Causal Forecasting: Generalization Bounds for Autoregressive Models - Supplementary Material

Leena Chennuru Vankadara*[1] **Philipp Michael Faller**[2] **Michaela Hardt**[2] **Lenon Minorics**[2]
Debarghya Ghoshdastidar[3] **Dominik Janzing**[2]

[1]University of Tübingen
[2]Amazon Research
[3]Technical University of Munich, Munich Data Science Institute.

1 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 1.1 (**Vector Autoregressive Model**). A vector autoregressive model (VAR(p)) of dimension d and order p is defined as

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

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 1.2 (**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 $t, \tau \in \mathbb{Z}$

$$\mathbb{E}[x_t] = \mathbb{E}[x_{t+\tau}], \;\; \mathbb{C}_x(t, t+\tau) = \mathbb{C}_x(0, \tau), \tag{2}$$

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 (1) as a VAR(1) model, $y_t = A y_{t-1} + e_t$, 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)***

*Part of this work was completed while the author was at Amazon Research.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

companion matrix defined as:

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{pmatrix}. \tag{3}$$

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

$$\det|\mathbb{I}_d\lambda^p - A_1\lambda^{p-1} - A_2\lambda^{p-2} - \cdots - A_p| = 0, \tag{4}$$

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).

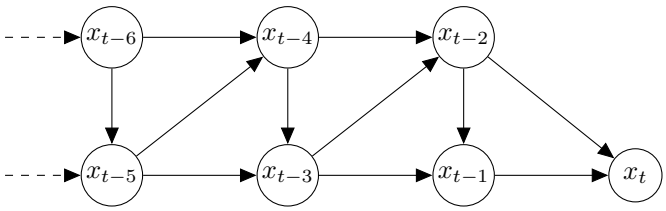

Figure 1: Causal DAG of an AR(2) model

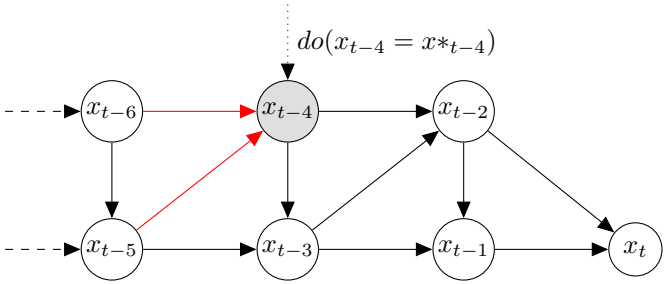

Figure 2: Graphical representation of the effect of an intervention $do(x_{t-4} = x*_{t-4})$ on an AR(2) model. Incoming edges into x_{t-4} are removed in the new DAG which are in red.

Definition 1.3 (**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 $\{-1, +1\}$.

2 PROOFS OF MAIN RESULTS

Lemma 1 (**Expressing powers of a companion matrix using symmetric polynomials**). *For a companion matrix A with distinct eigenvalues and for any $k \in [p]$, the $(1, k)$th element of A^ω, can be expressed as a Schur polynomial of the eigenvalues $\lambda = \{\lambda_1, \lambda_2, \cdots \lambda_p\}$ of A. in particular, $|A^\omega_{1,k}| = S_{\mu_{\omega,k},\lambda}$ where $S_{\mu_{\omega,k},\lambda}$ refers to the Schur polynomial over λ indexed by $\{\omega, 1, \cdots k-1$ times $\cdots, 1, 0, \cdots, 0\}$.*

Proof. For convenience, we use the notation λ and λ/λ_i to denote the sets $\{\lambda_1, \lambda_2, \cdots, \lambda_p\}$ and $\{\lambda_1, \lambda_2, \cdots, \lambda_{i-1}, \lambda_{i+1}, \cdots, \lambda_p\}$ respectively.

Assuming that the eigenvalues $\lambda = \{\lambda_i\}_{i=1}^p$ 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 (Brand 1964) given by

$$V_\lambda = \begin{pmatrix} \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_p^{p-2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1 & \lambda_2 & \cdots & \lambda_p \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{5}$$

For any $i \in [p]$, let $e_k(\lambda/\lambda_i)$ denote the elementary symmetric polynomial of order k with variables in λ/λ_i and let

$$\alpha_i = \frac{1}{\prod_{j \neq i}(\lambda_i - \lambda_j)}. \tag{6}$$

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

$$V^{-1} = \begin{pmatrix} \alpha_1 & -\alpha_1 e_1(\lambda/\lambda_1) & \cdots & (-1)^{p-1}\alpha_1 e_{p-1}(\lambda/\lambda_1) \\ \alpha_2 & -\alpha_2 e_1(\lambda/\lambda_2) & \cdots & (-1)^{p-1}\alpha_2 e_{p-1}(\lambda/\lambda_2) \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_p & -\alpha_p e_1(\lambda/\lambda_p) & \cdots & (-1)^{p-1}\alpha_p e_{p-1}(\lambda/\lambda_p) \end{pmatrix}, \tag{7}$$

