma)(A_{j:}^\omega - \widehat{A}_{j:}^\omega)| \leq (2\lambda_{\max}(\Sigma) - \lambda_{\min}(\Sigma)) \left\| A_{j:}^\omega - \widehat{A}_{j:}^\omega \right\|^2. \tag{54}$$

Case 2. $(A_{j:}^\omega - \widehat{A}_{j:}^\omega)^T(\Gamma - \Sigma)(A_{j:}^\omega - \widehat{A}_{j:}^\omega) \leq 0$.

$$|(A_{j:}^\omega - \widehat{A}_{j:}^\omega)^T(\Gamma - \Sigma)(A_{j:}^\omega - \widehat{A}_{j:}^\omega)| = (A_{j:}^\omega - \widehat{A}_{j:}^\omega)^T(\Sigma - \Gamma)(A_{j:}^\omega - \widehat{A}_{j:}^\omega), \tag{55}$$

$$\leq (\lambda_{\max}(\Sigma) - \lambda_{\min}(\Gamma)) \left\| A_{j:}^\omega - \widehat{A}_{j:}^\omega \right\|^2. \tag{56}$$

Using the same arguments used in deriving upper bounds for $\lambda_{\max}(\Gamma)$, we can show that $\lambda_{\min}(\Gamma) \geq \lambda_{\min}(\Sigma)$. Therefore, we have

$$|\mathcal{G}_{do_{\omega,i}}(f) - \mathcal{S}_\omega(f)| \leq \sum_{j\in[d]} (2\lambda_{\max}(\Sigma) - \lambda_{\min(\Sigma)}) \left\| A_{j:}^\omega - \widehat{A}_{j:}^\omega \right\|^2 \tag{57}$$

$$\leq (2\lambda_{\max}(\Sigma) - \lambda_{\min(\Sigma)}) \sum_{j\in[d]} \left\| A_{j:}^\omega - \widehat{A}_{j:}^\omega \right\|^2 \tag{58}$$

$$\leq (2\kappa(\Sigma) - 1)(\mathcal{S}_\omega(f) - \sigma_\epsilon^2). \tag{59}$$

To see why (59) holds, observe that

$$\mathcal{S}_\omega(f) - \sigma_\epsilon^2 = \sum_{j=1}^d (A_{j:}^\omega - \widehat{A}_{j:}^\omega)^T \Sigma (A_{j:}^\omega - \widehat{A}_{j:}^\omega) \tag{60}$$

$$\geq \sum_{j=1}^d (A_{j:}^\omega - \widehat{A}_{j:}^\omega)^T \Sigma (A_{j:}^\omega - \widehat{A}_{j:}^\omega) \tag{61}$$

$$\geq \sum_{j=1}^d \lambda_{\min}(\Sigma) \left\| A_{j:}^\omega - \widehat{A}_{j:}^\omega \right\|^2. \tag{62}$$

We now show that we can construct AR(2) processes such that the bound in Proposition 1 is tight upto a small constant factor. Consider an AR(2) process with true model parameters a_1 and a_2. The autocorrelation matrix Σ_2 of this process is given by $\Sigma_p = \begin{pmatrix} 1, \gamma \\ \gamma, 1 \end{pmatrix}$ where $\gamma = \frac{a_1}{1-a_2}$. The eigenvalues of Σ_2 are given by $\lambda_1 = 1 + \gamma$ and $\lambda_2 = 1 - \gamma$ corresponding to eigenvectors u_1 and u_2 respectively. Without loss of generality assume $\gamma > 0$ which yields $\lambda_1 \geq \lambda_2$. Denote vectors

$a = (a_1, a_2)$ and $\hat{a} = (\hat{a}_1, \hat{a}_2)$. Consider an AR(2) process with parameters $\hat{a}_1, \hat{a_2}$ such that $(a - \hat{a}) = u_2$. Then assuming $\omega = 1$, we have that

$$\frac{\mathcal{G}_{do_1} - \mathcal{S}_1}{\mathcal{S}_1 - \sigma_\epsilon^2} = \frac{\|a - \hat{a}\|^2 - (a - \hat{a})^T \Sigma (a - \hat{a})}{(a - \hat{a})^T \Sigma (a - \hat{a})} = \frac{\gamma}{1 - \gamma} = (\kappa(\Sigma) - 1)/2.$$

As a approaches the boundary of the stability domain, the process gets more strongly correlated and $\lambda_{\min}$ approaches 0 and the relative difference in causal and statistical errors diverges. □

Lemma 6 (**Bounds on** a_k)**.** *For any AR(p) model such that the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$,

$$|a_k| \leq \binom{p}{k} \delta^k. \tag{63}$$

Proof of Lemma 6. From Lemma 3, we know that

$$\begin{aligned}
|a_k| &= |S_{\{1,1,\cdots k \; times \; 1,0,\cdots,0\}}(\{\lambda_1, \lambda_2, \cdots, \lambda_p\})| \\
&= |\sum_{\{i_1 < i_2 < \cdots < i_k\} \in [p]} \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_k}| \\
&\leq \sum_{\{i_1 < i_2 < \cdots < i_k\} \in [p]} |\lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_k}| && (|x + y| \leq |x| + |y|) \\
&\leq \sum_{\{i_1 < i_2 < \cdots < i_k\} \in [p]} \delta^k && (|\lambda_i| \leq \delta) \\
&= \binom{p}{k} \delta^k.
\end{aligned}$$

