Causal Regularization

On the trade-off between in-sample risk and out-of-sample risk guarantees

Lucas Kania*1 and Ernst Wit^{†2}

¹Carnegie Mellon University ²Universitá della Svizzera italiana

Abstract

In recent decades, several data analytic ways of dealing with causality have been introduced, such as propensity score matching, the PC algorithm and invariant causal prediction. Although originally hailed for their interpretational appeal, here we study the identification of causal-like models from in-sample data that provide out-of-sample risk guarantees when predicting a target variable from a set of covariates.

Whereas ordinary least squares provides the best in-sample risk with limited out-of-sample guarantees, causal models have the best out-of-sample guarantees by sacrificing in-sample risk performance. We introduce causal regularization, by defining a trade-off between these properties. As the regularization increases, causal regularization provides estimators whose risk is more stable at the cost of increasing their overall in-sample risk. The increased risk stability is shown to result in out-of-sample risk guarantees. We provide finite sample risk bounds for all models and prove the adequacy of cross-validation for attaining these bounds.

^{*}lucaskania@cmu.edu

[†]ernst.jan.camiel.wit@usi.ch

1 Introduction

In many settings, the goal is to obtain models from in-sample data that are highly predictive for out-of-sample data. Heterogeneous in-sample data is oftentimes detrimental to discovering models that do not overfit [Hernan and Robins, 2010]. However, when the heterogeneity in the sample can be attributed to a covariate shift that does not affect the functional relationship between the target and the covariates, the invariance of that structure implies that the causal model provides predictions that are robust to other future covariate shifts. Hence, the functional causal model is an interesting inferential goal also from a purely predictive point of view.

If the source of the covariate shift is known, instrumental variables [Didelez et al., 2010, Imbens, 2014] can be constructed. Alternatively, if the sample can be split into sub-samples that isolate different instances of the covariate shift, the causal model provides invariant predictions regardless of the sub-sample. Under linearity and non-confounding assumptions, Peters et al. [2016] noted that regressing the target on its direct causes returns a model that is invariant to the covariate shift across the sub-samples. They proposed an algorithm that regresses the target and all possible subsets of covariates in each sub-sample and tests which regression returns the same model regardless of the sub-sample. Under several heterogeneity sources, the authors proved that the algorithm identifies the causal model. Unfortunately, their search algorithm suffers from a combinatorial explosion and does not allow for confounding.

Arjovsky et al. [2019] resolves some of these issues by approximating the search algorithm, which results in more restricted guarantees regarding the identification of the causal model under linearity [Rosenfeld et al., 2020]. Rothenhäusler et al. [2019] avoids the combinatorial search by limiting the source of heterogeneity. They model the covariate shift as a system of structural equations (SEM) being shifted by an instrument. In that scenario, the covariance between the covariates and the residuals under the causal model remains invariant across datasets. Given access to two datasets with different shifts, the difference between these invariant covariances provides a moment condition that uniquely identifies the causal model as long as the covariate shifts affect enough variables in the SEM. Estimating the moment condition gives rise to the causal Dantzig estimator. If the solution is not unique, the authors proposed an L_1 penalty to select a sparse approximation of the causal model. Under some stringent assumptions on the regularization parameter, finite sample bounds can be obtained for the distance between the causal model and the regularized estimator, but no risk guarantees were given for it.

Rojas-Carulla et al. [2018] showed that the prediction invariance of the causal model implies that it minimizes the maximum risk over all possible covariate shifts. Rothenhäusler et al. [2021] generalized that notion under the assumption that the covariate shift originates from an instrumental variable. Their key result is that the out-of-sample risk depends on the correlation between the residuals and the instrument. The more uncorrelated they are,

the stronger the out-of-sample risk guarantees under unseen covariate shifts, whereby the causal model provides the best guarantees. Assuming access to the instrumental variable that generates the covariate shift, the authors propose to build a regularized estimator, called anchor regression, that regulates the amount of correlation between the residuals and the instrumental variable. From that correlation, the out-of-sample guarantees and finite sample bounds of all regularized models follow. Ideally, a regularized model should be chosen based on subject matter knowledge about the maximum future shift of the data. If it is not known, cross-validation is recommended, albeit no proof was provided regarding its adequacy for the proposed loss. Follow-up work extended the method to noisy instrumental variables [Oberst et al., 2021] and discrete and censored outcomes [Kook et al., 2021].

In this work, we consider the multiple datasets setting used in Rothenhäusler et al. [2019], but one in which there is no access to instrumental variables. We propose *causal regularization* for obtaining estimators that progressively decrease the risk bound on shifted out-of-sample datasets, akin to anchor regression. Unlike it, the instrument can be unknown, and the estimators are identifiable even when the unknown instrument does not shift all covariates. We provide finite sample risk bounds for all regularized models and prove the adequacy of cross-validation and data splitting for attaining these bounds.

It is worth remarking that the search for models that provide out-of-sample guarantees as a consequence of in-sample invariance parallels the idea of algorithmic stability [Villa et al., 2013]. The main difference is that the latter studies out-of-sample risk bounds under the assumption that the training and test data come from the same distribution. Algorithmic stability requires model stability [Devroye and Wagner, 1979]. This means that under sample perturbation (usually by leaving one observation out) the algorithm returns models that are similar, in the sense that their predictions do not differ much on average. Kearns and Ron [1999] and Bousquet and Elisseeff [2002] relaxed the stringent requirement of model stability to the requirement of risk stability, i.e., an algorithm returns models that have similar risk given perturbed samples. The work of Peters et al. [2016] follows the model stability approach, where they look for a model that has similar predictions under perturbed sub-samples, while Rothenhäusler et al. [2021], when using a categorical instrument that indicates the sub-samples, looks for a model that provides similar risk across sub-samples. Although the connection with risk stability is not immediate from their results, our work makes the requirement explicit and helps to elucidate the relationship between the two subfields.

2 Structural equation model with shifts

In this paper, we assume that the data generating process is given by a system of linear structural equations (SEM) and that the covariate shift across different datasets is produced by an exogenous random variable.

Definition 1 (Structural equation model). A SEM(B, A) is defined as the stochastic solution (X, Y) of

$$\begin{bmatrix} Y \\ X \end{bmatrix} = \underbrace{\mathbb{B}}_{\substack{constant \\ structure}} \cdot \begin{bmatrix} Y \\ X \end{bmatrix} + \underbrace{\epsilon}_{\substack{noise}} + \underbrace{\begin{bmatrix} 0 \\ \mathbf{A} \end{bmatrix}}_{\substack{covariate \\ shift}} s.t. \quad \mathbf{B} \coloneqq \begin{bmatrix} 0 & \beta_{\mathrm{PA}}^{\phantom{\mathrm{T}}} \\ \beta_{\mathrm{CH}} & \mathbf{B}_{\mathrm{X}} \end{bmatrix} \ and \ \epsilon \coloneqq \begin{bmatrix} \epsilon_{Y} \\ \epsilon_{X} \end{bmatrix}$$

where $X, \epsilon_X, A \in \mathbb{R}^p$ and $Y, \epsilon_Y \in \mathbb{R}$ are random vectors and variables, respectively. The matrix $B_X \in \mathbb{R}^{p \times p}$ consists of the interactions among the covariates X, the vector $\beta_{PA} \in \mathbb{R}^p$ describes the causal effects of X on the target Y, and the vector $\beta_{CH} \in \mathbb{R}^p$ are the downstream effects of Y on X.

The noise is identically distributed irrespective of the shift, i.e., for all A we have $\epsilon \sim \mathcal{L}(0,\Sigma)$, some distribution with zero first moment and a finite second moment. Additionally, the shift A is assumed to be uncorrelated with the noise, i.e., $\mathbb{C}[A,\epsilon] = 0$. Finally, I - B is required to be non-singular in order for SEM(B, A) to be uniquely defined.

Let \mathcal{A} be the set of all distributions that have a finite second moment, then SEM(B) := $\bigcup_{\tilde{A} \in \mathcal{A}} SEM(B, \tilde{A})$ is the set of all the possible realizations of the SEM for a constant structure B with noise distributed as $\mathcal{L}(0, \Sigma)$. As our focus is on providing risk guarantees when predicting Y from X, the SEM definition does not allow for a target shift. This is known as the exclusion restriction in the potential outcomes framework [Rubin, 1974]. This assumption cannot be easily relaxed since a direct target shift affects the identifiability of the SEM. Conversely, all proved results in this work can be generalized to the case where the shift and noise variables are correlated and have a constant covariance over all possible shifts, i.e., $\mathbb{C}[A, \epsilon] = \mathbb{C}[\tilde{A}, \epsilon] \ \forall A, \tilde{A} \in \mathcal{A}$.

Since the target is never directly shifted and the noise is identically distributed regardless of the shifted distribution, the residuals under the causal model β_{PA} are identically distributed for any shift distribution, i.e., $\forall A \in \mathcal{A}$

$$\mathbf{Y}^{\mathbf{A}} - \beta_{\mathbf{P}\mathbf{A}}^{T} \mathbf{X}^{\mathbf{A}} = \epsilon_{Y} \sim \mathcal{L}_{Y}(0, \sigma^{2}) \text{ independent of A}.$$
 (1)

Denote the risk of the using the model β for $(X^A, Y^A) := SEM(B, A)$ as

$$R_{A}(\beta) = \mathbb{E}[(Y^{A} - \beta^{T} X^{A})^{2}]$$

then due to (1) the risk using the causal parameters $R_A(\beta_{PA})$ is invariant over \mathcal{A} . In other words, the risk of the causal model is invariant to all shifts $A \in \mathcal{A}$. Other models do not have this property since for $\beta \neq \beta_{PA}$, there exists always a sequence of shifts such that the risk under β is arbitrarily bad, i.e., $\exists \{A_k\}_{k=1}^{\infty}$ s.t. $\lim_{k\to\infty} R_{A_k}(\beta) = \infty$. Hence, if the causal model were able to be estimated from in-sample data, then under that model the out-of-sample risk would be constant, which would be the best possible guarantee for the out-of-sample risk.

3 Causal regularization

Studies with data from well-designed randomized interventions are able to extract causal parameters [Fisher, 1935]. In this paper, we assume we have much more unstructured data. We consider the setting, in which data from two distributions are available, for example from two separate studies. In particular, we assume that data is available from the observational distribution, i.e. $(X^0, Y^0) = SEM(B, 0)$, and a shifted distribution, i.e. $(X^A, Y^A) = SEM(B, A)$ s.t. $A \not\equiv 0$. There is no direct access to the instrumental variable, i.e., the intervention, that generated the covariate shift A.

3.1 Causal Dantzig

Rothenhäusler et al. [2019] noted that as the noise and shift are uncorrelated, the correlation between the predictors and the residual distribution under the causal model is invariant, i.e., the covariance $\mathbb{C}[X, Y - \beta_{PA}^T X]$ is constant for all $(X, Y) \in SEM(B)$. Consequently, the difference in this quantity for two arbitrary distributions in SEM(B) is zero. This insight characterizes the causal parameter β_{PA} , i.e.,

$$G_{\Delta} \beta_{PA} - Z_{\Delta} = 0 \tag{2}$$

where $G_{\Delta} := \mathbb{E}[X^A X^{A^T}] - \mathbb{E}[X^0 X^{0^t}]$ captures the shift in the covariates, and $Z_{\Delta} := \mathbb{E}[X^A Y^A] - \mathbb{E}[X^0 Y^0]$ captures the shift of the target due to the covariate shift. The causal Dantzig estimator [Rothenhäusler et al., 2019] is defined as the solution of

$$\beta_{\mathrm{CD}} \coloneqq \underset{\beta \in \mathbb{R}^{\mathrm{p}}}{\min} \left\| \mathbf{G}_{\Delta} \, \beta - \mathbf{Z}_{\Delta} \right\|_{2} \tag{3}$$

It is identifiable and unique, if and only if G_{Δ} is full-rank, which happens if and only if $\mathbb{C}[A]$ is full-rank (see proposition 3 in appendix A). In this case $\beta_{\rm CD} = G_{\Delta}^{-1} Z_{\Delta} = \beta_{\rm PA}$. The authors use the infinity norm instead of the L_2 norm but the solutions are equivalent insofar G_{Δ} is not singular.

If random samples $(X^0, Y^0) \in \mathbb{R}^{n_0 \times (p+1)}$ and $(X^A, Y^A) \in \mathbb{R}^{n_A \times (p+1)}$ from the observational and shifted distributions are available, then the plug-in causal Dantzig estimator can be defined as

$$\hat{\beta}_{\mathrm{CD}} \coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\mathrm{p}}} \left\| \mathbb{G}_{\Delta} \, \beta - \mathbb{Z}_{\Delta} \right\|_{2}$$

where $\mathbb{G}_{\Delta} = \mathbb{X}^{A^T} \mathbb{X}^A / n_A - \mathbb{X}^0 \mathbb{X}^0 / n_0$ and $\mathbb{Z}_{\Delta} = \mathbb{X}^{A^T} \mathbb{Y}^A / n_A - \mathbb{X}^{0^T} \mathbb{Y}^0 / n_0$ are the plug-in estimators of G_{Δ} and Z_{Δ} . If \mathbb{G}_{Δ} or G_{Δ} are rank deficient, the authors add an L_1 penalty, but no out-of-sample risk guarantees are given for the regularized estimators.