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

$$A^\omega = V\Lambda^\omega V^{-1} \tag{8}$$

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

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

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

For any $\mu = \{\mu_1, \mu_2, \cdots, \mu_p\}$ such that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_p$ consider the generalized Vandermonde matrix $V_{\mu,\lambda}$ defined as

$$V_{\mu,\lambda} = \begin{pmatrix} \lambda_1^{p-1+\mu_1} & \lambda_2^{p-1+\mu_1} & \cdots & \lambda_p^{p-1+\mu_1} \\ \lambda_1^{p-2+\mu_2} & \lambda_2^{p-2+\mu_2} & \cdots & \lambda_p^{p-2+\mu_2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{1+\mu_{p-1}} & \lambda_2^{1+\mu_{p-1}} & \cdots & \lambda_p^{1+\mu_{p-1}} \\ \lambda_1^{\mu_p} & \lambda_2^{\mu_p} & \cdots & \lambda_p^{\mu_p} \end{pmatrix}. \tag{9}$$

The Bilaternant formulation defines Schur polynomial $S_{\mu,\lambda}$ as

$$S_{\mu,\lambda} = \frac{\det(V_{\mu,\lambda})}{\det(V_\lambda)}. \tag{10}$$

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

$$\det(V_\lambda) = \prod_{1 \leq i < j \leq n} (\lambda_i - \lambda_j). \tag{11}$$

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

For any $i, k \in [p]$, consider the generalized Vandermonde matrix $V_{\mu_k,\lambda/\lambda_i}$, where $\mu_k = \{1, 1, \cdots k-1\text{times} \cdots 1,\ 0, 0, \cdots, 0\}$. That is,

$$V_{\mu_k,\lambda/\lambda_i} = \begin{pmatrix} \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_{i-1}^{p-1} & \lambda_{i+1}^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_{i-1}^{p-2} & \lambda_{i+1}^{p-2} & \cdots & \lambda_p^{p-2} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \lambda_1^{p-(k-1)} & \lambda_2^{p-(k-1)} & \cdots & \lambda_{i-1}^{p-(k-1)} & \lambda_{i+1}^{p-(k-1)} & \cdots & \lambda_p^{p-(k-1)} \\ \lambda_1^{p-(k+1)} & \lambda_2^{p-(k+1)} & \cdots & \lambda_{i-1}^{p-(k+1)} & \lambda_{i+1}^{p-(k+1)} & \cdots & \lambda_p^{p-(k+1)} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{12}$$

From (10), we know that

$$\det(V_{\mu_k,\lambda/\lambda_i}) = \det(V_{\lambda/\lambda_i}) S_{\mu_k,\lambda/\lambda_i},$$

where $S_{\mu_k,\lambda/\lambda_i}$ is the Schur polynomial of variables λ/λ_i indexed by $\mu_k = \{1, 1, \cdots k-1\text{times} \cdots 1,\ 0, 0, \cdots, 0\}$. Using a combinatorial definition of a Schur polynomial as a summation over semi-standard representations over a Young's Tableaux (see Macdonald (1998) for an exposition), it is easy to verify that

$$S_{\mu_k,\lambda/\lambda_i} = e_{k-1}(\lambda/\lambda_i). \tag{13}$$

Therefore, combining (11) and (13) we can write

$$\det(V_{\mu_k,\lambda/\lambda_i}) = \det(V_{\lambda/\lambda_i}) e_{k-1}(\lambda/\lambda_i) = e_{k-1}(\lambda/\lambda_i) \prod_{\substack{1 \le l < l' \le p \\ l,l' \neq i}} (\lambda_l - \lambda_{l'})$$

Now, observe that we can rewrite A_{1k}^{ω} as

$$\begin{aligned} A_{1k}^{\omega} &= (-1)^{k-1} \sum_{i=1}^{p} \alpha_i \lambda_i^{p+\omega-1} e_{k-1}(\lambda/\lambda_i), \\ &= (-1)^{k-1} \sum_{i=1}^{p} (-1)^{i+1} \lambda_i^{p+\omega-1} e_{k-1}(\lambda/\lambda_i) \prod_{\substack{1 \le l < l' \le p \\ l,l' \neq i}} (\lambda_l - \lambda_{l'}) / \det(V_\lambda), \\ &= (-1)^{k-1} \sum_{i=1}^{p} (-1)^{i+1} \lambda_i^{p+\omega-1} \det(V_{\mu_k,\lambda/\lambda_i}) / \det(V_\lambda). \end{aligned}$$

Finally, letting $\mu_{\omega,k} = \{\omega, 1, 1, \cdots k-1\text{times} \cdots 1,\ 0, 0, \cdots, 0\}$, consider the generalized Vandermonde matrix $V_{\mu_{\omega,k},\lambda}$ given by

$$V_{\mu_{\omega,k},\lambda} = \begin{pmatrix} \lambda_1^{p-1+\omega} & \lambda_2^{p-1+\omega} & \cdots & \lambda_p^{p-1+\omega} \\ \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_p^{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_1^{p-(k-1)} & \lambda_2^{p-(k-1)} & \cdots & \lambda_p^{p-(k-1)} \\ \lambda_1^{p-(k+1)} & \lambda_2^{p-(k+1)} & \cdots & \lambda_p^{p-(k+1)} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{14}$$

Using the Laplace expansion to compute the determinant along the first row of $V_{\mu_{\omega,k},\lambda}$ and observing that for any $i \in [p]$, the minor of $V_{\mu_{\omega,k},\lambda}(1,i)$ is given by $\det(V_{\mu_k,\lambda/\lambda_i})$, we have

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

and once again by invoking the bialternant formulation for Schur polynomials, we have

$$|A_{1k}^{\omega}| = \sum_{i=1}^{p}\alpha_i\lambda_i^{p+\omega-1}e_{k-1}(\lambda_i) = \frac{\det(V_{\mu_{\omega,k},\lambda})}{\det(V_\lambda)} = S_{\mu_{\omega,k},\lambda}.$$