□

Lemma 7 (**Bounds on** γ_k)**.** *For any stochastic process* $\{x_t\}_{t \in \mathbb{Z}}$ *following an AR(p) model the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$

$$|\gamma_k| \leq \frac{C \sigma_\epsilon^2 \delta^k}{1 - \delta^2}$$

Proof of Lemma 7. Using the infinite-moving average representation of X_t (See Brockwell et al. (1991)), we have

$$x_t = \sum_{i=0}^{\infty} A_{11}^i \epsilon_{t-i} \tag{64}$$

$$|\mathbb{E}[x_l, x_r]| = |\mathbb{E}[(\sum_{i_1=0}^{\infty} A_{11}^{i_1} \epsilon_{l-i_1})(\sum_{i_2=0}^{\infty} A_{11}^{i_2} \epsilon_{r-i_2})]| \tag{65}$$

$$= |\sum_{i=0}^{\infty} A_{11}^i A_{11}^{i+|l-r|} \mathbb{E}[\epsilon_t \epsilon_t^T]| \tag{66}$$

$$= |\sigma_\epsilon^2 \sum_{i=0}^{\infty} A_{11}^i A_{11}^{i+|l-r|}| \tag{67}$$

$$\leq K_p \delta^{|l-r|} \sigma_\epsilon^2 \sum_{i=0}^{\infty} \delta^{2i} \tag{68}$$

$$\leq K_p \sigma_\epsilon^2 \frac{\delta^{|l-r|}}{1 - \delta^2} \tag{69}$$

To see why (68) holds observe that, from Lemma 3,

$$A_{11}^i = S_{\{i,0,\cdots,0\}} \leq \sum_{\{i_1 \leq i_2 \leq \cdots \leq i_k\} \in [p]} |\lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_k}| \leq p^p \delta^i$$

□

Lemma 8 (**Lower Bounds on** $\lambda_{\min}(\Sigma)$)**.** *For any stochastic process* $\{x_t\}_{t\in\mathbb{Z}}$ *following an AR(p) model the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$

$$\lambda_{\min}(\Sigma) \geq \frac{\sigma_\epsilon^2}{(1+\delta)^{2p}}$$

Proof. First, note that

$$(1+\sum_{k=1}^{p}|a_k|) \leq \sum_{k=0}^{p}\binom{p}{k}\delta^k = (1+\delta)^p \text{(Binomial Theorem)}.$$

Combining this with the results from Lemma 11 and Proposition 3, we have

$$\lambda_{\min}(\Sigma) \geq 2\pi \inf_{\omega} f(\omega) \geq \frac{\sigma_\epsilon^2}{\nu_{\max}(\mathcal{A})} \geq \frac{\sigma_\epsilon^2}{(1+\sum_{k=1}^{p}|a_k|)^2}$$

$$\lambda_{\min}(\Sigma) \geq \frac{\sigma_\epsilon^2}{(1+\sum_{k=1}^{p}|a_k|)^2} \geq \frac{\sigma_\epsilon^2}{(1+\delta)^{2p}}.$$

□

Lemma 9 (**Upper Bounds on** $\lambda_{\max}(\Sigma)$)**.** *For any stochastic process* $\{x_t\}_{t\in\mathbb{Z}}$ *following an AR(p) model the non-zero eigenvalues of the companion matrix are distinct and satisfy* $|\lambda| \leq \delta < 1$

$$\lambda_{\max}(\Sigma) \leq 2K_p\sigma_\epsilon^2 n\frac{1}{1-\delta^2}$$

Proof. By Gershgorin's theorem (Varga 2010), we can derive an upper bound on the maximum eigenvalue of Σ_n as follows:

$$\lambda_{\max}(\Sigma_n) \leq \max_{i\in[n]}(\Sigma_{ii} + \sum_{j\neq i}|\Sigma_{ij}|).$$

Note that the autocovariance matrix of an AR process which is defined as $\Sigma_{i,j} = \gamma_{|i-j|}$ (the autocovariance of lag $|i-j|$) has a Toeplitz structure. Due to this Toeplitz structure of the autocovariance matrix, we can see that

$$\lambda_{\max}(\Sigma_n) < 2\sum_{i=1}^{n}|\gamma_{i-1}| < 2K_p\sigma_\epsilon^2\sum_{i=1}^{n}\frac{\delta^{i-1}}{1-\delta^2} \leq 2K_p n\sigma_\epsilon^2\frac{1}{1-\delta^2}$$

□

Corollary 4 (**Stability Controls Causal Generalization (AR(p))**)**.** *Consider an AR(q) process, such that eigenvalues of its companion matrix satisfy* $|\lambda| < \delta < 1$*. For any AR(q) model* f*,*

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq K_p\mathcal{S}_\omega(f)\frac{\max\{p,q\}\,(1+\delta)^{2\max\{p,q\}}}{(1-\delta^2)}, \tag{70}$$

where K_p *is some finite constant that depends on the order* p *of the underlying process.*

Proof of Corollary 4. From Proposition 2, we already know that

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq 2\kappa(\Sigma_{\max\{p,q\}})(\mathcal{S}_\omega(f) - \sigma_\epsilon^2), \tag{71}$$

From Lemma 8 and Lemma 9, we have that

$$\lambda_{\min}(\Sigma_{\max\{p,q\}}) \geq \frac{\sigma_\epsilon^2}{(1+\delta)^{2p}}$$

and

$$\lambda_{\max}(\Sigma_{\max\{p,q\}}) \leq 2K_p \max\{p,q\}\, \sigma_\epsilon^2 \frac{1}{1-\delta^2}$$

.