3.2 Causal regularization with risk guarantees

The aim of this work is to define a regularized version of the causal Dantzig estimator that (1) does not require the existence of a shift variable with a full-rank second moment and (2) has explicit risk guarantees over certain subsets of out-of-sample distributions. In other words, given data from SEM(B, 0) and SEM(B, A), we want explicit guarantees for our performance on SEM(B, \tilde{A}), where \tilde{A} is larger than A, in some precise sense. Hence, we consider the out-of-sample risk over the set C_{γ} of shifts \tilde{A} that are at most γ -times stronger than A, introduced in Rothenhäusler et al. [2021].

Definition 2 (Set of γ -times stronger shifts). For a fixed shift-distribution $A \in \mathcal{A}$, let C_{γ} be the set of shifts that are γ -times stronger than A, i.e.,

$$C_{\gamma} \coloneqq \left\{ \tilde{\mathbf{A}} \in \mathcal{A} \ \middle| \ \begin{cases} \mathbb{E}[\tilde{\mathbf{A}} \tilde{\mathbf{A}}^T] \leq \gamma \, \mathbb{E}[\mathbf{A} \, \mathbf{A}^T] & \text{if } \gamma \in [0, \infty) \\ \text{supp}(\tilde{\mathbf{A}}) = \text{supp}(\mathbf{A}) & \text{if } \gamma = \infty \end{cases} \right\}$$

where $M \leq N$ if M - N is positive semi-definite.

It is possible to decompose the worst out-of-sample risk over shifts in $C_{1+\tau}$ into a weighted sum of the pooled risk and the risk difference without requiring the second moment of the shift variables to be full-rank. This is formalized in the following lemma.

Lemma 1 (Worst risk decomposition).

$$\forall \beta \in \mathbb{R}^{P} \quad \sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta) = \frac{1}{2} R_{+}(\beta) + \frac{1+2\tau}{2} R_{\Delta}(\beta),$$

where
$$R_{+}(\beta) := R_{A}(\beta) + R_{0}(\beta)$$
 is the pooled risk $R_{\Delta}(\beta) := R_{A}(\beta) - R_{0}(\beta)$ is the risk difference

The risk difference directly measures the risk stability of the model across distributions, and it is related to the covariance invariance exploited by the causal Dantzig in the following sense: the smaller the risk difference, the more uncorrelated the residuals are of the model with the covariate shift generated by A. This worst risk decomposition motivates the definition of causal regularization.

Definition 3 (causal regularization). For $\lambda \in [0, \infty)$, the causal regularizer β_{λ} is defined as the minimizer of the worst risk over $C_{(1+\lambda)/2}$,

$$\beta_{\lambda} := \underset{\beta \in \mathbb{R}^{p}}{\operatorname{arg \, min}} \frac{1}{2} \operatorname{R}_{+}(\beta) + \frac{\lambda}{2} \operatorname{R}_{\Delta}(\beta)$$

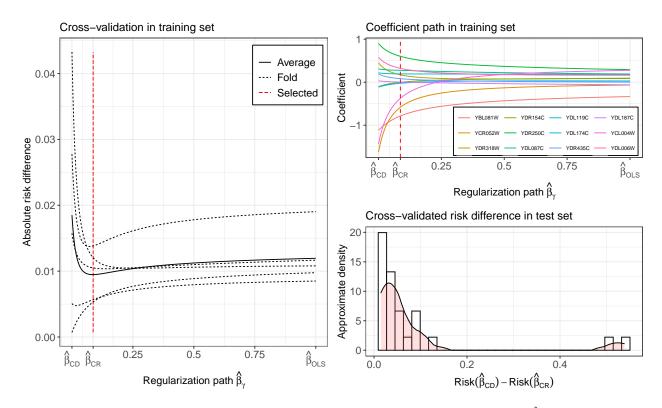


Figure 1: Causal regularization on gene knock-out experiments. (Left) $\hat{\beta}_{CR}$ denotes the model selected by 5-fold cross-validation. (Upper right) Coefficient path generated by causal regularization. (Lower right) Cross-validated risk difference on test set between the causal Dantzig and causal regularization. Observe that in all cases causal regularization achieves a smaller risk as evidence by Risk($\hat{\beta}_{CD}$) – Risk($\hat{\beta}_{CR}$) being always positive.

By increasing λ , causal regularization increases the pooled risk and reduces its risk difference across distributions, thereby increasing its risk stability. It is easy to check with the help of lemma 1 that, by definition, $\beta_{1+2\tau}$ optimizes the worst risk over $C_{1+\tau}$ where $\tau \in [0, \infty)$. Letting the set of shifts $C_{1+\tau}$ increase in magnitude towards the limit set C_{∞} , we recover that the normalized worst risk is equal to the risk difference,

$$\forall \beta \in \mathbb{R}^{P} \quad \lim_{\tau \to \infty} \frac{\sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta)}{\tau} = R_{\Delta}(\beta)$$
 (4)

which means that the pooled risk under β_{λ} increases and the normalized worst out-of-sample risk decreases as λ increases. In other words, the set of shifts, $C_{1+\tau}$, for which β_{λ} provides risk guarantees increases as λ increases.

Equation (4) also asserts that the regularizer $R_{\Delta}(\beta)$ is non-negative for $\beta \in \mathbb{R}^{P}$. This fact

is elucidated by noting that it can be rewritten as a quadratic form centred at β_{PA} , i.e.,

$$R_{\Delta}(\beta) = (\beta - \beta_{PA})^T G_{\Delta}(\beta - \beta_{PA})$$

as shown in proposition 4 in the appendix, which together with the positive semi-definiteness of G_{Δ} imply the non-negativity of $R_{\Delta}(\beta)$.

Based on the quadratic form, we can rewrite the regularizer in a way that clarifies the connection between causal Dantzig and causal regularization. The regularizer corresponds to the causal Dantzig objective multiplied by a pre-conditioner. Hence, by increasing the regularization, causal regularization approaches the causal Dantzig.

Proposition 1 (Convex regularizer).

$$R_{\Delta}(\beta) = R_{||\cdot||}(\beta) \quad s.t. \quad R_{||\cdot||}(\beta) := \left\| (G_{\Delta}^{g/2})^T (G_{\Delta} \beta - Z_{\Delta}) \right\|_2^2$$

where $G_{\Delta}^{g/2}$ is the Moore-Penrose pseudoinverse of the square root of G_{Δ} , i.e., the matrix that satisfies $G_{\Delta} = (G_{\Delta}^{g/2})^T G_{\Delta}^{g/2}$.

It follows that causal regularization interpolates between the causal Dantzig and ordinary least squares applied to both the observational and shifted distributions. The following corollary formalizes this notion.

Corollary 1 (Interpolation by causal regularization). Let $\lambda \in [0, \infty)$ and $G_+ := \mathbb{E}(X^A X^{A^T}) + \mathbb{E}(X^0 X^{O^T})$, if G_+ is non-singular, then

$$\beta_{\lambda} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{p}} \frac{1}{2} \operatorname{R}_{+}(\beta) + \frac{\lambda}{2} \operatorname{R}_{\Delta}(\beta) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{p}} \left\| \operatorname{G}_{\lambda} \beta - \operatorname{Z}_{\lambda} \right\|_{2}^{2} = \operatorname{G}_{\lambda}^{-1} \operatorname{Z}_{\lambda}$$

where $G_{\lambda} := G_{+} + \lambda G_{\Delta}$ and $Z_{\lambda} := Z_{+} + \lambda Z_{\Delta}$. In particular, $\beta_{0} = G_{+}^{-1} Z_{+}$ is the population ordinary least squares. Additionally, if G_{Δ} is non-singular, then $\beta_{\infty} := \lim_{\lambda \to \infty} \beta_{\lambda} = G_{\Delta}^{-1} Z_{\Delta}$ is the causal Dantzig.

3.3 Empirical causal regularization

Consider the plugin estimates $\mathbb{Z}_{+} \coloneqq \mathbb{X}^{A^{T}} \mathbb{Y}^{A} / n_{A} + \mathbb{X}^{0^{T}} \mathbb{Y}^{0} / n_{0}$, $\mathbb{G}_{+} \coloneqq \mathbb{X}^{A^{T}} \mathbb{X}^{A} / n_{A} + \mathbb{X}^{0^{T}} \mathbb{X}^{0} / n_{0}$, $\mathbb{G}_{\lambda} \coloneqq \mathbb{G}_{+} + \lambda \mathbb{G}_{\Delta}$ and $\mathbb{Z}_{\lambda} \coloneqq \mathbb{Z}_{+} + \lambda \mathbb{Z}_{\Delta}$. Furthermore, let the empirical risk of β on $(\mathbb{X}^{A}, \mathbb{Y}^{A})$ be $\hat{R}_{A}(\beta) = \|\mathbb{Y}^{A} - \mathbb{X}^{A}\beta\|_{2}^{2} / n_{A}$, and define the empirical pooled risk and risk difference as

$$\hat{\mathbf{R}}_{+}(\beta) := \hat{\mathbf{R}}_{\mathbf{A}}(\beta) + \hat{\mathbf{R}}_{\mathbf{0}}(\beta)$$
 and $\hat{\mathbf{R}}_{\Delta}(\beta) := \hat{\mathbf{R}}_{\mathbf{A}}(\beta) - \hat{\mathbf{R}}_{\mathbf{0}}(\beta)$

Every minimizer of $\frac{1}{2} \hat{R}_{+}(\beta) + \frac{\lambda}{2} \hat{R}_{\Delta}(\beta)$ satisfies the optimality condition $\mathbb{G}_{\lambda} \beta = \mathbb{Z}_{\lambda}$; hence, we define the causal regularization estimator $\hat{\beta}_{\lambda}$ as the minimum norm least squares solution of the optimality condition, i.e., $\hat{\beta}_{\lambda} = \mathbb{G}_{\lambda}^{g} \mathbb{Z}_{\lambda}$ where \mathbb{G}_{λ}^{g} denote the pseudoinverse of \mathbb{G}_{λ} .

Definition 4 (Empirical causal regularization). Given data from an observational and shifted environment, the empirical causal regularizer for $\lambda \in [0, \infty)$ is defined as

$$\hat{\beta}_{\lambda} := \underset{\beta \in M}{\operatorname{arg \, min}} \|\beta\|_{2} \quad s.t. \quad M := \underset{\beta \in \mathbb{R}^{p}}{\operatorname{arg \, min}} \frac{1}{2} \, \hat{R}_{+}(\beta) + \frac{\lambda}{2} \, \hat{R}_{\Delta}(\beta) \tag{5}$$

or equivalently $\hat{\beta}_{\lambda} = \mathbb{G}_{\lambda}^{g} \mathbb{Z}_{\lambda}$.

Its uniqueness is always guaranteed by the minimum norm condition [Planitz, 1979]. Furthermore, if \mathbb{G}_{Δ} is full rank, the empirical causal regularization interpolates between the OLS estimator on all the data and the causal Dantzig estimator. Its consistency depends only on whether the population matrix $G_{+} = \mathbb{E}(X^{A}X^{AT}) + \mathbb{E}(X^{O}X^{OT})$ is full rank, while the consistency of $\hat{\beta}_{\infty}$, the causal Dantzig estimator, depends on the non-singularity of G_{Δ} .

Proposition 2. If G_+ is positive definite, then $\hat{\beta}_{\lambda} \xrightarrow{P} \beta_{\lambda}$ for $\lambda \in [0, \infty)$

4 Finite sample bound for out-of-sample risk

We derive a finite sample bound for the out-of-sample risk that can be used for any $\beta \in \mathbb{R}^p$, and in particular for $\hat{\beta}_{\lambda}$. The crux of the proof is to decompose the out-of-sample risk using lemma 1, and exploit the concentration of $\hat{R}_A(\beta)$ and $\hat{R}_0(\beta)$ around $R_A(\beta)$ and $R_0(\beta)$, respectively (see proposition 5 in the appendix). The lemma assumes that covariates are normally distributed to avoid boundedness assumptions. The condition can be relaxed to sub-gaussian variables, and a distribution-free bound can be recovered by assuming bounded variables.

Lemma 2 (Finite sample bound for worst population risk). If \mathbb{X}^0 , \mathbb{Y}^0 , \mathbb{X}^A and \mathbb{Y}^A are multivariate centred Gaussian variables

$$\forall \beta \in \mathbb{R}^{P} : \sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta) \le \frac{1}{2} \hat{R}_{+}(\beta) + \frac{1+2\tau}{2} \hat{R}_{\Delta}(\beta) + \eta(\mathbf{n})$$
 (6)

with probability exceeding $1 - 2e^{-q}$ for q > 0, where

$$\eta(\mathbf{n}) := (1+\tau)(\|\beta\|_1^2 + 1) \varphi_+(p, \mathbf{n})
\varphi_+(p, \mathbf{n}) := \varphi(p, \mathbf{n}_{\mathsf{A}}) + \varphi(p, \mathbf{n}_{\mathsf{0}}) \quad and
\varphi(p, \mathbf{n}_{\mathsf{U}}) := \mathbb{V}[\mathbf{Y}^{\mathsf{U}}] \left(\max_{1 \le k \le p} \mathbb{V}[\mathbf{X}_{\mathsf{k}}^{\mathsf{U}}] \right) \left(\sqrt{\frac{4q + 8\log(p)}{n_{\mathsf{U}}}} + \frac{4q + 8\log(p)}{n_{\mathsf{U}}} \right)$$

Hence $\eta(\mathbf{n}) = o_p \left(1 / \sqrt{\min(\mathbf{n}_A, \mathbf{n}_0)} \right)$

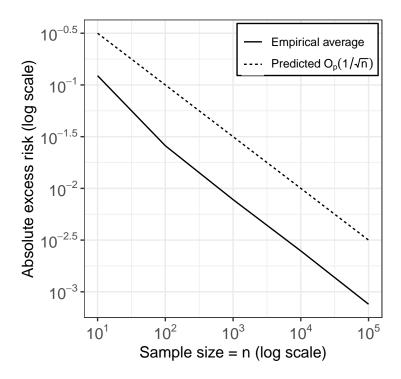


Figure 2: Absolute excess risk of β_{PA} as the sample size is increased. Note that both lines have approximately the same slope, indicating that the convergence rate matches the predicted one.