□

Lemma 2 (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 $\mathbb{P}_{do_\omega}(\mathbf{x}_{t-\omega}^n)$ is given by*

$$\Gamma'_n(i,j) = \begin{cases} 0 & \text{if } i \neq j, i = l_m,\ j = l_m\ \forall m \in [r] \\ x*^2_{t-\omega,l_m} & \text{if } i = j = l_m\ \forall m \in [r] \\ \Sigma_n(i,j) & \text{otherwise} \end{cases}. \tag{15}$$

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,

$$\Gamma_n(i,j) = \begin{cases} 0 & \text{if } i \neq j, i = l_m,\ j = l_m\ \forall m \in [r] \\ \Sigma_n(i,j) & \text{otherwise} \end{cases}. \tag{16}$$

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

Proof of Lemma 2.. 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 $\mathbf{x}_{t-\omega-1}^n$ 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 3 (**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* $\{\widehat{A}_1, \widehat{A}_2, \cdots \widehat{A}_p\}$,

$$|\mathcal{G}_{do_\omega}(f) - \mathcal{S}(f)| = \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T (\Gamma - \Sigma)(A^\omega_{i:} - \widehat{A}^\omega_{i:}), \tag{17}$$

Proof of Lemma 3. 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 (3) 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 (1) recursively, we can write

$$y_t^{(\max\{p,q\})} = A^\omega y_{t-\omega}^{(\max\{p,q\})} + A^\omega e_{t-\omega+1}^{(\max\{p,q\})} + A^{\omega-1} e_{t-\omega+2}^{(\max\{p,q\})} + \cdots + A e_{t-1}^{(\max\{p,q\})} + e_t^{(\max\{p,q\})} \tag{18}$$

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

$$Y_t = A^\omega y_{t-\omega} + \zeta_t. \tag{19}$$

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:

$$\begin{aligned}
\mathcal{O}_\omega &= \mathbb{E}_{\mathbb{P}(\mathbf{x}^n_{t-\omega}, x_t)}[\|x_t - \hat{x}_t\|^2] \\
&= \sum_{i=1}^{d} \mathbb{E}[x_{t,i} - \hat{x}_{t,i}]^2 && \text{(Subscript omitted for convenience)} \\
&= \sum_{i=1}^{d} \mathbb{E}[A^\omega_{i,:} y_{t-\omega} + \zeta_{t,\omega,i} - \hat{A}^\omega_{i,:} y_{t-\omega}]^2 \\
&= \sum_{i=1}^{d} (A^\omega_{i:} - \hat{A}^\omega_{i:})^T \mathbb{E}[y_{t-\omega} y^T_{t-\omega}](A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}[\zeta^2_{t,\omega,i}] && (\mathbb{E}[x_{t-i}\epsilon_t^T] = 0,\ \forall i \in \mathbb{N}\ ) \\
&= \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T \Sigma_{\max\{p,q\}} (A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}[\zeta^2_{t,\omega,i}]
\end{aligned}$$

Similarly,

$$\mathcal{G}_{do_\omega} = \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega}, x_t)}(\|x_t - \hat{x}_t\|^2) \tag{20}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^n_{t-\omega})}\left[x_{t,i} - \hat{x}_{t,i}\right]^2 \tag{21}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^n_{t-\omega})}\left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 \tag{22}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}_{do_\omega}(x_t|\mathbf{x}^q_{t-\omega})}\left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 \tag{23}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})\mathbb{P}(X_t|\mathbf{x}^q_{t-\omega})}\left[x^2_{t,i} + \hat{x}^2_{t,i} - 2x_{t,i}\hat{x}_{t,i}\right]^2 \tag{24}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^q_{t-\omega})}\left[\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x^2_{t,i}] + (\widehat{A}^\omega_{i:})^T y_{t-\omega})^2 - 2\mathbb{E}_{\mathbb{P}(x_t|\mathbf{x}^q_{t-\omega})}[x_{t,i}](\widehat{A}^\omega_{i:})^T y_{t-\omega}\right]^2 \tag{25}$$

$$= \sum_{i=1}^{d} \mathbb{E}_{\mathbb{P}_{do_\omega}(\mathbf{x}^n_{t-\omega})}\left[((A^\omega_{i:})^T y_{t-\omega} + \zeta_t)^2 + (\widehat{A}^\omega_{i:})^T y_{t-\omega})^2 - 2((A^\omega_{i:})^T y_{t-\omega})((\widehat{A}^\omega_{i:})^T y_{t-\omega})\right]^2 \tag{26}$$

$$= \sum_{i=1}^{d} \left((A^\omega_{i:} - \widehat{A}^\omega_{i:})^T \mathbb{E}_{do_\omega}(y_{t-\omega} y^T_{t-\omega})(A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}(\zeta^2_{t,i})\right) \qquad (\mathbb{E}(x_{t-i}\epsilon^T_t) = 0,\ \forall i \in \mathbb{N}\ ) \tag{27}$$

$$= \sum_{i=1}^{d} (A^\omega_{i:} - \widehat{A}^\omega_{i:})^T \Gamma'_{\max\{p,q\}} (A^\omega_{i:} - \widehat{A}^\omega_{i:}) + \mathbb{E}(\zeta^2_{t,i}) \tag{28}$$

To see why Equation (24) 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 2.