Combining these results, we have the desired result. □

Theorem 2 (**Finite sample bounds for VAR(p) models**). *Let $\mathcal{F}$ denote the family of all VAR models of dimension d and order p. For any $n > \max\{p,q\} \in \mathbb{N}$, let $\mu, m > 0$ be integers such that $2\mu m = n$ and $\delta > 2(\mu-1)\rho^m$ for a fixed constant $0 < \rho < 1$ determined by the underlying process. Let $\{x_1, x_2, \cdots x_n\} \in \mathbb{R}^d$ be a finite sample drawn from a VAR(q) process. Then, simultaneously for every $f \in \mathcal{F}$, under the square loss truncated at M, with probability at least $1-\delta$,*

$$\mathcal{G}_{\omega,i} \leq \zeta \hat{\mathcal{S}}_\omega + \zeta \widehat{\mathfrak{R}}_\mu(\mathcal{F}) + 3\zeta M \sqrt{\frac{\log \frac{4}{\delta'}}{2\mu}} \tag{72}$$

where $\zeta = 1 + \kappa(\Sigma^\nu)$, $\delta' = \delta - 2(\mu-1)\rho^m$, and $\widehat{\mathfrak{R}}_\mu(\mathcal{F})$ denotes the empirical Rademacher complexity of $\mathcal{F}$.

Proof of Theorem 2. From Proposition , we already have that

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq 2\kappa(\Sigma_{\max\{p,q\}})(\mathcal{S}_\omega(f) - \sigma_\epsilon^2). \tag{73}$$

Additionally, processes that follow VAR models are known to be β mixing and in particular, they are geometrically completely regular, that is, there exists some $0 < \rho < 1$ such that $\beta(k) = C\rho^k$ for some constant C, where $\beta(k)$ denotes the β mixing coefficient of the process (Mokkadem 1988). Theorem 2 then follows by applying Rademacher bounds (Mohri et al. 2009, Theorem 1) for generalization in time-series under mixing conditions. □

C RELATIVE INTERVENTIONS

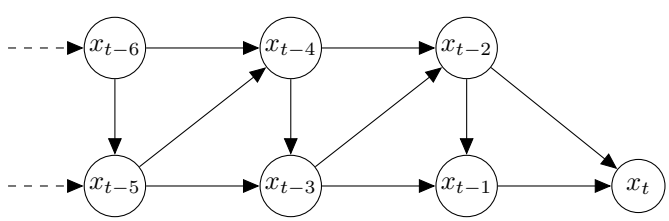

Figure 7: Causal DAG of an AR(2) model

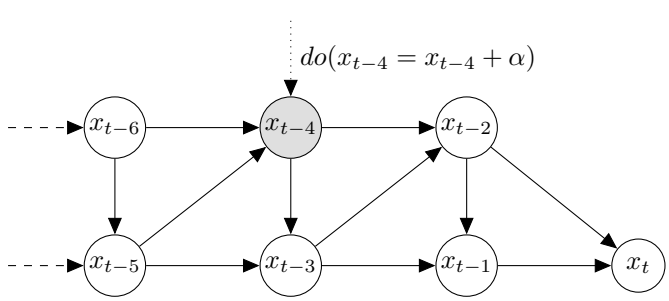

Figure 8: Graphical representation of the effect of an intervention $do(x_{t-4} = x_{t-4} + \alpha)$ on an AR(2) model. Dependencies are retained.

Assume for simplicity $p = q$ and $d = 1$. Let A and $\widehat{A}$ denote the companion matrices corresponding to the true and estimated parameters respectively. Then, rewriting the VAR(p) model as a VAR(1) model, we have

$$x_t = A_{11}^\omega x_{t-\omega} + A_{12}^\omega x_{t-\omega-1} + \cdots + A_{1p}^\omega x_{t-\omega-p+1} + A_{11}^{\omega-1}\epsilon_{t-\omega+1} + \cdots + A_{11}\epsilon_{t-1} + \epsilon_t. \tag{74}$$

Let $\zeta_t = A_{11}^{\omega-1}\epsilon_{t-\omega+1} + \cdots + A_{11}\epsilon_{t-1} + \epsilon_t$. Then, Statistical error S_ω can be computed as

$$\mathbb{E}[x_t - \hat{x}_t]^2 = \mathbb{E}[\sum_{i=1}^{p}(A_{1i}^{\omega} - \widehat{A}_{1i}^{\omega})x_{t-\omega-i+1} + \zeta_t^2] \tag{75}$$

$$= \sum_{ij=1}^{p}(A_{1i}^{\omega} - \widehat{A}_{1i}^{\omega})(A_{1j}^{\omega} - \widehat{A}_{1j}^{\omega})\Sigma_{ij} + \mathbb{E}[\zeta_t^2] \tag{76}$$

The causal error $\mathcal{G}_{do_\omega}$ due to the effect of an intervention $do(x_{t-\omega} = x_{t-\omega} + \alpha)$ can be computed as

$$\mathbb{E}_{do_\omega}[x_t - \hat{x}_t]^2 = \mathbb{E}[\sum_{i=1}^{p}(A_{1i}^{\omega} - \widehat{A}_{1i}^{\omega})x_{t-\omega-i+1} + (A_{11}^{\omega} - \widehat{A}_{11}^{\omega})\alpha + \zeta_t^2] \tag{77}$$

$$= \sum_{ij=1}^{p}(A_{1i}^{\omega} - \widehat{A}_{1i}^{\omega})(A_{1j}^{\omega} - \widehat{A}_{1j}^{\omega})\Sigma_{ij} + (A_{11}^{\omega} - \widehat{A}_{11}^{\omega})^2\alpha^2 + \mathbb{E}[\zeta_t^2] \tag{78}$$

To see why (78) holds, recall that $\mathbb{E}[x_t] = 0, \mathbb{E}[\epsilon_t] = 0, \mathbb{E}[x_{t-i}\epsilon_t] = 0\ \forall i \in \mathbb{N}$.