We remark that due to the decomposition given in lemma 1, the inequality becomes an equality in the limit. Furthermore, $\hat{\beta}_{1+2\tau}$ can be motivated as the model that minimizes the first two terms of the RHS in equation (6) while controlling the norm $\|\hat{\beta}_{1+2\tau}\|_2$ which in turns controls $\eta(n)$.

Generally, the maximum future shift τ for which one wants guarantees is unknown. Thus, lemma 2 is mostly useful for controlling the normalized excess risk, i.e.,

$$\bar{\varphi}(\beta) := \frac{\left(\frac{1}{2} \,\hat{\mathbf{R}}_{+}(\beta) + \frac{1+2\tau}{2} |\,\hat{\mathbf{R}}_{\Delta}(\beta)|\right) - \left(\sup_{\tilde{\mathbf{A}} \in \mathbf{C}_{1+\tau}} \mathbf{R}_{\tilde{\mathbf{A}}}(\beta)\right)}{(1+\tau)(||\beta||_{1}^{2}+1)} \le o_{p}\left(1/\sqrt{\min(\mathbf{n}_{A}, \mathbf{n}_{0})}\right)$$
(7)

which is composed by the normalized excess pooled risk and the normalized excess risk difference

$$\bar{\varphi}(\beta) = \frac{1}{2} \left(\frac{\hat{R}_{+}(\beta) - R_{+}(\beta)}{(1+\tau)(\|\beta\|_{1}^{2} + 1)} \right) + \frac{1+2\tau}{2} \left(\frac{|\hat{R}_{\Delta}(\beta)| - R_{\Delta}(\beta)}{(1+\tau)(\|\beta\|_{1}^{2} + 1)} \right)$$

Figure 2 displays the empirical and predicted normalized excess risk as the sample sizes

grow for the example shown in figure 4. The experiment details are deferred to section 7.1. Finally, the bound can be modified to provide guarantees for out-of-sample datasets.

Corollary 2 (Finite sample bound for worst sample risk). If \mathbb{X}^0 , \mathbb{Y}^0 , \mathbb{X}^{A} , \mathbb{Y}^{A} , $\mathbb{X}^{\tilde{A}}$, and $\mathbb{Y}^{\tilde{A}}$ are multivariate centred Gaussian variables, and $\tilde{A} \in C_{1+\tau}$

$$\forall \beta \in \mathbb{R}^{P}: \quad \hat{R}_{\bar{A}}(\beta) \leq \frac{1}{2} \hat{R}_{+}(\beta) + \frac{1+2\tau}{2} |\hat{R}_{\Delta}(\beta)| + \eta(\mathbf{n})$$

with probability exceeding $1 - 2e^{-q}$ for q > 0, where

$$\eta(\mathbf{n}) := (||\beta||_1^2 + 1)((1 + \tau)\varphi_+(p, n_A, n_o) + \varphi(p, n_{\tilde{A}}))$$

The proof follows from upper-bounding $\hat{R}_{\tilde{A}}$ by $R_{\tilde{A}}$ via a concentration argument. Then, since $R_{\tilde{A}}(\beta) \leq \sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta)$, the corollary follows by applying lemma 2.

5 Model selection

So far, we have only considered out-of-sample risk evaluations for a fixed value of the regularization parameter λ . In this section, we consider selecting the regularization parameter based on resampling. The main idea is to design a model selection procedure that asymptotically chooses the same model as an oracle would with access to the unknown distributions (X^0, Y^0) and (X^A, Y^A) .

Given a set of models $\{\hat{\beta}_{\lambda_1}, \dots, \hat{\beta}_{\lambda_K}\}$ fitted on the in-sample data, a selector λ is a choice of one of the models, i.e., $\lambda \in \Lambda := \{\lambda_1, \dots, \lambda_K\}$. Two selectors $\hat{\lambda}$ and λ are asymptotically equivalent under some loss ℓ , if their difference vanishes in probability asymptotically,

$$\ell(\hat{\lambda}) - \ell(\lambda) \stackrel{\mathrm{p}}{\to} 0.$$

Dudoit and van der Laan [2005] showed the asymptotic equivalence of the sample and population cross-validation selectors (defined for our situation in equations (9) and (8) below) under boundedness assumptions of both the target and covariates. Under further stringent conditions, they prove that the population cross-validation risk of the sample cross-validation selector is equivalent to the risk of the optimal selector, defined in equation (10) below. However, they did not prove that the optimal selector, which has access to models fitted with the full sample rather than a sub-sample, and the sample cross-validation selector are asymptotically equivalent. Van der Vaart et al. [2006] relaxed the assumptions required for the asymptotic equivalence of the sample and population cross-validation selectors, replacing boundedness with pairs of Bernstein numbers. Nevertheless, they also did not prove their asymptotic equivalence with the optimal selector.

In this section, we prove, for a particular loss, that (i) the optimal selector, (ii) the sample cross-validation selector, and (iii) the population cross-validation selector are all

asymptotically equivalent under normal covariates, i.e., without boundedness assumptions. The results follow from the concentration inequality described in lemma 2 (see proposition 5 in the appendix). The result can be generalized to sub-gaussian covariates. As before, distribution-free results can be recovered by assuming bounded covariates.

If the maximum future shift is unknown, one may want to choose a model that will perform well under arbitrarily strong shifts. Corollary 2 states that for arbitrary out-of-sample datasets, $\sup_{\tilde{A}\in C_{1+\tau}} \hat{R}_{\tilde{A}}(\beta)$ concentrates around $\sup_{\tilde{A}\in C_{1+\tau}} R_{\tilde{A}}(\beta)$. Thus, the bound given in lemma 2 can be used to perform model selection. Equation (4) tells us that the normalized worst out-of-sample risk is controlled by the population risk difference corresponding to the in-sample datasets, i.e., $R_{\Delta}(\beta)$. Hence, minimizing $R_{\Delta}(\beta)$ provides out-of-sample guarantees. Given that we only have access to random samples, performing sample-splitting or cross-validation on the absolute in-sample risk difference $|\hat{R}_{\Delta}(\beta)|$ provides a useful surrogate loss. In the following, we show the validity of this idea.

We formally define data resampling in order to later recover the cross-validation and sample-splitting losses. Let $S^{\tilde{\mathbf{A}}} = (S_1^{\tilde{\mathbf{A}}}, \dots, S_{n_{\tilde{\mathbf{A}}}}^{\tilde{\mathbf{A}}}) \in \{0,1\}^{n_{\tilde{\mathbf{A}}}}$ be a random variable that indicates if an observation from the dataset $(\mathbb{X}^{\tilde{\mathbf{A}}}, \mathbb{Y}^{\tilde{\mathbf{A}}})$ is in the train or test set. That is, $S_i^{\tilde{\mathbf{A}}} = 0$ means that $(x_i^{\tilde{\mathbf{A}}}, y_i^{\tilde{\mathbf{A}}})$ is in the training set, while $S_i^{\tilde{\mathbf{A}}} = 1$ indicates that it belongs to the test set. The distribution of $S^{\tilde{\mathbf{A}}}$ determines the methodology. For instance, if its distribution is a point mass at a unique binary string, i.e.,

$$\exists s \in \{0, 1\}^{\tilde{n}}$$
 s.t. $P(S^{\tilde{n}} = s) = 1$

we recover sample-splitting. Alternatively, if there are V binary strings among which the probability mass is homogeneously distributed, i.e.,

$$\exists s^1, \dots, s^V \in \{0, 1\}^{\tilde{n}_{\tilde{A}}} \text{ s.t. } P(S^{\tilde{A}} = s^j) = 1/V \text{ for } j = 1, \dots, V$$

where $\sum_{j=1}^{V} s_i^j = 1$ and $\sum_{i=1}^{n_{\tilde{\Lambda}}} s_i^j = n_{\tilde{\Lambda}}/V$, we recover V-fold cross-validation. Note that $S^{\tilde{\Lambda}}$ need not to be random but it must be stochastically independent from all observations, including those of other datasets. The notion can be formalized using extended conditional independence across the set S of values that $S^{\tilde{\Lambda}}$ can assume [Constantinou and Dawid, 2017], defined as

$$\forall s, \tilde{s} \in \mathcal{S} \quad (\mathbb{X}^{\mathsf{U}}, \mathbb{Y}^{\mathsf{U}}) | S^{\tilde{\mathsf{A}}} = s \quad \sim \quad (\mathbb{X}^{\mathsf{U}}, \mathbb{Y}^{\mathsf{U}}) | S^{\tilde{\mathsf{A}}} = \tilde{s} \quad \text{for } \tilde{\mathsf{A}}, \mathsf{U} \in \{\mathsf{A}, \mathsf{0}\}$$

The sub-empirical distributions $\mathbb{P}_{S,1}^{\tilde{\mathbf{A}}}$ and $\mathbb{P}_{S,0}^{\tilde{\mathbf{A}}}$ are defined by restricting the empirical distribution $\mathbb{P}_{n_{\tilde{\lambda}}}^{\tilde{\mathbf{A}}}$ to the corresponding set,

$$\mathbb{P}_{\mathbf{S}^{\tilde{\mathbf{A}}},\mathbf{j}}^{\tilde{\mathbf{A}}} \coloneqq \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}(\mathbf{S}^{\tilde{\mathbf{A}}},\mathbf{j})} \sum_{i:\mathbf{S}_{i}^{\tilde{\mathbf{A}}}=j}^{\mathbf{n}_{\tilde{\mathbf{A}}}(\mathbf{S}^{\tilde{\mathbf{A}}},\mathbf{j})} \delta_{i}^{\tilde{\mathbf{A}}} \text{ where } \mathbf{n}_{\tilde{\mathbf{A}}}(\mathbf{S}^{\tilde{\mathbf{A}}},\mathbf{j}) \coloneqq \#\{1 \leq i \leq \mathbf{n}_{\tilde{\mathbf{A}}} : \mathbf{S}_{i}^{\tilde{\mathbf{A}}}=j\} \text{ for } j \in \{0,1\} \quad ,$$

Let $S = (S^{\tilde{A}}, S^0)$ be the pair of indicators for each sample, and $(\mathbb{X}_{S^{\tilde{A}},j}^{\tilde{A}}, \mathbb{Y}_{S^{\tilde{A}},j}^{\tilde{A}}) \in \mathbb{R}^{n_{\tilde{A}}(S^{\tilde{A}},j) \times (p+1)}$ for $j \in {0,1}$ denote the design matrix and target vector corresponding to the previous subempirical distributions, then the in-sample risk on the test set is

$$\hat{\mathbf{R}}_{\tilde{\mathbf{A}}}^{S,1}(\beta) = \hat{\mathbf{R}}_{\tilde{\mathbf{A}}}(\beta, \mathbb{P}_{\mathbf{S}^{\tilde{\mathbf{A}}},1}^{\tilde{\mathbf{A}}}) = \left\| \mathbb{Y}_{\mathbf{S}^{\tilde{\mathbf{A}}},1}^{\tilde{\mathbf{A}}} - \mathbb{X}_{\mathbf{S}^{\tilde{\mathbf{A}}},1}^{\tilde{\mathbf{A}}} \beta \right\| / n_{\tilde{\mathbf{A}}}(\mathbf{S}^{\tilde{\mathbf{A}}}, \mathbf{j})$$

and the estimator fitted in the in-sample training set is given by equation (5), where \mathbb{G}_{λ} and \mathbb{Z}_{λ} are computed on the sub-samples $\{(\mathbb{X}_{S^{\tilde{A}},0}^{\tilde{A}},\mathbb{X}_{S^{\tilde{A}},0}^{\tilde{A}})\}_{\tilde{A}\in\{A,0\}}$ rather than the full samples $\{(\mathbb{X}^{\tilde{A}}, \mathbb{Y}^{\tilde{A}})\}_{\tilde{A} \in \{A,0\}}$

$$\hat{\beta}_{\lambda}^{S,0} = \hat{\beta}_{\lambda} (\mathbb{P}_{S^{A},0}^{A}, \mathbb{P}_{S^{0},0}^{0})$$

We define the population and sample selectors as

$$\tilde{\lambda}_{S} := \underset{\lambda \in \Lambda}{\arg \min} \, \tilde{\theta}_{S}(\lambda) \quad \text{s.t.} \quad \tilde{\theta}_{S}(\lambda) := \mathbb{E}_{S} \, \mathcal{R}_{\Delta}(\hat{\beta}_{\lambda}^{S,0}) \quad \text{Population selector}$$
 (8)

$$\tilde{\lambda}_{S} := \underset{\lambda \in \Lambda}{\arg \min} \, \tilde{\theta}_{S}(\lambda) \quad \text{s.t.} \quad \tilde{\theta}_{S}(\lambda) := \mathbb{E}_{S} \, R_{\Delta}(\hat{\beta}_{\lambda}^{S,0}) \quad \text{Population selector}$$

$$\hat{\lambda}_{S} := \underset{\lambda \in \Lambda}{\arg \min} \, \hat{\theta}_{S}(\lambda) \quad \text{s.t.} \quad \hat{\theta}_{S}(\lambda) := \mathbb{E}_{S} \, |\, \hat{R}_{\Delta}^{S,1}(\hat{\beta}_{\lambda}^{S,0})| \quad \text{Sample selector}$$
(9)

where the expectation is with respect to the sub-sampling distribution S and where $\hat{R}^{S,1}_{\Delta}(\beta) := \hat{R}^{S,1}_{\Delta}(\beta) - \hat{R}^{S,1}_{0}(\beta)$ is in-sample risk difference on the test set. It holds that the sample selector is asymptotically equivalent to the population selector

Lemma 3 (Asymptotic equivalence of sample and population selectors). If \mathbb{X}^0 , \mathbb{Y}^0 , \mathbb{X}^A and \mathbb{Y}^A are multivariate centred Gaussian variables

$$|\tilde{\theta}_{S}(\hat{\lambda}_{S}) - \tilde{\theta}_{S}(\tilde{\lambda}_{S})| \stackrel{p}{\rightarrow} 0$$

The result can be interpreted as follows. Define the optimal selector for a particular realization of $\{S^{\tilde{A}}\}_{\tilde{A}\in\{A,0\}}$, called the S-optimal selector, as

$$\tilde{\lambda}_{S,0} = \operatorname*{arg\,min}_{\lambda \in \Lambda} \tilde{\theta}_{S,0}(\lambda) \quad \text{ s.t. } \quad \tilde{\theta}_{S,0}(\lambda) \coloneqq \mathrm{R}_{\Delta}(\hat{\beta}_{\lambda}^{S,0}) \quad S\text{-optimal selector}$$

The sample selector is asymptotically equivalent to the S-optimal selector for the average train-test split, as shown in the following corollary.