□

Corollary 1 (Difference in Causal and Statistical errors (AR)). *Let $\{x_t\}$ follow an AR(q) process. Then, for any AR(p) model f with parameters $\{\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p\}$,*

$$|\mathcal{G}_{do_\omega}(f) - \mathcal{S}_\omega(f)| = 2\left|(A^\omega_{1,1} - \widehat{A}^\omega_{1,1}) \sum_{k=2}^{\max\{p,q\}} (A^\omega_{1,k} - \widehat{A}^\omega_{1,k})\gamma_{k-1}\right|, \tag{29}$$

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 3.

Proof of Corollary 1. Corollary 1 directly follows from Lemmas 2 and 3. □

Proposition 1 (Stability Controls Causal Generalization (VAR)). *Consider a VAR(q) process. Assuming $n > \max\{p, q\}$, for any VAR(p) model f,*

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq 2\kappa(\Sigma_{\max\{p,q\}})(\mathcal{S}_\omega(f) - \sigma^2_\epsilon), \tag{30}$$

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

Proof. From Lemma 3, 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

$$|(A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:})^T(\Gamma-\Sigma)(A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:})| \leq (2\lambda_{\max}(\Sigma))\left\|A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:}\right\|^2. \tag{31}$$

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

$$|(A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:})^T(\Gamma-\Sigma)(A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:})| = (A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:})^T(\Gamma-\Sigma)(A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:}), \tag{32}$$

$$\leq (\lambda_{\max}(\Gamma)-\lambda_{min}(\Sigma))\left\|A^{\omega}_{j:}-\widehat{A}^{\omega}_{j:}\right\|^2. \tag{33}$$

where (33) 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

$$\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}, \qquad \Gamma = \begin{pmatrix}\Gamma_{11} & \Gamma_{12}\\ \Gamma_{21} & \Gamma_{22}\end{pmatrix}. \tag{34}$$

From Lemma 2, 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:

$$\Gamma = \Gamma_1' + \Gamma_2', \tag{35}$$

where

$$\Gamma_1' = \begin{pmatrix}\sigma^2 & \mathbf{0}_{1\times d\max\{p,q\}-1}\\ \mathbf{0}_{d\max\{p,q\}-1\times 1} & \mathbf{0}_{d\max\{p,q\}-1\times d\max\{p,q\}-1}\end{pmatrix}, \tag{36}$$

and

$$\Gamma_2' = \begin{pmatrix}\mathbf{0}_{1\times 1} & \mathbf{0}_{1\times d\max\{p,q\}-1}\\ \mathbf{0}_{d\max\{p,q\}-1\times 1} & \Sigma_{22}\end{pmatrix}. \tag{37}$$

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

$$\lambda_{\max}(\Gamma_2') \leq \lambda_{\max}(\Sigma). \tag{38}$$

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

$$\lambda_{\max}(\Gamma_2') = \sigma^2 = \Sigma_{11} = e_1^T\Sigma e_1 \leq \lambda_{\max}(\Sigma), \tag{39}$$

where e_i denotes the ith standard basis vector. Combining (38) and (39) we have

$$\lambda_{\max}(\Gamma) \leq 2\lambda_{\max}(\Sigma) \tag{40}$$

and

$$|(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T(\Gamma - \Sigma)(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})| \leq (2\lambda_{\max}(\Sigma) - \lambda_{\min}(\Sigma)) \left\|A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}\right\|^2. \tag{41}$$

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

$$|(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T(\Gamma - \Sigma)(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})| = (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T(\Sigma - \Gamma)(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}), \tag{42}$$

$$\leq (\lambda_{\max}(\Sigma) - \lambda_{\min}(\Gamma)) \left\|A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}\right\|^2. \tag{43}$$

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

$$|\mathcal{G}_{do_{\omega,i}}(f) - \mathcal{S}_{\omega}(f)| \leq \sum_{j\in[d]} (2\lambda_{\max}(\Sigma) - \lambda_{\min(\Sigma)}) \left\|A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}\right\|^2 \tag{44}$$

$$\leq (2\lambda_{\max}(\Sigma) - \lambda_{\min(\Sigma)}) \sum_{j\in[d]} \left\|A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}\right\|^2 \tag{45}$$

$$\leq (2\kappa(\Sigma) - 1)(\mathcal{S}_{\omega}(f) - \sigma^2_{\epsilon}). \tag{46}$$

To see why (46) holds, observe that

$$\mathcal{S}_{\omega}(f) - \sigma^2_{\epsilon} = \sum_{j=1}^{d} (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T\Sigma(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}) \tag{47}$$

$$\geq \sum_{j=1}^{d} (A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:})^T\Sigma(A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}) \tag{48}$$

$$\geq \sum_{j=1}^{d} \lambda_{\min}(\Sigma) \left\|A^{\omega}_{j:} - \widehat{A}^{\omega}_{j:}\right\|^2. \tag{49}$$