Lemma 10 (**Difference in Causal and Statistical errors (AR) under Relative Interventions**). *Let $\{X_t\}$ follow an AR(q) process. Then, for any AR(p) model f with parameters $\{\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p\}$,*

$$\mathcal{G}_{do_\omega}(f) - \mathcal{S}_\omega(f) = (A_{1,1}^{\omega} - \widehat{A}_{1,1}^{\omega})^2\alpha^2, \tag{79}$$

where, A and $\widehat{A}$ are the corresponding companion matrices of the model and estimated parameters.

D OTHER RESULTS

Proposition 3. *(Basu et al. 2015, Proposition 2.2) Consider a (matrix-valued) polynomial $\mathcal{A}(z) = I_d - \sum_{k=1}^{p} A_k z^k, x \in \mathbb{C},\ p \in \mathbb{N}$, satisfying $det(\mathcal{A}(z)) \neq 0$ for all $|z| < 1$, $\mu_{\max}(\mathcal{A}) \leq (1 + (\nu_{row} + \nu_{col})/2)^2$, where*

$$\nu_{row} = \sum_{k=1}^{p}\max_{1\leq i\leq d}\sum_{j=1}^{d}|A_k(i,j)|, \quad \nu_{col} = \sum_{k=1}^{p}\max_{1\leq i\leq d}\sum_{i=1}^{d}|A_k(i,j)|.$$

Lemma 11 (**Bounds on spectrum of Σ**). *Let $\{X_t\}$ be a second-order stationary time series with spectral density $f(\omega)$ and let Σ_n denote the autocorrelation matrix of size $n \times n$ given by $\Sigma_n(i,j) = \gamma_{|i-j|} = \mathbb{E}(x_{t+i}, x_{t+j})$ for any $i, j \in \mathbb{Z}$. Then the extremal eigenvalues of Σ are bounded as follows.*

$$\lambda_{\min}(\Sigma_n) \geq 2\pi \inf_{\omega} f(\omega) \quad \textit{and} \quad \lambda_{\max}(\Sigma_n) \leq 2\pi \sup_{\omega} f(\omega)\ \forall n \in \mathbb{N}$$

Furthermore, the bound holds uniformly for all $n \in \mathbb{N}$. See Brockwell et al. (1991, Proposition 4.5.3) for a proof of the Lemma.

E ADDITIONAL EXPERIMENTAL RESULTS

In section 5 we described experiments with simulated autoregressive processes. Here, we provide additional plots from these experiments.

E.1 STATISTICAL AND CAUSAL ERRORS

In the main paper we have seen that even in very simple AR models the causal error of an OLS regression estimator can be several times larger than its statistical error. In Figures 9, 10 and 11 we can see that this is also the case for OLS, Lasso and ElasticNet regression and different process orders. All methods can be seen as the solution to an optimization problem, minimizing the empirical statistical error plus some penalty term $\Omega(\hat{a})$, that is, $\sum_{y_i,\hat{y}_i}(y_i - \hat{y}_i)^2 + \lambda\Omega(\hat{a})$, where $\hat{y}_i$ denotes

the model prediction with estimated parameters $\hat{a}$ and $\lambda > 0$ the strength of the regularization. For OLS, the penalty term is zero. For Ridge and Lasso the penalty is the l_2 and l_1 norm respectively, i.e. $\Omega(\hat{a}) = \|\hat{a}\|_2$ for Ridge and $\Omega(\hat{a}) = \|\hat{a}\|_1$ for Lasso. For ElasticNet we have $\Omega(\hat{a}) = \mu \cdot \|\hat{a}\|_1 + (1-\mu) \cdot \|\hat{a}\|_2$, where μ is a parameter balancing the l_1 and l_2 penalty.

We used standard grid-search and 5-fold cross-validation to find the optimal regularization strength. For ElasticNet, we additional optimized μ with the grid search. Except for Figures 12, we use 100 training and 1000 test samples. For all experiments, we simulate our processes with noise variance $\sigma^2 = 1$.