Corollary 3 (Asymptotic equivalence of sample and S-optimal selectors). If \mathbb{X}^0 , Y^0 , X^A and Y^A are multivariate centred Gaussian variables

$$|\mathbb{E}_{S}[\tilde{\theta}_{S,0}(\hat{\lambda}_{S}) - \tilde{\theta}_{S,0}(\tilde{\lambda}_{S,0})]| \stackrel{p}{\to} 0$$

Consequently, if the model selection methodology is such that the sub-empirical distributions $\{\mathbb{P}_{S,0}^{\tilde{A}}\}_{\tilde{A}\in\{A,0\}}$ converge in law to the empirical distributions $\{\mathbb{P}_{n_{\tilde{A}}}^{\tilde{A}}\}_{\tilde{A}\in\{A,0\}}$ as n_A and n_0 grow to infinity, then $\hat{\beta}_{\lambda}^{S,0}$ converges to $\hat{\beta}_{\lambda}(\mathbb{P}_{n_A}^A,\mathbb{P}_{n_0}^0)$ due to the continuity of the estimator, and the sample selector is on average asymptotically equivalent to the optimal selector, defined as

$$\tilde{\lambda} := \arg\min_{\lambda \in \Lambda} \tilde{\theta}(\lambda) \quad \text{s.t.} \quad \tilde{\theta}(\lambda) := R_{\Delta}(\beta_{\lambda}(\mathbb{P}_{n_{A}}^{A}, \mathbb{P}_{n_{0}}^{0})) \quad \text{Optimal}$$
(10)

This is the case for sample splitting, leave-one-out cross-validation, and V-fold cross-validation insofar as the number of folds grows towards infinity as the sample sizes increase to infinity.

In other words, under a reasonable model selection strategy the selected estimator from the causal regularization path will be asymptotically equivalent to the optimal selector.

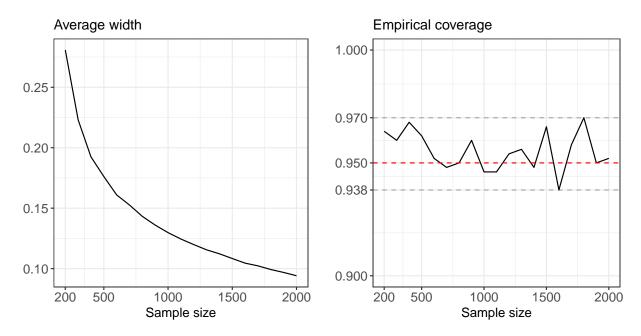


Figure 3: Empirical average width and coverage of the 95% bootstrap confidence interval for the normalized worst risk of $\hat{\beta}_{0.3}$.

6 Normalized worst risk confidence interval via Bootstrap

Equation (4) states that the normalized worst risk of a model $\lim_{\tau\to\infty} \sup_{\tilde{A}\in C_{1+\tau}} R_{\tilde{A}}(\hat{\beta}_{\lambda})/\tau$ is equal to its risk stability $R_{\Delta}(\beta)$. Thus, a confidence interval for the later is a confidence

interval for the former. In this section, we construct a bootstrap pivotal confidence interval for $R_{\Delta}(\beta)$ [Wasserman, 2006].

In order to simplify the notation, let $\hat{\lambda}_*$ denote the selected model and assume that we have a fresh sample that is independent from the one used for model selection. Such condition can always be satisfied via sample splitting. Henceforth, let $(\mathbb{X}^{\tilde{\Lambda}}, \mathbb{Y}^{\tilde{\Lambda}})$ for $\tilde{\Lambda} \in \{A, 0\}$ denote the fresh sample, and $(\mathbb{P}^0, \mathbb{P}^{\Lambda})$ the corresponding empirical distributions.

Given a distribution \mathbb{P} , let \mathbb{P}_b denote a bootstrap resample of \mathbb{P} for $b \in \{1, \dots, B\}$. Then, $\hat{\beta}_{\hat{\lambda}_*}^b := \hat{\beta}_{\hat{\lambda}_*}(\mathbb{P}_b^A, \mathbb{P}_b^0)$ is the selected model refitted in the *b*-bootstrap resample, and $\hat{\beta}_{\hat{\lambda}_*} := \hat{\beta}_{\hat{\lambda}_*}(\mathbb{P}^A, \mathbb{P}^0)$ is the same model refitted in the whole sample. Let $\hat{R}(\beta, \mathbb{P})$ denote the squared prediction error of the model β on the distribution \mathbb{P} , we define the risk difference in the *b*-bootstrap resample and the whole sample as

$$\Delta^b \coloneqq \hat{\mathbf{R}}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_*}^b, \mathbb{P}_b^{\mathbf{A}}) - \hat{\mathbf{R}}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_*}^b, \mathbb{P}_b^{\mathbf{0}}) \quad \Delta_n \coloneqq \hat{\mathbf{R}}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_*}, \mathbb{P}^{\mathbf{A}}) - \hat{\mathbf{R}}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_*}, \mathbb{P}^{\mathbf{0}})$$

An asymptotic $(1 - \alpha)$ confidence interval for the normalized work risk is given by the $1 - \alpha$ bootstrap pivotal confidence interval of Δ restricted to the positive real line, that is

$$[0,\infty) \cap (2\Delta_n - \Delta_{1-\alpha/2}^B, 2\Delta_n - \Delta_{\alpha/2}^B)$$

where Δ_{α}^{B} is the α -quantile of the set $\{\Delta^{b}\}_{b=1}^{B}$. Figure 3 displays the empirical width and coverage of the confidence interval as the sample size is increased. The details of this example are deferred to section 7.2.

7 Simulation studies

In this section, we empirically study causal regularization, and compare it to the causal Dantzig estimator. For all the simulations, the structural matrix B of the structural equation model is displayed in figure 4. In all cases the noise is sampled from a multivariate standard normal. The shift random variable is specified in the following subsections. Moreover, when we mention a sample size n, we mean that both the observational and shifted distribution are sampled n times. The code for reproducing the results presented in section 7 and 8 can be found at github.com/lkania/causal-regularization

7.1 Convergence of out-of-sample risk at $O_p(1/\sqrt{n})$

Figure 2 displays the empirical and predicted normalized excess risk for β_{PA} (see equation (7)) as the sample size is increased. For every sample size, 1000 experiments are realized and their resulting empirical normalized excess risk averaged. The shifted random variable is chosen as $A \sim \mathcal{N}(0, I)$. The computation of the population worst case risk is detailed in appendix B. The fact that the slope of both theoretical and empirical excess risk are the same means that the convergence rate of the empirical risk matches the predicted one.

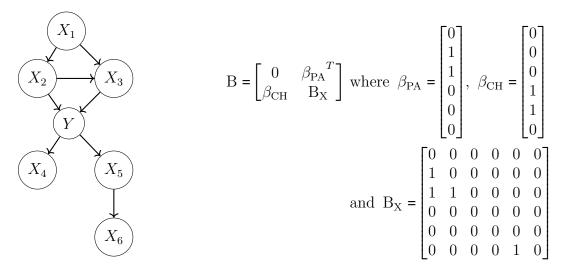


Figure 4: SEM(B, A) used for the simulation study.

7.2 Bootstrap confidence interval

Figure 3 presents the empirical width and coverage of the bootstrap confidence intervals for $\hat{\beta}_{0.3}$, proposed in section 6, as the sample size is increased. The shift is taken to be standard normal, $A \sim \mathcal{N}(0, I)$, and the corresponding population worst risk computation is detailed in section B of the appendix. For every sample size, 500 repetitions are conducted in order to compute the empirical coverage, and for each experiment 1000 bootstrap resamples were used to compute the confidence interval. We observe that even for small sample sizes, the confidence interval possesses, approximately, the desired coverage.

7.3 Causal regularization vs causal Dantzig

Recall that the causal Dantzig is an extreme point in the regularized path generated by causal regularization. In section 5, we saw that causal regularization coupled with cross-validation will be asymptotically equivalent to the optimal selector. Since by definition, the causal Dantzig minimizes the population worst out-of-sample risk, cross-validation will converge to the causal Dantzig for large sample sizes. Nevertheless, for smaller samples the variance of the causal Dantzig might be so high that its actual out-of-sample risk is less than optimal, whereas other estimators in the causal regularization path perform much better. Figure 5 exemplifies such behaviour. It displays the median normalized out-of-sample risk for out-of-sample datasets that are shifted 10, 100 and 1000 times more than the in-sample datasets. The data is generated from the SEM structure shown in figure 4, and the in-sample shift is given by $A \sim \mathcal{N}(0, I)$, while the out-of-sample shifts are chosen as $A \sim \mathcal{N}(0, \lambda I)$ such that $\lambda \in \{10, 100, 1000\}$. The size of both in-sample datasets n are 250 and 1000 while the

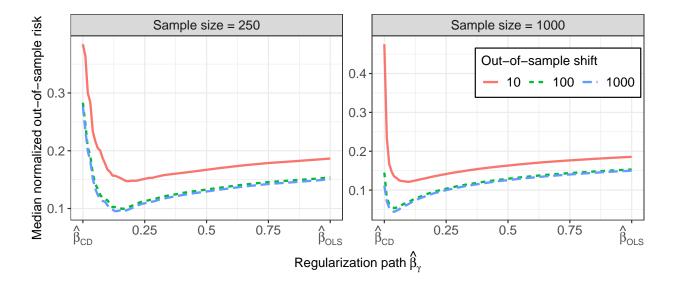


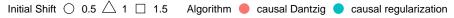
Figure 5: Median normalized risk of the regularization path produced by causal regularization as the out-of-sample datasets are 10, 100 and 1000 times more than the in-sample datasets. Models on the left panel were trained on in-sample datasets consisting of 250 observations while on the right 1000 observation were used. The median was chosen rather than the mean due to the high variance of the causal Dantzig estimator.

out-of-sample datasets always have sample size 10^6 . It is clear that for moderate sample sizes the causal Dantzig estimator can be dramatically improved by causal regularization.

Consider the out-of-sample risk of the causal Dantzig and causal regularization as we increase separation between the in-sample distributions, that is, as the magnitude of the in-sample shift is enlarged. Figure 6 compares the in-sample and out-of-sample risk for models obtained via causal regularization and causal Dantzig. All shifted in-sample datasets consist of n examples where the shift was normally distributed i.e., $A \sim \mathcal{N}(0, \lambda I)$ such that $\lambda \in \{0.5, 1, 1.5\}$. Model selection for causal regularization was done by 3-fold cross-validation. As expected, in all cases the out-of-sample risk corresponding to causal regularization is lower than causal Dantzig.

8 Applications

Below we describe two illustrations of causal regularization in more or less realistic settings. This involves both appropriate transformations of variables, the definition of environments, as well as the definition of the system under consideration.



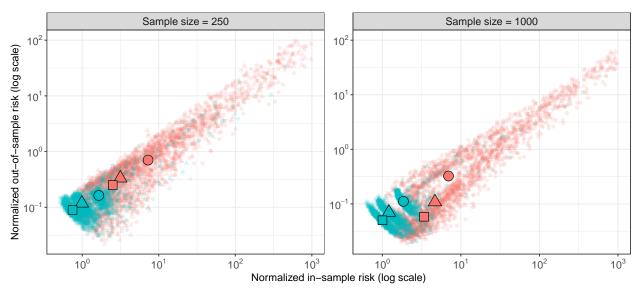


Figure 6: Comparison between causal regularization and the causal Dantzig as the training datasets are progressively shifted. Each faded point is the normalized in-sample and out-of-sample risk achieved by the models. The solid points are medians. The median rather than the mean was chosen due to the high variance of the causal Dantzig estimator.