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

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

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

$$|a_k| \leq \binom{p}{k} \delta^k. \tag{50}$$

Proof of Lemma 4. From Lemma 1, we know that

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

□

Lemma 5 (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 5. Using the infinite-moving average representation of X_t (See Brockwell et al. (1991)), we have

$$x_t = \sum_{i=0}^{\infty} A_{11}^i \epsilon_{t-i} \tag{51}$$

$$|\mathbb{E}[x_l, x_r]| = |\mathbb{E}[(\sum_{i_1=0}^{\infty} A_{11}^{i_1}\epsilon_{l-i_1})(\sum_{i_2=0}^{\infty} A_{11}^{i_2}\epsilon_{r-i_2})]| \tag{52}$$

$$= |\sum_{i=0}^{\infty} A_{11}^i A_{11}^{i+|l-r|}\mathbb{E}[\epsilon_t\epsilon_t^T]| \tag{53}$$

$$= |\sigma_\epsilon^2 \sum_{i=0}^{\infty} A_{11}^i A_{11}^{i+|l-r|}| \tag{54}$$

$$\leq K_p\delta^{|l-r|}\sigma_\epsilon^2 \sum_{i=0}^{\infty}\delta^{2i} \tag{55}$$

$$\leq K_p\sigma_\epsilon^2\frac{\delta^{|l-r|}}{1-\delta^2} \tag{56}$$

To see why (55) holds observe that, from Lemma 1,

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

□

Lemma 6 (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 9 and Proposition 2, 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 7 (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 2 (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,*

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq K_p\mathcal{S}_\omega(f)\frac{\max\{p,q\}(1+\delta)^{2\max\{p,q\}}}{(1-\delta^2)}, \tag{57}$$

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

Proof of Corollary 2. From Proposition 1, we already know that

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq 2\kappa(\Sigma_{\max\{p,q\}})(\mathcal{S}_\omega(f) - \sigma_\epsilon^2), \tag{58}$$

From Lemma 6 and Lemma 7, 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 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 $1-\delta$,*

$$\mathcal{G}_{\omega,i} \leq \zeta\hat{\mathcal{S}}_\omega + \zeta\widehat{\mathfrak{R}}_\mu(\mathcal{F}) + 3\zeta M\sqrt{\frac{\log\frac{4}{\delta'}}{2\mu}} \tag{59}$$

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 $\mathcal{F}$.

Proof of Theorem 1. From Proposition , we already have that

$$|\mathcal{G}_{\omega,i}(f) - \mathcal{S}_\omega(f)| \leq (2\kappa(\Sigma_{\max\{p,q\}}) - 1)(\mathcal{S}_\omega(f) - \sigma_\epsilon^2). \tag{60}$$

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 1 then follows by applying Rademacher bounds (Mohri et al. 2009, Theorem 1) for generalization in time-series under mixing conditions. □

3 RELATIVE INTERVENTIONS

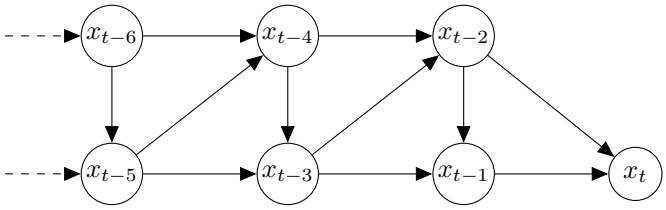

Figure 3: Causal DAG of an AR(2) model

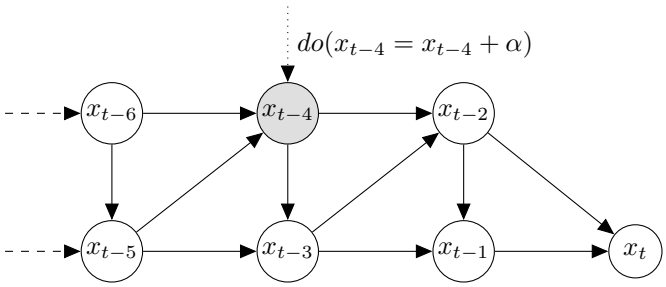

Figure 4: Graphical representation of the effect of an intervention $do(x_{t-4} = x_{t-4} + \alpha)$ on an AR(2) model. Dependencies are retained.

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

$$x_t = A_{11}^\omega x_{t-\omega} + A_{12}^\omega x_{t-\omega-1} + \cdots + A_{1p}^\omega x_{t-\omega-p+1} + A_{11}^{\omega-1}\epsilon_{t-\omega+1} + \cdots + A_{11}\epsilon_{t-1} + \epsilon_t. \tag{61}$$

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

$$\mathbb{E}[x_t - \hat{x}_t]^2 = \mathbb{E}[\sum_{i=1}^{p}(A_{1i}^\omega - \widehat{A}_{1i}^\omega)x_{t-\omega-i+1} + \zeta_t^2] \tag{62}$$

$$= \sum_{ij=1}^{p}(A_{1i}^\omega - \widehat{A}_{1i}^\omega)(A_{1j}^\omega - \widehat{A}_{1j}^\omega)\Sigma_{ij} + \mathbb{E}[\zeta_t^2] \tag{63}$$

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

$$\mathbb{E}_{do_\omega}[x_t - \hat{x}_t]^2 = \mathbb{E}[\sum_{i=1}^{p}(A_{1i}^\omega - \widehat{A}_{1i}^\omega)x_{t-\omega-i+1} + (A_{11}^\omega - \widehat{A}_{11}^\omega)\alpha + \zeta_t^2] \tag{64}$$

$$= \sum_{ij=1}^{p}(A_{1i}^\omega - \widehat{A}_{1i}^\omega)(A_{1j}^\omega - \widehat{A}_{1j}^\omega)\Sigma_{ij} + (A_{11}^\omega - \widehat{A}_{11}^\omega)^2\alpha^2 + \mathbb{E}[\zeta_t^2] \tag{65}$$

To see why (65) 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 8 (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 $\{\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p\}$,*

$$\mathcal{G}_{do_\omega}(f) - \mathcal{S}_\omega(f) = (A^\omega_{1,1} - \widehat{A}^\omega_{1,1})^2 \alpha^2, \tag{66}$$

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

4 OTHER RESULTS

Proposition 2. *(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 9 (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.