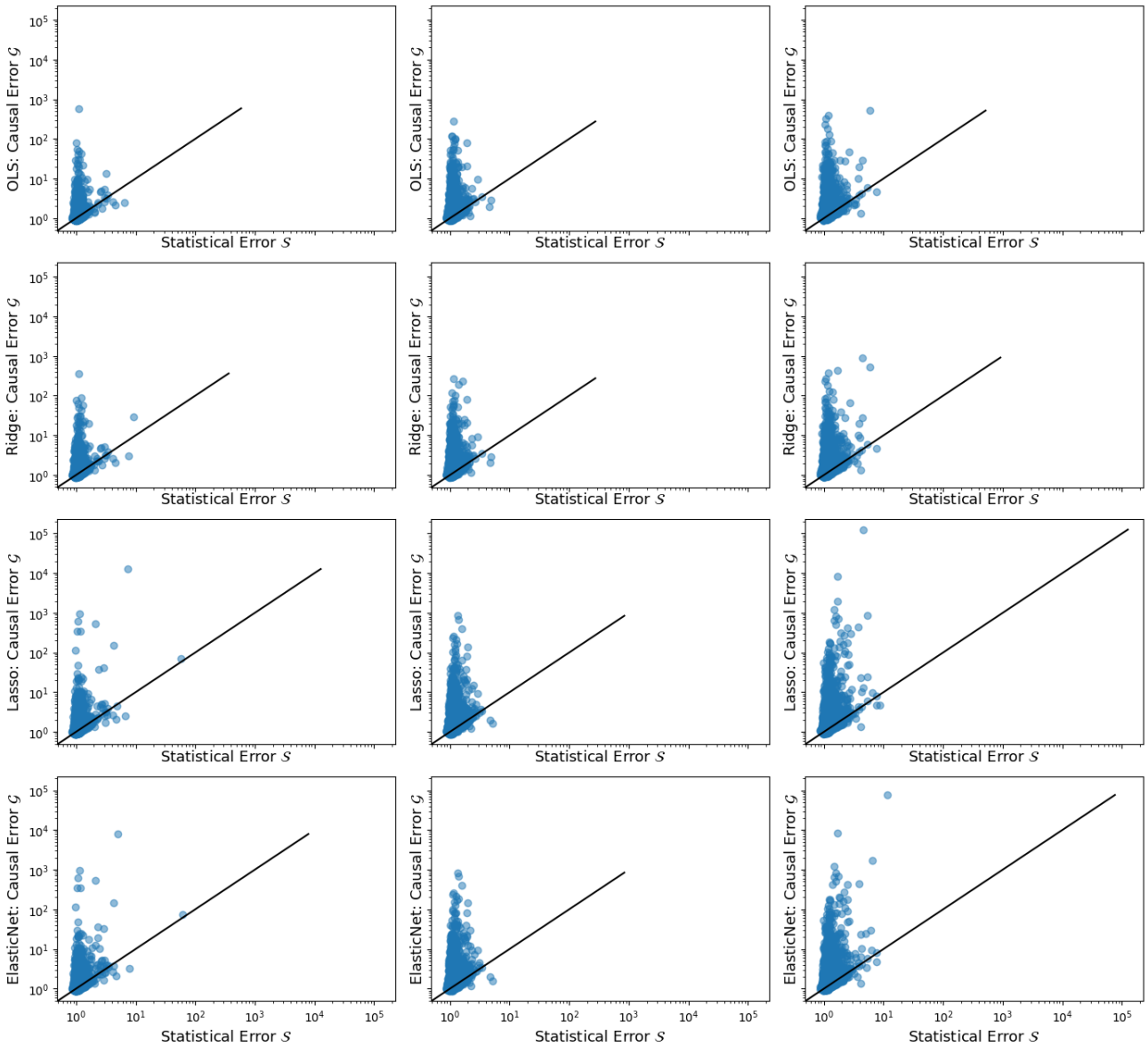

Figure 9: The causal error $\mathcal{G}$ plotted against the statistical error $\mathcal{S}$ for process orders $p = 3, 5, 7$ (from left to right) and estimators OLS, Lasso and ElasticNet (from top to bottom).

Increasing sample size. As one would expect, in Figure 12, we can see that the absolute difference of the errors decreases for larger training samples. We show this result for the ridge regression estimator. Results for other estimators are similar. The respective means are 13.28, 0.48 and 0.18 from left to right and the standard deviations are 264.54, 4.35 and 0.27, which is hard to read from the plot due to the scale of the outliers.

Violations of causal sufficiency. In Figure 14 we violated the causal sufficiency assumption by introducing a hidden confounder. To this end we draw a two-dimensional AR(1) process by drawing each entry of the parameter matrix A

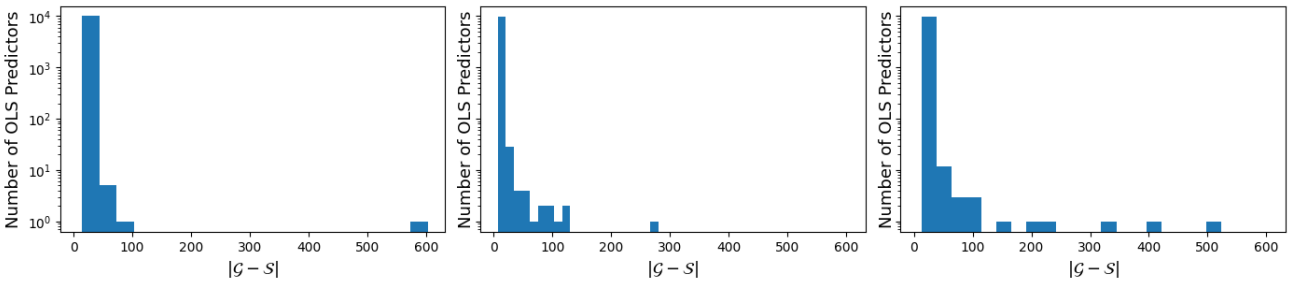

Figure 10: Histogram of the difference $|\mathcal{G} - \mathcal{S}|$ for orders $p = 3, 5, 7$.