8.1 Fulton fish market

We apply our methodology to predict the demand in a fish market from the price [Graddy, 1995, Imbens, 2014]. In equilibrium, it has been hypothesized that

$$\log(\text{quantity}) = \beta_0 + \beta_1 \log(\text{price}) + \epsilon$$

where quantity is the daily total quantity of fish in pounds, and price is the average daily price in cents.

We are interested in measuring the prediction quality of the causal regularization and causal Dantzig across the week (Mondays, Tuesdays, Wednesdays and Thursdays). Hence, we split the dataset in a training dataset consisting of Mondays, Tuesdays and Thursdays, and a test dataset composed by Wednesdays. We further divide the training dataset into two datasets: one for stormy days and one for fair days. Stormy days are those when the wind speed is greater than 18 knots and wave height is higher than 4.5 feats. Figure 7 visualizes the covariate and response for each one of the datasets.

We fit causal regularization and the causal Dantzig on the training dataset. Model selection is done by cross-validation with 3 folds. Then, we proceed to compute their risk on 1000 resamples of the test dataset. Figure 8 shows that in each each resample causal

regularization obtains a smaller out-of-sample risk than the causal Dantzig.

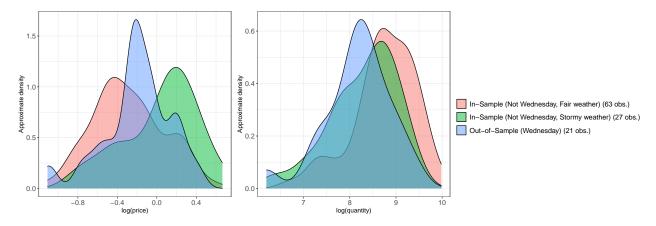


Figure 7: Estimated densities for the amount of fish and its price for each dataset in the Fulton fish market example. This shows that for the covariate, log(price), the observational, fair weather, distribution is indeed different from the shifted, stormy weather, distribution.

8.2 Gene knockout experiments

Next, we consider a dataset of gene expression in yeast under deletion of single genes [Kemmeren et al., 2014, Meinshausen et al., 2016]. There are 262 non-interventional observations, which consists of no gene deletions, and 1479 observations where in each one a different gene is perturbed. In all cases, 6170 genes are measures. The goal is predict the gene expression of one of the genes based on the others. We take as the target the gene that was not directly intervened and whose mean was most shifted between the observation and interventional datasets. Analogously, we choose 10 intervened genes whose mean was most shifted between the interventional and observation datasets as the predictors.

We analyze the performance of causal regularization and the causal Dantzig by nested cross-validation. The interventional data was split in 25 folds; of those 24 were used as a training set, and the remaining as a test set. On the training set, model selection was done via 5-fold cross-validation. Finally, the risk of the estimator was evaluated in the test set. Figure 1 displays the cross-validation selection and the corresponding point in the coefficient path for one of the iterations. The process was repeated leaving one of the 25 folds out at a time. The right-lower pane of figure 1 shows that in all cases, causal regularization achieved a smaller risk than the causal Dantzig.

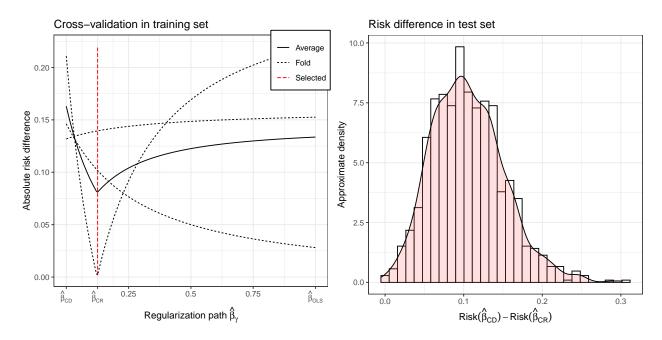


Figure 8: Fulton fish market. (Left) 3-fold cross-validation for causal regularization. The red dashed line indicates the chosen model. (Right) Out-of-sample risk difference between the causal Dantzig and causal regularization. In all cases, the risk achieved by causal regularization is smaller as evidenced by $\operatorname{Risk}(\hat{\beta}_{CD}) - \operatorname{Risk}(\hat{\beta}_{CR})$ always being positive.

9 Conclusion

In this paper, we have introduced *causal regularization*, a technique for trading off in-sample and out-of-sample risk guarantees by exploiting the heterogeneity across in-sample datasets. The method provides out-of-sample guarantees, for any regularization parameter, against specific covariates shifts in the same sense as the causal Dantzig estimator while being identifiable in a broader set of circumstances. Furthermore, we showed theoretically and empirically that cross-validation and sample-splitting can be used to do model selection for causal regularization, and obtain estimators with better out-of-sample performance than the causal Dantzig estimator.

The ideas developed in this work can be extended to some generalized linear models by appropriately defining the risk so that we have risk stability under the causal model. For instance, consider the Poisson regression analogue of definition 1: $Y^A \mid X^A \sim \text{Poisson}(m(\beta))$ where $m(\beta) := \exp(\beta^T \mid X^A \mid)$, and consider the Pearson χ^2 -risk: $R_A(\beta) := \mathbb{E}\left[(Y^A - m(\beta))^2/m(\beta)\right]$. It follows that $R_A(\beta_{PA})$ is constant w.r.t. A, thus we can use causal regularization, as in equation (3), to trade-off overall in-sample fit for risk stability across sub-samples. The out-of-sample risk guarantees are, however, not trivial to derive due to the non-linearity and require further assumptions.

Causal regularization is effective against additive covariate shifts, as specified in definition 1. If the shift between datasets does not follow this pattern, then causal regularization only protects against the projection of the shift to the set of additive shifts.

Acknowledgement

The authors acknowledge funding from the Swiss National Science Foundation (SNSF 188534).

References

- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv:1907.02893, 2019.
- Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
- Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618 2653, 2017. doi: 10.1214/16-AOS1537.
- L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979. doi: 10.1109/TIT. 1979.1056087.
- Vanessa Didelez, Sha Meng, and Nuala A. Sheehan. Assumptions of iv methods for observational epidemiology. *Statist. Sci.*, 25(1):22–40, 02 2010. doi: 10.1214/09-STS316.
- Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005. ISSN 1572-3127. doi: 10.1016/j.stamet.2005.02.003.
- RA Fisher. The design of experiments. Oliver and Boyd, Edinburgh, 1935.
- Kathryn Graddy. Testing for imperfect competition at the fulton fish market. *The RAND Journal of Economics*, 26(1):75–92, 1995. ISSN 07416261. URL http://www.jstor.org/stable/2556036.
- Miguel Hernan and James Robins. Causal inference. CRC Taylor & Francis distributor, Boca Raton, Fla. London, 2010. ISBN 1420076167.
- Guido W. Imbens. Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358, 2014. ISSN 08834237, 21688745.

- Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999. doi: 10.1162/089976699300016304.
- Patrick Kemmeren, Katrin Sameith, Loes A. L. van de Pasch, Joris J. Benschop, Tineke L. Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W. Ko, Sebastiaan van Heesch, Mehdi M. Kashani, Giannis Ampatziadis-Michailidis, Mariel O. Brok, Nathalie A. C. H. Brabers, Anthony J. Miles, Diane Bouwmeester, Sander R. van Hooff, Harm van Bakel, Erik Sluiters, Linda V. Bakker, Berend Snel, Philip Lijnzaad, Dik van Leenen, Marian J. A. Groot Koerkamp, and Frank C. P. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell, 157(3):740–752, April 2014. ISSN 0092-8674. doi: 10.1016/j.cell.2014.02.054.
- Lucas Kook, Beate Sick, and Peter Bühlmann. Distributional anchor regression, 2021. URL https://arxiv.org/abs/2101.08224.
- Nicolai Meinshausen, Alain Hauser, Joris M. Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016. doi: 10.1073/pnas.1510493113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1510493113.
- Michael Oberst, Nikolaj Thams, Jonas Peters, and David Sontag. Regularizing towards causal invariance: Linear models with proxies. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8260–8270. PMLR, 18–24 Jul 2021.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society:* Series B (Statistical Methodology), 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.
- M. Planitz. Inconsistent systems of linear equations. *The Mathematical Gazette*, 63(425): 181–185, 1979. ISSN 00255572. URL http://www.jstor.org/stable/3617890.
- Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *J. Mach. Learn. Res.*, 19(1):1309–1342, January 2018. ISSN 1532-4435.
- Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *CoRR*, abs/2010.05761, 2020.

- Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society:* Series B (Statistical Methodology), 83(2):215–246, 2021.
- Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal dantzig: Fast inference in linear structural equation models with hidden variables under additive interventions. *Ann. Statist.*, 47(3):1688–1722, 06 2019. doi: 10.1214/18-AOS1732.
- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. SIAM Review, 19(4):634–662, 1977. ISSN 00361445. URL http://www.jstor.org/stable/2030248.
- Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- Aad W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006. doi: doi: 10.1524/stnd.2006.24.3.351.
- Silvia Villa, Lorenzo Rosasco, and Tomaso Poggio. On learnability, complexity and stability. In *Empirical Inference*, pages 59–69. Springer, 2013.
- Larry Wasserman. All of Nonparametric Statistics (Springer Texts in Statistics). Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387251456.

A Proofs

A.1 Population statements

Lemma 1 (Worst risk decomposition).

$$\forall \beta \in \mathbb{R}^{P} \quad \sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta) = \frac{1}{2} R_{+}(\beta) + \frac{1+2\tau}{2} R_{\Delta}(\beta),$$

where
$$R_{+}(\beta) := R_{A}(\beta) + R_{0}(\beta)$$
 is the pooled risk $R_{\Delta}(\beta) := R_{A}(\beta) - R_{0}(\beta)$ is the risk difference

Proof. Let $\tilde{A} \in C_{1+\tau}$, by the SEM definition 1, it holds that

$$\begin{bmatrix} \mathbf{Y}^{\tilde{\mathbf{A}}} \\ \mathbf{X}^{\tilde{\mathbf{A}}} \end{bmatrix} = (I - B)^{-1} \left(\epsilon + \begin{bmatrix} 0 \\ \tilde{\mathbf{A}} \end{bmatrix} \right) \tag{11}$$

Let $P_X \in \mathbb{R}^{p \times (p+1)}$ and $P_Y \in \mathbb{R}^{1 \times (p+1)}$ be the projections of the X and Y coordinates, i.e., $P_X = \begin{bmatrix} \mathbb{O}_{p \times 1} & I_{p \times p} \end{bmatrix}$ and $P_Y = \begin{bmatrix} 1 & \mathbb{O}_{1 \times p} \end{bmatrix}$, the population residuals can be rewritten as a projection of equation (11)

$$\mathbf{Y}^{\tilde{\mathbf{A}}} - \boldsymbol{\beta}^T \mathbf{X}^{\tilde{\mathbf{A}}} = u \left(\epsilon + \begin{bmatrix} 0 \\ \tilde{\mathbf{A}} \end{bmatrix} \right)$$
 where $u := (\mathbf{P}_{\mathbf{Y}} - \boldsymbol{\beta}^T \mathbf{P}_{\mathbf{X}})(I - B)^{-1}$

It follows that the residuals under the shift \tilde{A} are

$$R_{\tilde{A}}(\beta) = \mathbb{E}[(Y^{\tilde{A}} - \beta^T X^{\tilde{A}})^2] = u \left(\mathbb{C}[\epsilon] + \begin{bmatrix} 0 & \mathbb{O}^T \\ \mathbb{O} & \mathbb{E}[\tilde{A} \tilde{A}^T] \end{bmatrix} \right) u^T$$
 (12)

where we have exploited the fact that the shift variable and the noise are uncorrelated, i.e., $\mathbb{C}[\tilde{A}, \epsilon] = 0$. Taking the supremum over $C_{1+\tau}$, we obtain

$$\sup_{\tilde{\mathbf{A}} \in \mathcal{C}_{1+\tau}} \mathcal{R}_{\tilde{\mathbf{A}}}(\beta) = u \left(\mathbb{C}[\epsilon] + (1+\tau) \begin{bmatrix} 0 & \mathbb{O}^T \\ \mathbb{O} & \mathbb{E}[\mathbf{A} \, \mathbf{A}^T] \end{bmatrix} \right) u^T$$
 (13)

Noting that equation (12) implies that

$$R_{\Delta}(\beta) = u \left(\begin{bmatrix} 0 & \mathbb{O}^T \\ \mathbb{O} & \mathbb{E}[A A^T] \end{bmatrix} \right) u^T$$

and replacing it back into equation (13), we get

$$\sup_{\tilde{\mathbf{A}} \in \mathcal{C}_{1+\tau}} \mathbf{R}_{\tilde{\mathbf{A}}}(\beta) = \mathbf{R}_{\mathbf{A}}(\beta) + \tau \, \mathbf{R}_{\Delta}(\beta) = \frac{1}{2} \, \mathbf{R}_{+}(\beta) + \frac{1+2\tau}{2} \, \mathbf{R}_{\Delta}(\beta)$$

Proposition 3. G_{Δ} is positive semi-definite and rank $G_{\Delta} = \operatorname{rank} \mathbb{E}[A A^{T}]$