5 ADDITIONAL EXPERIMENTAL RESULTS

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

5.1 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 5, 6 and 7 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 8, we use 100 training and 1000 test samples. For all experiments, we simulate our processes with noise variance $\sigma^2 = 1$.

Increasing sample size. As one would expect, in Figure 8, 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 10 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 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 10.

REFERENCES

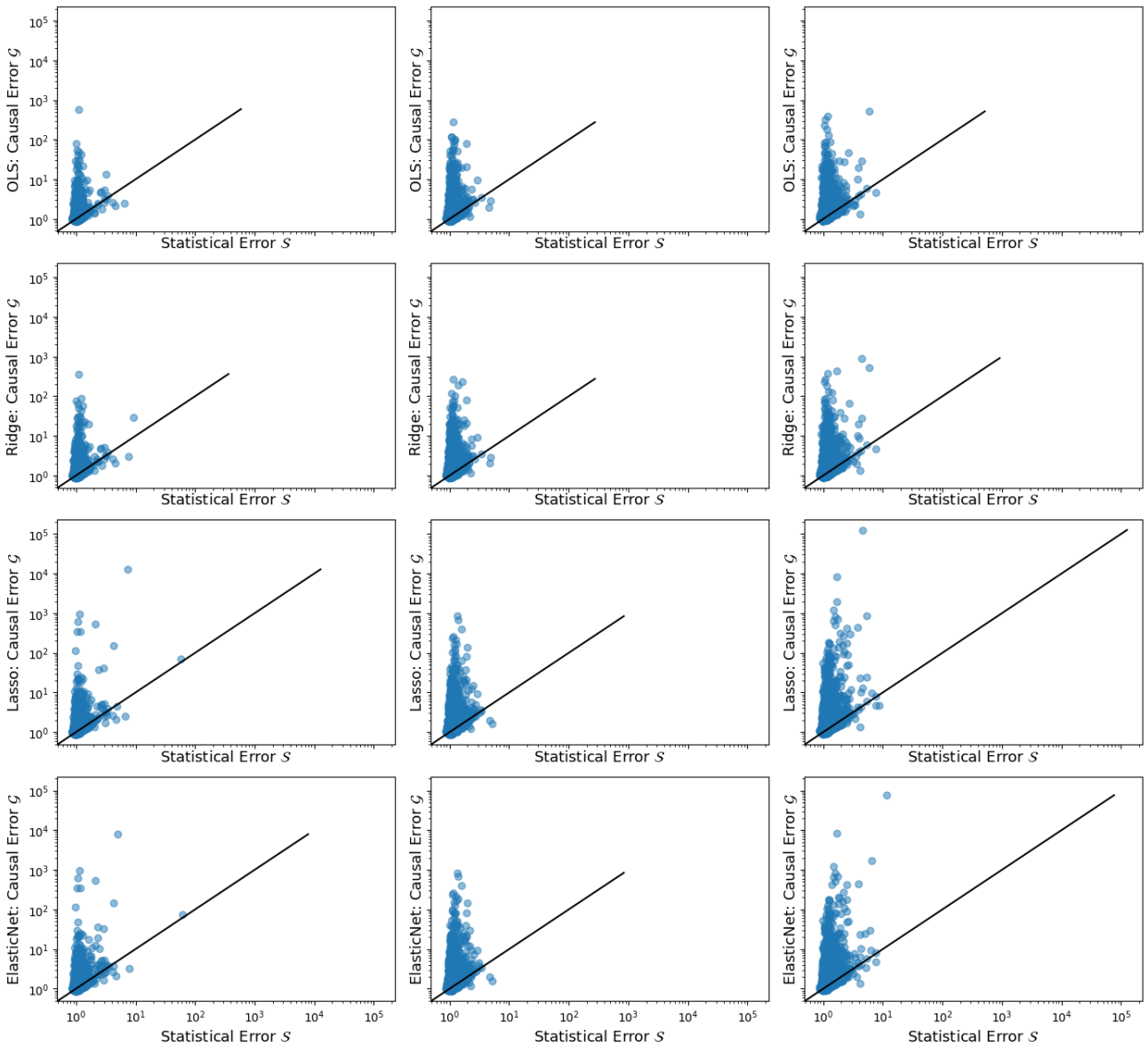

Figure 5: The causal error $\mathcal{G}$ plotted against the statistical error $\mathcal{S}$ for process orders $p = 3, 5, 7$ (from left to right) and estimators OLS, Lasso and ElasticNet (from top to bottom).

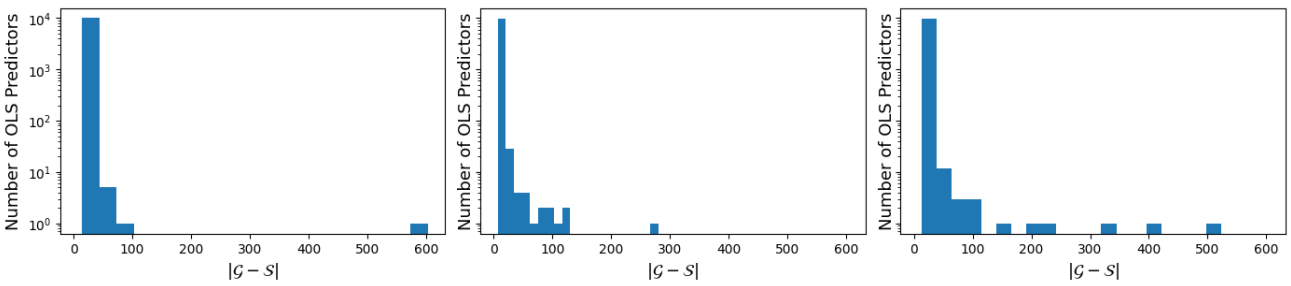

Figure 6: Histogram of the difference $|\mathcal{G} - \mathcal{S}|$ for orders $p = 3, 5, 7$.