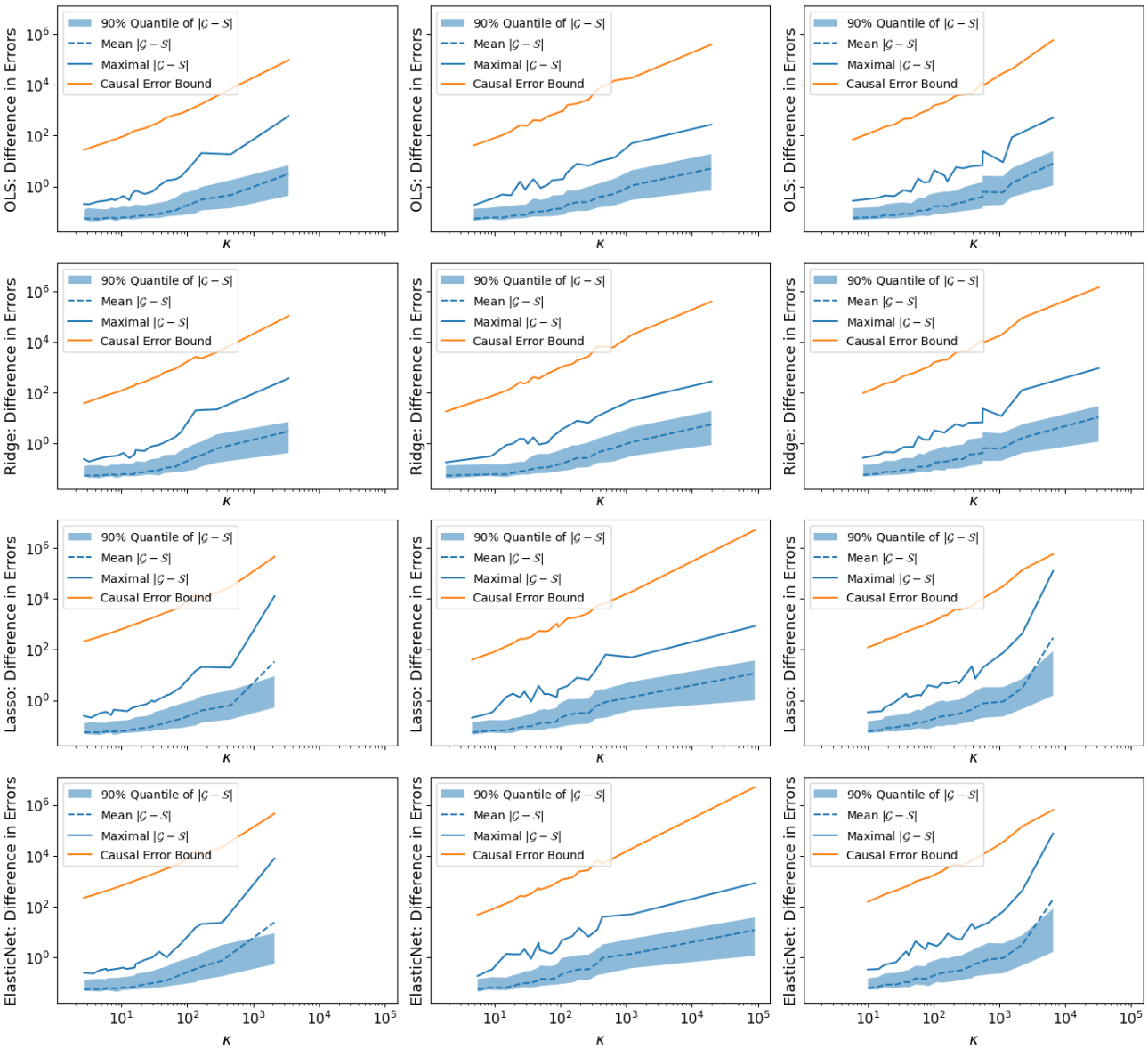

Figure 11: The maximal difference of statistical and causal error $|\mathcal{G} - \mathcal{S}|$ plotted against the condition number of the autocorrelation matrix κ for process orders $p = 3, 5, 7$ (from left to right) and estimators OLS, Lasso and ElasticNet (from top to bottom).

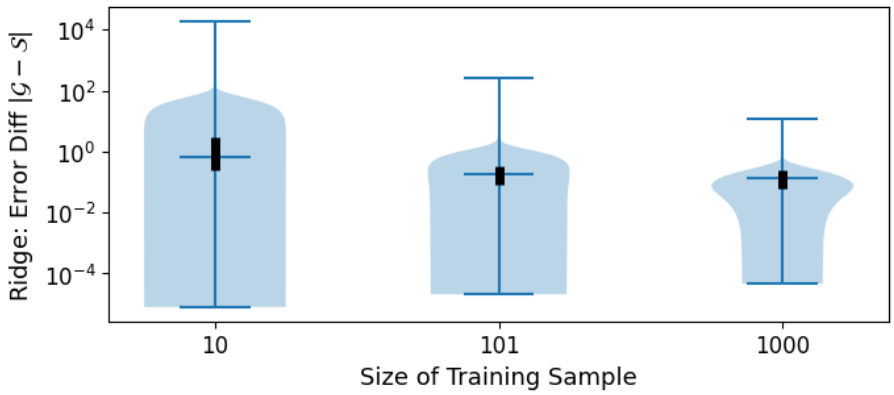

Figure 12: The absolute difference $|\mathcal{G} - \mathcal{S}|$ of causal and statistical error plotted against the sample size for process orders $p = 5$, sample sizes 10, 100, 1000 using Ridge regression. The blue bars mark the 0, 0.5 and 1 quantile and the black block goes from the 0.25 to the 0.75 quantile.