Proof. By definition

$$G_{\Delta} = \mathbb{E}[X^{A} X^{A^{T}}] - \mathbb{E}[X^{0} X^{0^{T}}]$$

Let $P_X \in \mathbb{R}^{p \times (p+1)}$ be the projection of the X coordinates, i.e., $P_X = \begin{bmatrix} \mathbb{O}_{p \times 1} & I_{p \times p} \end{bmatrix}$, then for $\tilde{A} \in \{A, 0\}$

$$\mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \mathbf{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}}] = u \left(\mathbb{C}[\epsilon] + \begin{bmatrix} 0 & \mathbb{O}^{T} \\ \mathbb{O} & \mathbb{E}[\tilde{\mathbf{A}} \tilde{\mathbf{A}}^{\mathrm{T}}] \end{bmatrix} \right) u^{T} \text{ where } u = \mathbf{P}_{\mathbf{X}} (I - B)^{-1}$$
(14)

since $\mathbb{C}[\tilde{A}, \epsilon] = 0$. Let $(I - B)^{-1} := \begin{bmatrix} 1 & v^T \\ w & M \end{bmatrix}$, then $u = P_X(I - B)^{-1} = \begin{bmatrix} w & M \end{bmatrix}$ and

$$\mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \mathbf{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}}] = u \, \mathbb{C}[\epsilon] u^{T} + M \, \mathbb{E}[\tilde{\mathbf{A}} \, \tilde{\mathbf{A}}^{\mathrm{T}}] \, M^{T}$$

Thus, $G_{\Delta} = M \mathbb{E}[A A^T] M^T$. We can explicitly write the inverse of M by noting that

$$(I - B)^{-1}(I - B) = I_{(p+1)\times(p+1)} \implies \begin{cases} -w \,\beta_{PA}^{T} + M(I - B_X) &= I_{p\times p} \\ w - M \,\beta_{CH} &= 0 \end{cases}$$
(15)

$$\implies M\left((I - \mathbf{B}_{\mathbf{X}}) - \beta_{\mathbf{CH}} \,\beta_{\mathbf{PA}}^{T}\right) = I_{p \times p} \tag{16}$$

Thus, M is square and has a right-inverse, hence M is invertible. In other words, M is a product of elementary matrices and doens't modify the column-rank or the row-rank of any conformal matrix. Ergo,

$$\operatorname{rank} \mathbf{G}_{\Delta} = \operatorname{rank} M \mathbb{E}[\mathbf{A} \mathbf{A}^{\mathrm{T}}] M^{T} = \operatorname{rank} \mathbb{E}[\mathbf{A} \mathbf{A}^{\mathrm{T}}]$$

Finally, note that under the assumption that $\mathbb{E}[A A^T]$ is positive semi-definite, it follows that G_{Δ} is positive semi-definite since

$$\forall \beta \in \mathbb{R}^{\mathbf{p}} \, \beta^T \, \mathbf{G}_{\Delta} \, \beta = \left(\boldsymbol{M}^T \boldsymbol{\beta} \right)^T \mathbb{E}[\mathbf{A} \, \mathbf{A}^T](\boldsymbol{M}^T \boldsymbol{\beta}) \geq 0$$

Proposition 4 (The risk difference is a centred quadratic form).

$$\forall \beta \in \mathbb{R}^{P} \ \mathrm{R}_{\Delta}(\beta) = (\beta - \beta_{\mathrm{PA}})^{T} \mathrm{G}_{\Delta}(\beta - \beta_{\mathrm{PA}})$$

Proof. For $\tilde{A} \in \{A, 0\}$

$$R_{\tilde{\mathbf{A}}}(\beta) = \mathbb{E}[(\mathbf{Y}^{\tilde{\mathbf{A}}} - \beta^T \mathbf{X}^{\tilde{\mathbf{A}}})^2] = \beta^T \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \mathbf{X}^{\tilde{\mathbf{A}}^T}] \beta - 2\beta^T \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \mathbf{Y}^{\tilde{\mathbf{A}}}] + \mathbb{E}[(\mathbf{Y}^{\tilde{\mathbf{A}}})^2]$$

Recalling the definitions

$$G_{\Delta} = \mathbb{E}[X^{A} X^{A^{T}}] - \mathbb{E}[X^{0} X^{0^{T}}]$$

$$Z_{\Delta} = \mathbb{E}[X^{A} Y^{A}] - \mathbb{E}[X^{0} Y^{0}] = G_{\Delta} \beta_{PA}$$
By equation (2)

It holds that

$$R_{\Delta}(\beta) = R_{A}(\beta) - R_{0}(\beta)$$

$$= \beta^{T} G_{\Delta} \beta - 2\beta^{T} G_{\Delta} \beta_{PA} + \mathbb{E}[(Y^{A})^{2}] - \mathbb{E}[(Y^{0})^{2}]$$
(17)

Notice that for $\tilde{A} \in \{A, 0\}$, $\mathbb{E}[(Y^{\tilde{A}})^2]$ can be expanded as follows

$$\begin{split} \mathbb{E}[(\mathbf{Y}^{\tilde{\mathbf{A}}})^{2}] &= \mathbb{E}[(\beta_{\mathrm{PA}}^{\mathrm{T}} \, \mathbf{X}^{\tilde{\mathbf{A}}} + \epsilon_{Y})^{2}] \\ &= \beta_{\mathrm{PA}}^{\mathrm{T}} \, \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}}] \, \beta_{\mathrm{PA}} + \mathbb{E}[\epsilon_{Y}^{2}] + \beta_{\mathrm{PA}}^{\mathrm{T}} \, \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \, \epsilon_{Y}] \\ &= \beta_{\mathrm{PA}}^{\mathrm{T}} \, \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}}] \, \beta_{\mathrm{PA}} + \mathbb{E}[\epsilon_{Y}^{2}] + \beta_{\mathrm{PA}}^{\mathrm{T}} \, u \, \mathbb{E}[(\epsilon + \tilde{\mathbf{A}}) \epsilon_{Y}] \\ &= \beta_{\mathrm{PA}}^{\mathrm{T}} \, \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}}] \, \beta_{\mathrm{PA}} + \mathbb{E}[\epsilon_{Y}^{2}] + \beta_{\mathrm{PA}}^{\mathrm{T}} \, u \, \mathbb{C}[\epsilon, \epsilon_{Y}] \qquad \text{since } \mathbb{E}[\tilde{\mathbf{A}} \, \epsilon_{Y}] = 0 \end{split}$$

hence

$$\mathbb{E}[(\mathbf{Y}^{\mathbf{A}})^{2}] - \mathbb{E}[(\mathbf{Y}^{\mathbf{0}})^{2}] = \beta_{\mathbf{P}\mathbf{A}}^{\mathbf{T}} \mathbf{G}_{\Delta} \beta_{\mathbf{P}\mathbf{A}}$$

plugging the result back into equation (17), we get

$$R_{\Delta}(\beta) = \beta^{T} G_{\Delta} \beta + \beta^{T} G_{\Delta} \beta_{PA} + \beta_{PA}^{T} G_{\Delta} \beta_{PA}$$
$$= (\beta - \beta_{PA})^{T} G_{\Delta}(\beta - \beta_{PA})$$

Theorem 1 (Moore–Penrose inverse). Given a $A \in \mathbb{R}^{p \times p}$, the Moore–Penrose inverse A is defined as the unique matrix $A^g \in \mathbb{R}^{p \times p}$ that satisfies the following properties

$$A^g A A^g = A \tag{1}$$

$$AA^gA = A \tag{2}$$

$$(AA^g)^T = AA^g \tag{3}$$

$$\left(A^g A\right)^T = A^g A \tag{4}$$

Proposition 1 (Convex regularizer).

$$R_{\Delta}(\beta) = R_{||\cdot||}(\beta) \quad s.t. \quad R_{||\cdot||}(\beta) := \left\| (G_{\Delta}^{g/2})^T (G_{\Delta} \beta - Z_{\Delta}) \right\|_2^2$$

where $G_{\Delta}^{g/2}$ is the Moore-Penrose pseudoinverse of the square root of G_{Δ} , i.e., the matrix that satisfies $G_{\Delta} = (G_{\Delta}^{g/2})^T G_{\Delta}^{g/2}$.

Proof. Note that

$$R_{\Delta}(\beta) = (\beta - \beta_{PA})^{T} G_{\Delta}(\beta - \beta_{PA}) = \left\| G_{\Delta}^{1/2}(\beta - \beta_{PA}) \right\|_{2}^{2}$$
(18)

where the first equality holds by proposition 4, and in the last equality we applied the square root definition. The existence of $G_{\Delta}^{1/2}$ positive semi-definite is guaranteed by the positive semi-definiteness of G_{Δ} , see proposition 3. Let $G_{\Delta}^{g/2}$ be the pseudoinverse of $G_{\Delta}^{1/2}$, we rewrite the regularizer as follows,

$$\begin{aligned} \mathbf{G}_{\Delta}^{1/2} &= \mathbf{G}_{\Delta}^{1/2} \, \mathbf{G}_{\Delta}^{g/2} \, \mathbf{G}_{\Delta}^{1/2} & \text{by theorem 1.2} \\ &= \left(\mathbf{G}_{\Delta}^{1/2} \, \mathbf{G}_{\Delta}^{g/2}\right)^T \, \mathbf{G}_{\Delta}^{1/2} & \text{by theorem 1.3} \\ &= \left(\mathbf{G}_{\Delta}^{g/2}\right)^T \left(\mathbf{G}_{\Delta}^{1/2}\right)^T \, \mathbf{G}_{\Delta}^{1/2} & \\ &= \left(\mathbf{G}_{\Delta}^{g/2}\right)^T \, \mathbf{G}_{\Delta} & \text{by square root def.} \end{aligned}$$

plugging the result back into equation (18), and using equation (2) in the last equality, we get

$$R_{\Delta}(\beta) = \left\| \left(G_{\Delta}^{g/2} \right)^T \left(G_{\Delta} \beta - G_{\Delta} \beta_{PA} \right) \right\|_2^2 = \left\| \left(G_{\Delta}^{g/2} \right)^T \left(G_{\Delta} \beta - Z_{\Delta} \right) \right\|_2^2$$

Corollary 1 (Interpolation by causal regularization). Let $\lambda \in [0, \infty)$ and $G_+ := \mathbb{E}(X^A X^{A^T}) + \mathbb{E}(X^0 X^{O^T})$, if G_+ is non-singular, then

$$\beta_{\lambda} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{p}} \frac{1}{2} \operatorname{R}_{+}(\beta) + \frac{\lambda}{2} \operatorname{R}_{\Delta}(\beta) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{p}} \left\| \operatorname{G}_{\lambda} \beta - \operatorname{Z}_{\lambda} \right\|_{2}^{2} = \operatorname{G}_{\lambda}^{-1} \operatorname{Z}_{\lambda}$$

where $G_{\lambda} := G_{+} + \lambda G_{\Delta}$ and $Z_{\lambda} := Z_{+} + \lambda Z_{\Delta}$. In particular, $\beta_{0} = G_{+}^{-1} Z_{+}$ is the population ordinary least squares. Additionally, if G_{Δ} is non-singular, then $\beta_{\infty} := \lim_{\lambda \to \infty} \beta_{\lambda} = G_{\Delta}^{-1} Z_{\Delta}$ is the causal Dantzig.

Proof. For $\lambda \in [0, \infty)$

$$\beta_{\lambda} := \underset{\beta \in \mathbb{R}^{p}}{\operatorname{arg \, min}} \frac{1}{2} \, \mathrm{R}_{+}(\beta) + \frac{\lambda}{2} \, \mathrm{R}_{\Delta}(\beta)$$

$$= \underset{\beta \in \mathbb{R}^{p}}{\operatorname{arg \, min}} \frac{1}{2} \beta^{T} \, \mathrm{G}_{\lambda} \, \beta - \beta^{T} (\mathrm{Z}_{+} + \lambda \, \mathrm{G}_{\Delta} \, \beta_{\mathrm{PA}}) \qquad \text{by proposition 4}$$

$$= \underset{\beta \in \mathbb{R}^{p}}{\operatorname{arg \, min}} \frac{1}{2} \beta^{T} \, \mathrm{G}_{\lambda} \, \beta - \beta^{T} (\mathrm{Z}_{+} + \lambda \, \mathrm{Z}_{\Delta}) \qquad \text{by equation 2}$$

$$= \mathrm{G}_{\lambda}^{-1} \, \mathrm{Z}_{\lambda}$$

Note that G_{λ}^{-1} always exists since G_{λ} is positive definite because G_{+} is assumed non-singular, and therefore positive definite, and G_{Δ} is positive semi-definite by proposition 3. Given the uniqueness of the solution, we can express β_{λ} as the least squares solution of $G_{\lambda}\beta = Z_{\lambda}$, i.e. $\beta_{\lambda} := \arg\min_{\beta \in \mathbb{R}^p} \|G_{\lambda}\beta - Z_{\lambda}\|_2$. Finally, if G_{Δ} is full rank, $\beta_{\infty} = \arg\min_{\beta \in \mathbb{R}^p} \|G_{\Delta}\beta - Z_{\Delta}\|_2$, that is, causal regularization at $\gamma = \infty$ and the causal Dantzig have the same solution, see equation (3).

A.2 Consistency

Definition 5. Let $A \in \mathbb{R}^{p \times p}$ be a symmetric matrix, its operator norm can be defined as

$$||A||_2 \coloneqq \max_{x \in \mathbb{R}^p} \max_{s.t. \ ||x||_2 \le 1} |x^T A x|$$

and it holds that

$$||A||_2 = \max_{1 \le i \le p} |\gamma_i(A)|$$

where $\gamma_i(A)$ is the i-th eigenvalue of A.