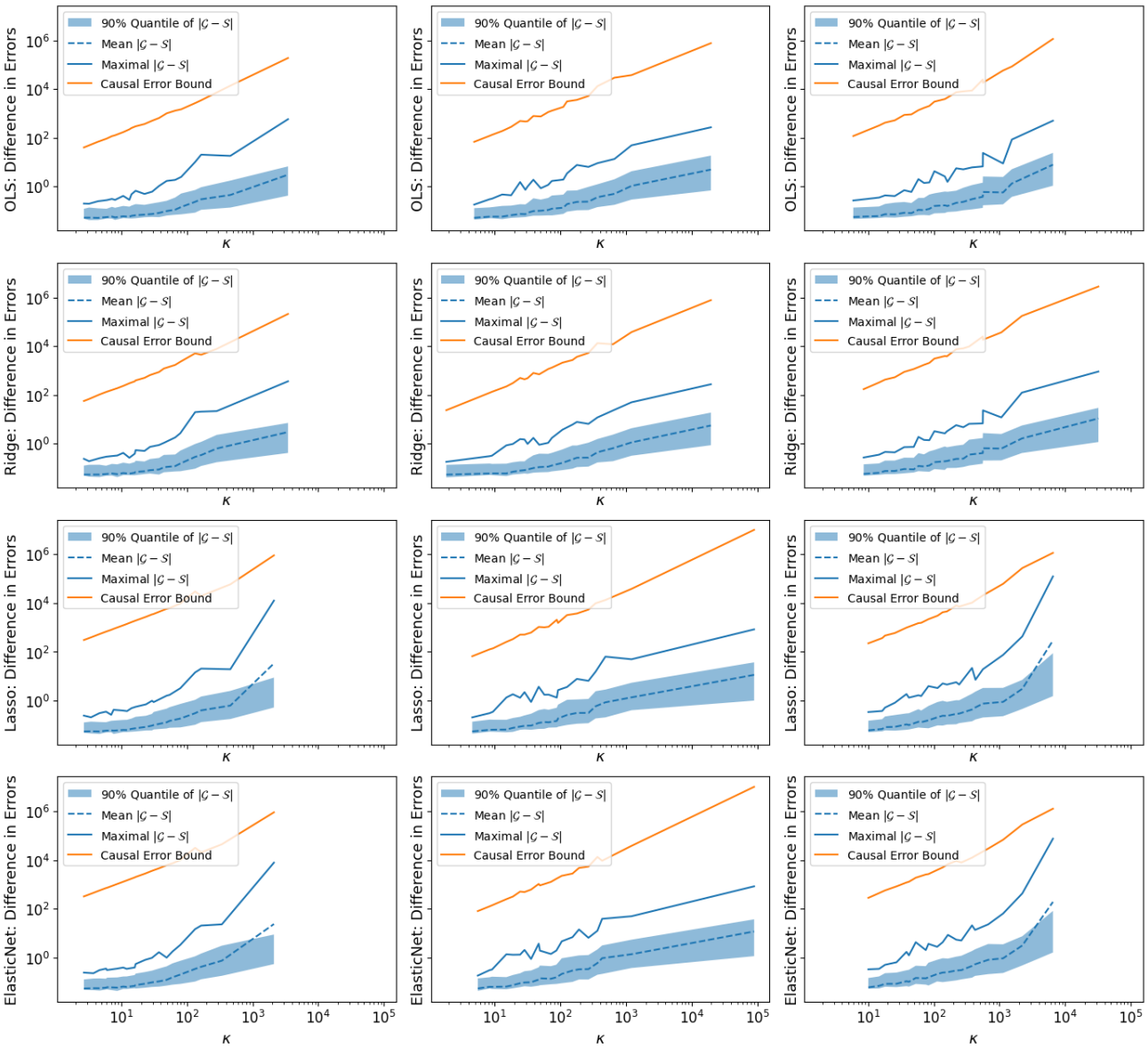

Figure 7: The maximal difference of statistical and causal error $|\mathcal{G} - \mathcal{S}|$ plotted against the condition number of the autocorrelation matrix κ for process orders $p = 3, 5, 7$ (from left to right) and estimators OLS, Lasso and ElasticNet (from top to bottom).