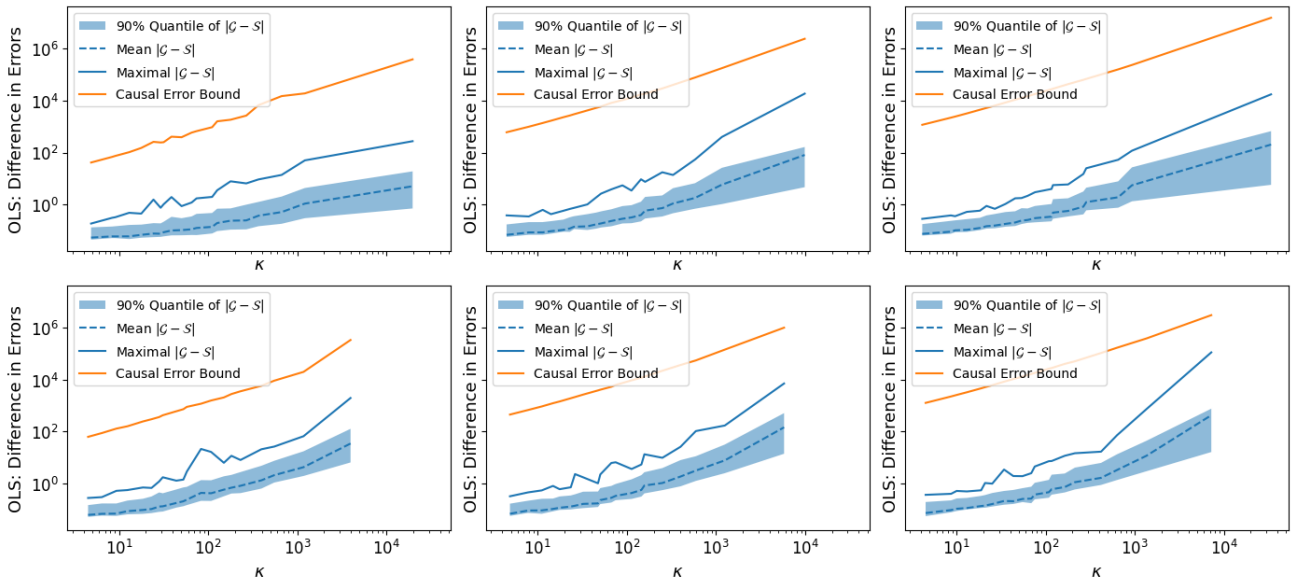

Figure 13: The maximal difference of errors $|\mathcal{G} - \mathcal{S}|$ as well as the generalization bound from Theorem 2 plotted against condition number of the autocorrelation matrix for process order $p = 5$, steps predicted ahead $\omega = 1, 5, 7$ (from left to right). The top row show interventions only on the most recent timestep x_{t-1} where the bottom row shows interventions on all previous timesteps before the prediction.

independently and uniformly from $[-2, 2]$ and reject matrices that yield non-stationary processes. We then only use one of the two dimensions as training and test sample. The other one acts as hidden confounder. We also use only the sample of the observed dimension to estimate the autocorrelation of the process, which is the x-axis of the plots in Figure 14.

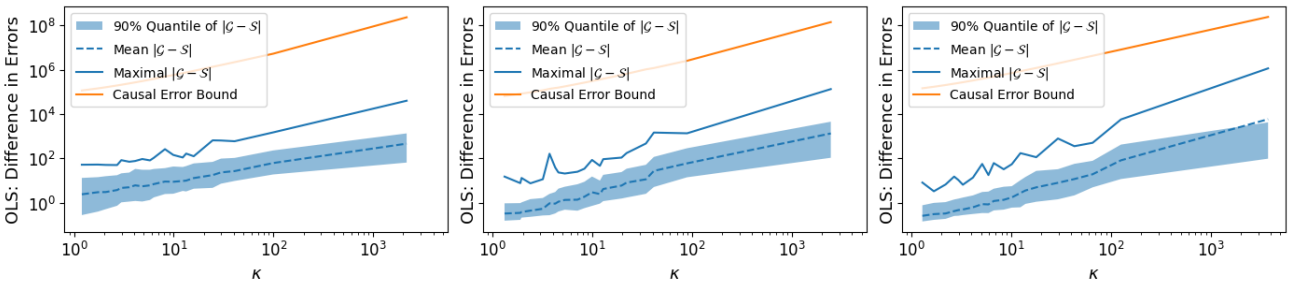

Figure 14: The maximal difference of errors $|\mathcal{G} - \mathcal{S}|$ as well as the generalization bound from Theorem 2 plotted against condition number of the autocorrelation matrix for process order $p = 5$, steps predicted ahead $\omega = 1, 5, 7$ (from left to right).

REFERENCES

Alexandrov, Alexander, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang (2019). *GluonTS: Probabilistic Time Series Models in Python*. arXiv: 1906.05264 [cs.LG].

Basu, Sumanta and George Michailidis (2015). "Regularized estimation in sparse high-dimensional time series models". In: *The Annals of Statistics* 43.4, pp. 1535–1567.

Ben-David, Shai, John Blitzer, Koby Crammer, Fernando Pereira, et al. (2007). "Analysis of representations for domain adaptation". In: *Advances in neural information processing systems* 19, p. 137.

Bennett, Andrew, Nathan Kallus, Lihong Li, and Ali Mousavi (2021). "Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1999–2007.

Bica, Ioana, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar (2020). "Estimating counterfactual treatment outcomes over time through adversarially balanced representations". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=BJg866NFvB.

Brand, Louis (1964). "The companion matrix and its properties". In: *The American Mathematical Monthly* 71.6, pp. 629–634.

Brockwell, Peter J, Richard A Davis, and Stephen E Fienberg (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

Cauchy, Augustin Louis (1815). "M 'e memory on functions which can obtain only two equal values é and of opposite signs as a result of the transpositions op 'e r é es between the variables that they contain". In: *Journal de l'Ecole polytechnique* 10.17, pp. 29–112.

Chaugule, Prasad, Mrinal Kumar, Nutan Limaye, Chandra Kanta Mohapatra, Adrian She, and Srikanth Srinivasan (2019). "Schur Polynomials do not have small formulas if the Determinant doesn't!" In: *arXiv preprint arXiv:1911.12520*.

Davison, Edward (1976). "The robust control of a servomechanism problem for linear time-invariant multivariable systems". In: *IEEE transactions on Automatic Control* 21.1, pp. 25–34.