Theorem 2 (Stewart [1977], Theorem 3.3). Let $A \in \mathbb{R}^{p \times p}$ and $B \in \mathbb{R}^{p \times p}$, then

$$||A^g - B^g||_2 \le C \cdot \max\{||A^g||_2^2, ||B^g||_2^2\} \cdot ||A - B||_2^2$$

where A, $\|\cdot\|_2$ denotes the opetor norm, and C is a positive universal constant.

Lemma 4. Let $A \in \mathbb{R}^{p \times p}$ be a positive symmetric define matrix, and $\{B_n\}_{n=1}^{\infty} \in \mathbb{R}^{p \times p}$ a sequence of symmetric matrices such that $\|A - B_n\|_2 \stackrel{p}{\to} 0$, then $\|A^g - B_n^g\|_2 \stackrel{p}{\to} 0$

Proof. By theorem 2, it holds that

$$||A^g - B_n^g||_2 \le C \cdot \max\{||A^g||_2^2, ||B_n^g||_2^2\} \cdot ||A - B_n||_2^2$$

Since A is a positive symmetric definite matrix, it has an inverse, and it follows that

$$||A^g||_2 = ||A^{-1}||_2 = \gamma_{\min}(A)^{-1}$$

where $\gamma_{\min}(A)$ is the smallest positive eigenvalue of A. Let UDV^T be the SVD decomposition of B_n , it holds that

$$B_n^g = UD^gV^T$$
 where $\gamma_i(D^g) = \begin{cases} 0 & \text{if } \gamma_i(B_n) = 0\\ |\gamma_i(B_n)|^{-1} & \text{otherwise} \end{cases}$

and we rewrite the operator norm of B_n as

$$\|B_n^g\|_2 = \|UD^gV^T\|_2 = \|D^g\|_2 = \begin{cases} 0 & \text{if } \gamma_i(B_n) = 0 \ \forall 1 \le i \le p \\ \max\{ \ |\gamma_i(B_n)|^{-1} \ \text{s.t.} \ |\gamma_i(B_n)| > 0 \ \} & \text{otherwise} \end{cases}$$

where the second equality is due to the operator norm being unitarily invariant.

Weyl's inequality says that if B_n converges to A in operator norm, then their eigenvalues converge uniformly,

$$|\gamma_i(A) - |\gamma_i(B_n)|| \le |\gamma_i(A) - \gamma_i(B_n)| \le ||A - B_n||_2$$
 for $1 \le i \le p$

Thus, a lowerbound for $|\gamma_i(B)|$ is

$$\gamma_{\min}(A) - ||A - B_n||_2 \le \gamma_i(A) - ||A - B_n||_2 \le |\gamma_i(B_n)|$$
 for $1 \le i \le p$

Define the event E_n as

$$E_n := \{\gamma_{\min}(A) - ||A - B_n||_2 > 0\}$$

it follows that on the event E_n ,

$$||B_n^g||_2 \le \frac{1}{\gamma_{\min}(A) - ||A - B_n||_2}$$

Hence, if the event E_n happens, it holds that

$$||A^{g} - B_{n}^{g}||_{2} \le C \cdot \max\{\frac{1}{\gamma_{\min}(A)^{2}}, \frac{1}{(\gamma_{\min}(A) - ||A - B_{n}||_{2})^{2}}\} \cdot ||A - B_{n}||_{2}^{2}$$

$$\le g(||A - B_{n}||_{2}) \quad \text{where } g(x) := \frac{x^{2}}{(\gamma_{\min}(A) - x)^{2}}$$
(19)

Putting all together we get that for $\epsilon > 0$ and $\delta > 0$

$$P(||A^{g} - B_{n}^{g}||_{2} \ge \epsilon) = P(||A^{g} - B_{n}^{g}||_{2} \ge \epsilon \land E_{n}|) + P(||A^{g} - B_{n}^{g}||_{2} \ge \epsilon \land E_{n}^{c})$$

$$\le P(|g(||A - B_{n}||_{2}) > \epsilon) + P(||E_{n}^{c}|)$$

$$= P(|g(||A - B_{n}||_{2}) > \epsilon) + P(||A - B_{n}||_{2} \ge \gamma_{\min}(A))$$

Where the inequality is due to equation (19). Note that g is continuous for $x \neq \gamma_{\min}(A)$ and g(0) = 0. Since $||A - B_n||_2 \stackrel{\text{p}}{\to} 0$ and $\gamma_{\min}(A) > 0$, by the continuous mapping theorem $g(||A - B_n||_2) \stackrel{\text{p}}{\to} 0$. Hence, we can choose $N := N(\epsilon, \gamma_{\min}(A), \delta) \in \mathbb{N}$ such that $P(g(||A - B_n||_2) > \epsilon) < \frac{\delta}{2}$ and $P(||A - B_n||_2 \ge \gamma_{\min}(A)) < \frac{\delta}{2}$ for all $n \ge N$, which implies that $P(||A^g - B_n^g||_2 \ge \epsilon) < \delta$ for all $n \ge N$ and consequently $||A^g - B_n^g||_2 \stackrel{\text{p}}{\to} 0$

Proposition 2. If G_+ is positive definite, then $\hat{\beta}_{\lambda} \xrightarrow{p} \beta_{\lambda}$ for $\lambda \in [0, \infty)$

Proof. By the weak law of large numbers and continuity, \mathbb{G}_+ , \mathbb{G}_Δ , \mathbb{Z}_+ , and \mathbb{Z}_Δ converge in probability to G_+ , G_Δ , Z_+ , and Z_Δ correspondingly. Hence, by continuity, $\mathbb{G}_\lambda := \mathbb{G}_+ + \lambda \, \mathbb{G}_\Delta$ and $\mathbb{Z}_\lambda := \mathbb{Z}_+ + \lambda \, \mathbb{Z}_\Delta$ converge in probability to G_λ and Z_λ .

Since p is finite, $\mathbb{G}_{\lambda} \stackrel{p}{\to} \mathbb{G}_{\lambda}$ implies $\|\mathbb{G}_{\lambda} - \mathbb{G}_{\lambda}\|_{2} \stackrel{p}{\to} 0$. Then, theorem 2 implies that $\|\mathbb{G}_{\lambda}^{g} - \mathbb{G}_{\lambda}^{g}\|_{2} \stackrel{p}{\to} 0$, and consequently $\mathbb{G}_{\lambda}^{g} \stackrel{p}{\to} \mathbb{G}_{\lambda}^{g}$. It follows that

$$\hat{\beta}_{\lambda} \coloneqq \mathbb{G}^{\mathrm{g}}_{\lambda} \, \mathbb{Z}_{\lambda} \overset{\mathrm{p}}{\to} \, \mathrm{G}^{\mathrm{g}}_{\lambda} \, \mathrm{Z}_{\lambda} = \mathrm{G}^{-1}_{\lambda} \, \mathrm{Z}_{\lambda} = \beta_{\lambda}$$

where the convergence in probability is due to continuity, and the penultimate equality is due to G_{λ} being positive definite since G_{+} is assumed positive definite, and G_{Δ} is positive semi-definite by proposition 3.

A.3 Finite sample bounds

Proposition 5. For $\tilde{A} \in A$,

$$||R_{\tilde{A}}(\beta) - \hat{R}_{\tilde{A}}(\beta)| \le (||\beta||_1^2 + 1) V[Y^{\tilde{A}}] \left(\max_{1 \le k \le p} V[X_k^{\tilde{A}}] \right) \left(\sqrt{\frac{4q + 8 \log(p)}{n_{\tilde{A}}}} + \frac{4q + 8 \log(p)}{n_{\tilde{A}}} \right)$$

with probability exceeding $1 - 2e^{-q}$ for q > 0, hence $|R_{\tilde{A}}(\beta) - \hat{R}_{\tilde{A}}(\beta)| = o_p(\frac{1}{\sqrt{n_{\tilde{A}}}})$

$$\begin{aligned} \textit{Proof.} \ \ \text{Let} \ \mathbf{V}^{\tilde{\mathbf{A}}} &:= \begin{bmatrix} \mathbf{Y}^{\tilde{\mathbf{A}}} \\ \mathbf{X}^{\tilde{\mathbf{A}}} \end{bmatrix}, \ \mathbf{S}^{\tilde{\mathbf{A}}} &= \mathbb{E}[\mathbf{V}^{\tilde{\mathbf{A}}} \mathbf{V}^{\tilde{\mathbf{A}}^T}], \ \text{and} \ \boldsymbol{\upsilon} := \begin{bmatrix} -1 \\ \boldsymbol{\beta} \end{bmatrix} \text{then} \\ & |\mathbf{R}_{+}(\boldsymbol{\beta}) - \hat{\mathbf{R}}_{+}(\boldsymbol{\beta})| = |\boldsymbol{\upsilon}^T (\hat{\mathbf{S}}^{\tilde{\mathbf{A}}} - \mathbf{S}^{\tilde{\mathbf{A}}}) \boldsymbol{\upsilon}| \leq ||\boldsymbol{\upsilon}||_{1}^{2} \left\| \hat{\mathbf{S}}^{\tilde{\mathbf{A}}} - \mathbf{S}^{\tilde{\mathbf{A}}} \right\|_{\infty} \leq (||\boldsymbol{\beta}||_{1}^{2} + 1) \left\| \hat{\mathbf{S}}^{\tilde{\mathbf{A}}} - \mathbf{S}^{\tilde{\mathbf{A}}} \right\|_{\infty} \end{aligned}$$

where $\|\hat{\mathbf{S}}^{\tilde{\mathbf{A}}} - \mathbf{S}^{\tilde{\mathbf{A}}}\|_{\infty} = \max_{i,j \in \{1,\dots,p+1\}} |\hat{S}_{ij}^{\tilde{\mathbf{A}}} - S_{ij}^{\tilde{\mathbf{A}}}|$. Note that due to their block structure, it follows that

$$m_{1} := \left\| \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \, \mathbb{X}^{\tilde{\mathbf{A}}^{\mathsf{T}}} \, \mathbb{X}^{\tilde{\mathbf{A}}} - \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{X}^{\tilde{\mathbf{A}}^{\mathsf{T}}}] \right\|_{\infty}$$
where $\left\| \hat{\mathbf{S}}^{\tilde{\mathbf{A}}} - \mathbf{S}^{\tilde{\mathbf{A}}} \right\|_{\infty} = \max(m_{1}, m_{2}, m_{3}) \text{ s.t. } m_{2} := \left\| \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \, \mathbb{X}^{\tilde{\mathbf{A}}^{\mathsf{T}}} \, \mathbb{Y}^{\tilde{\mathbf{A}}} - \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{Y}^{\tilde{\mathbf{A}}}] \right\|_{\infty}$

$$m_{3} := \left\| \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \, \mathbb{Y}^{\tilde{\mathbf{A}}^{\mathsf{T}}} \, \mathbb{Y}^{\tilde{\mathbf{A}}} - \mathbb{E}[(\mathbf{Y}^{\tilde{\mathbf{A}}})^{2}] \right\|_{\infty}$$

If $X^{\bar{A}}$ is a multivariate centred Gaussian random variable, it holds that [Van De Geer and Bühlmann, 2009]

$$m_{1} = \left\| \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}} \mathbb{X}^{\tilde{\mathbf{A}}} - \mathbb{E}[\mathbf{X}^{\tilde{\mathbf{A}}} \mathbf{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}}] \right\|_{\infty} \leq \max_{k} \mathbb{V}[\mathbf{X}^{\tilde{\mathbf{A}}}_{k}] \left(\sqrt{\frac{4q + 8\log(p)}{n_{\tilde{\mathbf{A}}}}} + \frac{4q + 8\log(p)}{n_{\tilde{\mathbf{A}}}} \right)$$

with probability exceeding $1-4e^{-q}$ for q>0. Analogously, if Y^A is a multivariate centred Gaussian random variable

$$m_3 = \left\| \frac{1}{n_{\tilde{A}}} \, \mathbb{Y}^{\tilde{A}^{\mathrm{T}}} \, \mathbb{Y}^{\tilde{A}} - \mathbb{E}[(\mathbf{Y}^{\tilde{A}})^2] \right\|_{\infty} \leq \mathbb{V}[\mathbf{Y}^{\tilde{A}}] \left(\sqrt{\frac{4q}{n_{\tilde{A}}}} + \frac{4q}{n_{\tilde{A}}} \right)$$