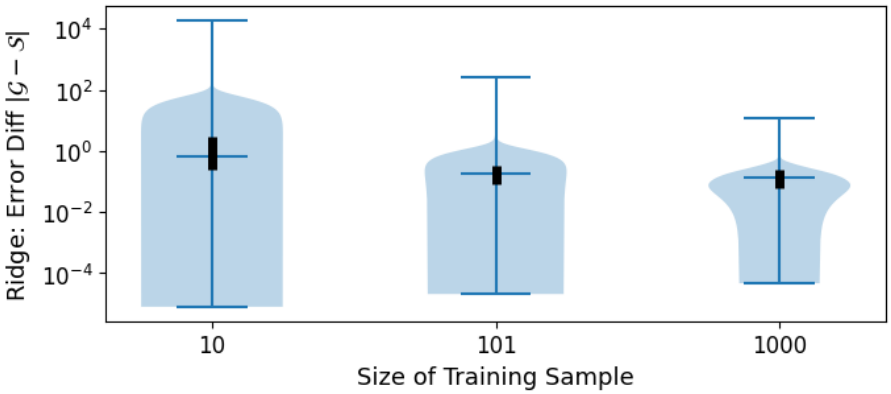

Figure 8: The absolute difference $|\mathcal{G} - \mathcal{S}|$ of causal and statistical error plotted against the sample size for process orders $p = 5$, sample sizes 10, 100, 1000 using Ridge regression. The blue bars mark the 0, 0.5 and 1 quantile and the black block goes from the 0.25 to the 0.75 quantile.

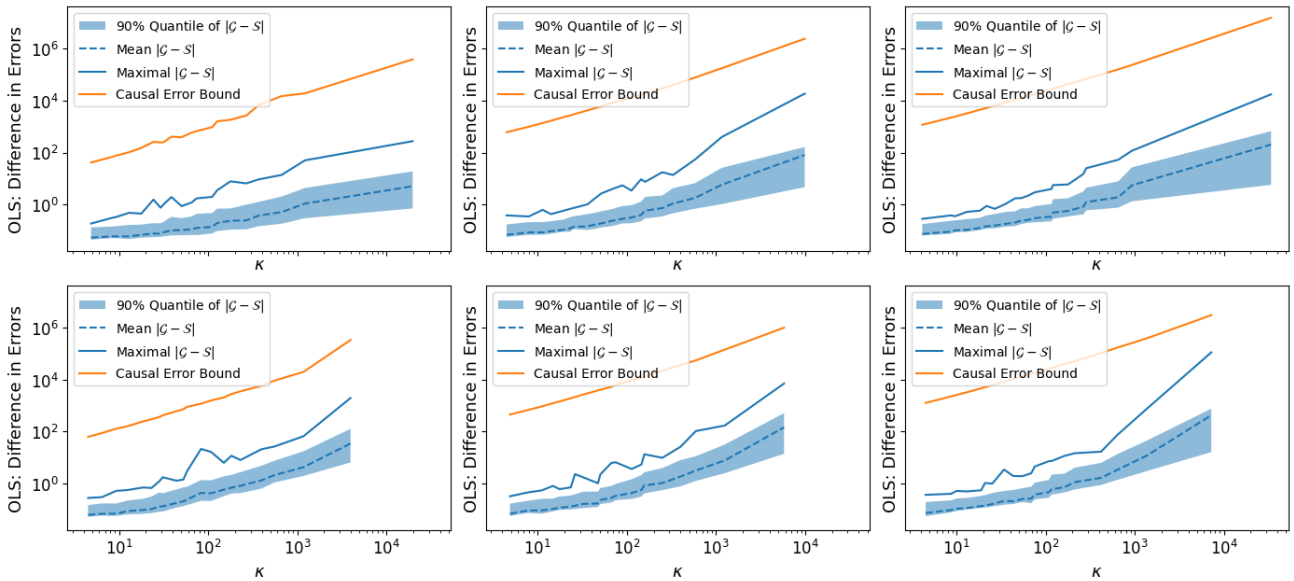

Figure 9: The maximal difference of errors $|\mathcal{G} - \mathcal{S}|$ as well as the generalization bound from Theorem 1 plotted against condition number of the autocorrelation matrix for process order $p = 5$, steps predicted ahead $\omega = 1, 5, 7$ (from left to right). The top row show interventions only on the most recent timestep x_{t-1} where the bottom row shows interventions on all previous timesteps before the prediction.

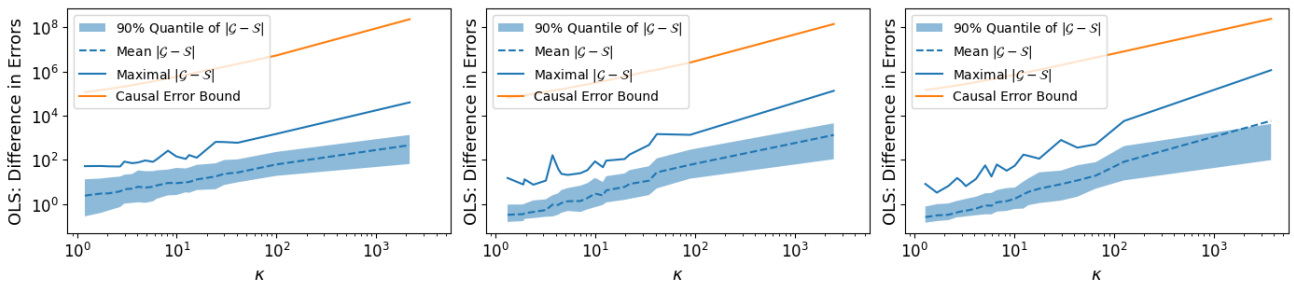

Figure 10: The maximal difference of errors $|\mathcal{G} - \mathcal{S}|$ as well as the generalization bound from Theorem 1 plotted against condition number of the autocorrelation matrix for process order $p = 5$, steps predicted ahead $\omega = 1, 5, 7$ (from left to right).

REFERENCES

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.

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.

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

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

Lütkepohl, Helmut (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.

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.

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

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