De Marchi, Stefano (2001). "Polynomials arising in factoring generalized Vandermonde determinants: an algorithm for computing their coefficients". In: *Mathematical and computer modelling* 34.3-4, pp. 271–281.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Faloutsos, Christos, Jan Gasthaus, Tim Januschowski, and Yuyang Wang (2018). "Forecasting Big Time Series: Old and New". In: 11.12. ISSN: 2150-8097. DOI: 10.14778/3229863.3229878. URL: https://doi.org/10.14778/3229863.3229878.

Fisk, Steve (2005). "A very short proof of Cauchy's interlace theorem for eigenvalues of Hermitian matrices". In: *arXiv preprint math/0502408*.

Gasthaus, Jan, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski (Apr. 2019). "Probabilistic Forecasting with Spline Quantile Function RNNs". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1901–1910. URL: https://proceedings.mlr.press/v89/gasthaus19a.html.

Grabowski, Wojciech and Aleksander Welfe (2020). "The Tobit cointegrated vector autoregressive model: An application to the currency market". In: *Economic Modelling* 89, pp. 88–100.

Hatt, Tobias and Stefan Feuerriegel (2021). *Sequential Deconfounding for Causal Inference with Unobserved Confounders*. arXiv: 2104.09323 [stat.ME].

Hill, Jennifer and Jerome P Reiter (2006). "Interval estimation for treatment effects using propensity score matching". In: *Statistics in medicine* 25.13, pp. 2230–2256.

Horn, Roger A and Charles R Johnson (2012). *Matrix analysis*. Cambridge university press.

Jacobi, Carl Gustav Jakob (1841). "De functionibus alternantibus earumque divisione per productum e differentiis elementorum conflatum." In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1841.22, pp. 360–371.

Januschowski, Tim, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot (2020). "Criteria for classifying forecasting methods". In: *International Journal of Forecasting* 36.1. M4 Competition, pp. 167–177. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.05.008. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301529.

Janzing, D., L. Minorics, and P. Bloebaum (Aug. 2020). "Feature relevance quantification in explainable AI: A causal problem". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. Online: PMLR, pp. 2907–2916.

Johansson, Fredrik D, Uri Shalit, Nathan Kallus, and David Sontag (2020). "Generalization bounds and representation learning for estimation of potential outcomes and causal effects". In: *arXiv preprint arXiv:2001.07426*.

Lim, Bryan, Ahmed Alaa, and Mihaela van der Schaar (2018). "Forecasting Treatment Responses over Time Using Recurrent Marginal Structural Networks". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 7494–7504.

Lundberg, S. and S. Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 4765–4774.

Lütkepohl, Helmut (2009). "Econometric analysis with vector autoregressive models". In: *Handbook of computational econometrics*, pp. 281–319.

– (2013). "Vector autoregressive models". In: *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing.

Macdonald, Ian Grant (1998). *Symmetric functions and Hall polynomials*. Oxford university press.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilis Assimakopoulos (June 2018). "The M4 Competition: Results, findings, conclusion and way forward". In: *International Journal of Forecasting* 34. DOI: 10.1016/j.ijforecast.2018.06.001.

Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh (2009). "Domain adaptation: Learning bounds and algorithms". In: *arXiv preprint arXiv:0902.3430*.

McDonald, Daniel J, Cosma Rohilla Shalizi, and Mark Schervish (2017). "Nonparametric risk bounds for time-series forecasting". In: *The Journal of Machine Learning Research* 18.1, pp. 1044–1083.

Meir, Ron (2000). "Nonparametric time series prediction through adaptive model selection". In: *Machine learning* 39.1, pp. 5–34.

Melnyk, Igor and Arindam Banerjee (2016). "Estimating structured vector autoregressive models". In: *International Conference on Machine Learning*. PMLR, pp. 830–839.

El-Mikkawy, Moawwad EA (2003). "Explicit inverse of a generalized Vandermonde matrix". In: *Applied mathematics and computation* 146.2-3, pp. 643–651.

Mohri, Mehryar and Afshin Rostamizadeh (2009). "Rademacher Complexity Bounds for Non-I.I.D. Processes". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc.

Mokkadem, Abdelkader (1988). "Mixing properties of ARMA processes". In: *Stochastic processes and their applications* 29.2, pp. 309–315.

Molnar, C. (2019). *Interpretable Machine Learning*. Molnar, C. URL: https://christophm.github.io/interpretable-ml-book/.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv: 1609.03499 [cs.SD].

Pearl, Judea (2009). *Causality*. Cambridge university press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference*. The MIT Press.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3, pp. 1181–1191. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301888.

Shalit, Uri, Fredrik D Johansson, and David Sontag (2017). "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Shi, Claudia, David M Blei, and Victor Veitch (2019). "Adapting neural networks for the estimation of treatment effects". In: *arXiv preprint arXiv:1906.02120*.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*. New York, NY: Springer-Verlag.

Sugiyama, Masashi and Motoaki Kawanabe (2012). *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press. ISBN: 0262017091.

Valdés-Sosa, Pedro A, Jose M Sánchez-Bornot, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García, and Erick Canales-Rodríguez (2005). "Estimating brain functional connectivity with sparse multi-

variate autoregression". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457, pp. 969–981.

Varga, Richard S (2010). *Geršgorin and his circles*. Vol. 36. Springer Science & Business Media.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, Jiaxuan, Jenna Wiens, and Scott Lundberg (2020). *Shapley Flow: A Graph-based Approach to Interpreting Model Predictions*. arxiv:2010.14592.

Yu, Bin (1994). "Rates of convergence for empirical processes of stationary mixing sequences". In: *The Annals of Probability*, pp. 94–116.

Zivot, Eric and Jiahui Wang (2006). "Vector autoregressive models for multivariate time series". In: *Modeling Financial Time Series with S-Plus®*, pp. 385–429.