with probability exceeding $1 - 2e^{-q}$ for q > 0, and

$$m_2 = \left\| \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \, \mathbb{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}} \, \mathbb{Y}^{\tilde{\mathbf{A}}} - \mathbb{E} \big[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{Y}^{\tilde{\mathbf{A}}} \big] \right\|_{\infty} \\ \leq \mathbb{V} \big[\mathbf{Y}^{\tilde{\mathbf{A}}} \big] \max_{k} \mathbb{V} \big[\mathbf{X}^{\tilde{\mathbf{A}}}_{k} \big] \left(\sqrt{\frac{4q + 4 \log(p)}{n_{\tilde{\mathbf{A}}}}} + \frac{4q + 4 \log(p)}{n_{\tilde{\mathbf{A}}}} \right) \right) \\ = \left\| \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \, \mathbb{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}} \, \mathbb{Y}^{\tilde{\mathbf{A}}} - \mathbb{E} \big[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{Y}^{\tilde{\mathbf{A}}} \big] \right\|_{\infty} \\ \leq \mathbb{V} \big[\mathbf{Y}^{\tilde{\mathbf{A}}} \big] \\ = \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}} \, \mathbb{Y}^{\tilde{\mathbf{A}}} - \mathbb{E} \big[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{Y}^{\tilde{\mathbf{A}}} \big] \right\|_{\infty} \\ \leq \mathbb{V} \big[\mathbf{Y}^{\tilde{\mathbf{A}}} \big] \\ = \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{X}^{\tilde{\mathbf{A}}^{\mathrm{T}}} \, \mathbb{Y}^{\tilde{\mathbf{A}}} - \mathbb{E} \big[\mathbf{X}^{\tilde{\mathbf{A}}} \, \mathbf{Y}^{\tilde{\mathbf{A}}} \big] \right\|_{\infty} \\ \leq \mathbb{V} \big[\mathbf{Y}^{\tilde{\mathbf{A}}} \big] \\ = \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} \Big] \\ = \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} \Big] \\ = \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} \Big] \\ = \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \mathbb{Y}^{\tilde{\mathbf{A}}} \Big] \\ = \frac{1}{\mathbf{n}_{\tilde{\mathbf{A}}}} \mathbb{Y}^{\tilde{\mathbf{A}}} + \mathbb{Y}^{\tilde{\mathbf{A}}} +$$

with probability exceeding $1 - 2e^{-q}$ for q > 0. Therefore,

$$\left\|\hat{\mathbf{S}}^{\tilde{\mathbf{A}}} - \mathbf{S}^{\tilde{\mathbf{A}}}\right\|_{\infty} \leq \mathbb{V}[\mathbf{Y}^{\tilde{\mathbf{A}}}] \max_{k} \mathbb{V}[\mathbf{X}^{\tilde{\mathbf{A}}}_{k}] \left(\sqrt{\frac{4q + 8\log(p)}{n_{\tilde{\mathbf{A}}}}} + \frac{4q + 8\log(p)}{n_{\tilde{\mathbf{A}}}}\right)$$

with probability exceeding $1 - 2e^{-q}$ for q > 0, and

$$\left\| \hat{\mathbf{S}}^{\tilde{\mathbf{A}}} - \mathbf{S}^{\tilde{\mathbf{A}}} \right\|_{\infty} = o_p \left(1 / \sqrt{n_{\tilde{\mathbf{A}}}} \right)$$

Proposition 6.

$$|R_{\Delta}(\beta) - \hat{R}_{\Delta}(\beta)| \le (||\beta||_{1}^{2} + 1) \varphi(p, n_{A}, n_{0})$$

 $|R_{+}(\beta) - \hat{R}_{+}(\beta)| \le (||\beta||_{1}^{2} + 1) \varphi(p, n_{A}, n_{0})$

with probability exceeding $1 - 2e^{-q}$ for q > 0, where

$$\varphi(p, \mathbf{n_A}, \mathbf{n_0}) \coloneqq \sum_{\tilde{\mathbf{A}} \in \{\mathbf{A}, \mathbf{0}\}} \mathbb{V}[\mathbf{Y}^{\tilde{\mathbf{A}}}] \max_{k} \mathbb{V}[\mathbf{X}^{\tilde{\mathbf{A}}}_{k}] \left(\sqrt{\frac{4q + 8\log(p)}{n_{\tilde{\mathbf{A}}}}} + \frac{4q + 8\log(p)}{n_{\tilde{\mathbf{A}}}} \right)$$

hence $\varphi(p, n_A, n_0) = o_p \left(1 / \sqrt{\min(n_A, n_0)} \right)$

Proof.

$$|R_{\Delta}(\beta) - \hat{R}_{\Delta}(\beta)| \le |R_{\Delta}(\beta) - \hat{R}_{\Delta}(\beta)| + |R_{0}(\beta) - \hat{R}_{0}(\beta)| \le (||\beta||_{1}^{2} + 1)\varphi(p, n_{\Delta}, n_{0})$$

where in the last inequality we used proposition 5. The proof for $|R_{+}(\beta) - \hat{R}_{+}(\beta)|$ is analogous.

Lemma 2 (Finite sample bound for worst population risk). If \mathbb{X}^0 , \mathbb{Y}^0 , \mathbb{X}^A and \mathbb{Y}^A are multivariate centred Gaussian variables

$$\forall \beta \in \mathbb{R}^{P} : \sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta) \leq \frac{1}{2} \hat{R}_{+}(\beta) + \frac{1+2\tau}{2} \hat{R}_{\Delta}(\beta) + \eta(\mathbf{n})$$
 (6)

with probability exceeding $1 - 2e^{-q}$ for q > 0, where

$$\eta(\mathbf{n}) \coloneqq (1+\tau)(\|\beta\|_1^2 + 1) \varphi_+(p, \mathbf{n})
\varphi_+(p, \mathbf{n}) \coloneqq \varphi(p, \mathbf{n}_{\mathbf{A}}) + \varphi(p, \mathbf{n}_{\mathbf{0}}) \quad and
\varphi(p, \mathbf{n}_{\mathbf{U}}) \coloneqq \mathbb{V}[\mathbf{Y}^{\mathbf{U}}] \left(\max_{1 \le k \le p} \mathbb{V}[\mathbf{X}_{\mathbf{k}}^{\mathbf{U}}] \right) \left(\sqrt{\frac{4q + 8\log(p)}{n_{\mathbf{U}}}} + \frac{4q + 8\log(p)}{n_{\mathbf{U}}} \right)$$

Hence $\eta(\mathbf{n}) = o_p \left(1 / \sqrt{\min(\mathbf{n}_A, \mathbf{n}_0)} \right)$

Proof.

$$\begin{split} \sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta) \\ &= \frac{1}{2} R_{+}(\beta) + \frac{1+2\tau}{2} R_{\Delta}(\beta) \\ &\leq \frac{1}{2} \hat{R}_{+}(\beta) + \frac{1+2\tau}{2} \hat{R}_{\Delta}(\beta) + (1+\tau) \varphi(\beta) \end{split} \qquad \text{by lemma 1}$$

A.4 Model selection

Corollary 4 (Loss concentration).

$$\forall \lambda \in \Lambda \quad |\tilde{\theta}_{S}(\lambda) - \hat{\theta}_{S}(\lambda)| \leq \mathbb{E}_{S} \, \tilde{\varphi}(\lambda)$$

with probability exceeding $1 - 2e^{-q}$ for q > 0,

where
$$\tilde{\varphi}(\lambda) := (\|\hat{\beta}_{\lambda}^{S,0}\|_{1}^{2} + 1)\varphi(p, n_{A}(S^{A}, 1), n_{O}(S^{O}, 1))$$

Proof.

$$\begin{split} \tilde{\theta}_{S}(\lambda) &= \mathbb{E}_{S} \, \mathrm{R}_{\Delta}(\hat{\beta}_{\lambda}^{\mathrm{S},0}) & \text{by definition} \\ &\leq \mathbb{E}_{S} \, \hat{\mathrm{R}}_{\Delta}(\hat{\beta}_{\lambda}^{\mathrm{S},0}) + \mathbb{E}_{S} \, \tilde{\varphi}(\lambda) & \text{by proposition 6} \\ &\leq \mathbb{E}_{S} \, |\, \hat{\mathrm{R}}_{\Delta}(\hat{\beta}_{\lambda}^{\mathrm{S},0})| + \mathbb{E}_{S} \, \tilde{\varphi}(\lambda) & \\ &= \hat{\theta}_{S}(\lambda) + \mathbb{E}_{S} \, \tilde{\varphi}(\lambda) & \text{by definition} \end{split}$$

П

Noting that

$$|R_{\Delta}(\beta) - |\hat{R}_{\Delta}(\beta)|| = ||R_{\Delta}(\beta)| - |\hat{R}_{\Delta}(\beta)|| \le |R_{\Delta}(\beta) - \hat{R}_{\Delta}(\beta)|$$

by the reverse triangle inequality. An analogous proof can be constructed showing that $\hat{\theta}_{S}(\lambda) \leq \tilde{\theta}_{S}(\lambda) + \mathbb{E}_{S} \tilde{\varphi}(\lambda)$ which concludes the proof.

Lemma 3 (Asymptotic equivalence of sample and population selectors). If \mathbb{X}^0 , \mathbb{Y}^0 , \mathbb{X}^A and \mathbb{Y}^A are multivariate centred Gaussian variables

$$|\tilde{\theta}_{S}(\hat{\lambda}_{S}) - \tilde{\theta}_{S}(\tilde{\lambda}_{S})| \stackrel{p}{\rightarrow} 0$$

Proof.

$$\begin{split} \tilde{\theta}_{S}(\tilde{\lambda}_{S}) &\leq \tilde{\theta}_{S}(\hat{\lambda}_{S}) & \text{since } \tilde{\theta}_{S}(\tilde{\lambda}_{S}) \leq \tilde{\theta}_{S}(\lambda) \forall \lambda \in \Lambda \\ &\leq \hat{\theta}_{S}(\hat{\lambda}_{S}) + \mathbb{E}_{S} \tilde{\varphi}(\hat{\lambda}_{S}) & \text{by corollary 4} \\ &\leq \hat{\theta}_{S}(\tilde{\lambda}_{S}) + \mathbb{E}_{S} \tilde{\varphi}(\hat{\lambda}_{S}) & \text{since } \hat{\theta}_{S}(\hat{\lambda}_{S}) \leq \hat{\theta}_{S}(\lambda) \forall \lambda \in \Lambda \\ &\leq \tilde{\theta}_{S}(\tilde{\lambda}_{S}) + \mathbb{E}_{S}[\tilde{\varphi}(\hat{\lambda}_{S}) + \tilde{\varphi}(\tilde{\lambda}_{S})] & \text{by corollary 4} \end{split}$$

Thus,

$$0 \le \tilde{\theta}_{S}(\hat{\lambda}_{S}) - \tilde{\theta}_{S}(\tilde{\lambda}_{S}) \le \mathbb{E}_{S}[\tilde{\varphi}(\hat{\lambda}_{S}) + \tilde{\varphi}(\tilde{\lambda}_{S})] = o_{p}\left(1/\sqrt{\min(n_{A}, n_{0})}\right)$$
(20)

which implies the stated lemma.

Corollary 3 (Asymptotic equivalence of sample and S-optimal selectors). If \mathbb{X}^0 , \mathbb{Y}^A and \mathbb{Y}^A are multivariate centred Gaussian variables

$$|\mathbb{E}_{S}[\tilde{\theta}_{S,0}(\hat{\lambda}_{S}) - \tilde{\theta}_{S,0}(\tilde{\lambda}_{S,0})]| \stackrel{p}{\rightarrow} 0$$

Proof.

$$\mathbb{E}_{S} \, \tilde{\theta}_{S,0}(\hat{\lambda}_{S}) = \tilde{\theta}_{S}(\hat{\lambda}_{S}) \qquad \text{by definition}$$

$$\leq \tilde{\theta}_{S}(\tilde{\lambda}_{S}) + \mathbb{E}_{S}[\tilde{\varphi}(\hat{\lambda}_{S}) + \tilde{\varphi}(\tilde{\lambda}_{S})] \qquad \text{by eq. (20)}$$

$$\leq \tilde{\theta}_{S}(\tilde{\lambda}_{S,0}) + \mathbb{E}_{S}[\tilde{\varphi}(\hat{\lambda}_{S}) + \tilde{\varphi}(\tilde{\lambda}_{S})] \qquad \text{since } \tilde{\theta}_{S}(\tilde{\lambda}_{S}) \leq \tilde{\theta}_{S}(\lambda) \forall \lambda \in \Lambda$$

$$= \mathbb{E}_{S} \, \tilde{\theta}_{S,0}(\tilde{\lambda}_{S,0}) + \mathbb{E}_{S}[\tilde{\varphi}(\hat{\lambda}_{S}) + \tilde{\varphi}(\tilde{\lambda}_{S})] \qquad \text{by definition}$$

Thus,

$$0 \leq \mathbb{E}_{S}[\tilde{\theta}_{S,0}(\hat{\lambda}_{S}) - \tilde{\theta}_{S,0}(\tilde{\lambda}_{S,0})] \leq \mathbb{E}_{S}[\tilde{\varphi}(\hat{\lambda}_{S}) + \tilde{\varphi}(\tilde{\lambda}_{S})] = o_{p}\left(1/\sqrt{\min(n_{A}, n_{o})}\right)$$

B Population worst risk computation

Using lemma 1, we want to compute the worst risk at the population level. Henceforth, we generate $(X^{\tilde{A}}, Y^{\tilde{A}}) = \text{SEM}(B, \tilde{A})$ for $\tilde{A} \in \{A, 0\}$ such that $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[A] = 0$, $\mathbb{C}[A] = I_{p \times p}$ and $\mathbb{C}[\epsilon] = I_{(p+1)\times(p+1)}$. For all $\beta \in \mathbb{R}^p$, we have that

$$\sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta) = \frac{1}{2} R_{+}(\beta) + \frac{1+2\tau}{2} R_{\Delta}(\beta)$$
 by lemma 1

s.t.
$$R_{+}(\beta) = (\beta - \beta_{OLS})^{T} G_{+}(\beta - \beta_{OLS})$$
 (21)

and
$$R_{\Delta}(\beta) = (\beta - \beta_{PA})^T G_{\Delta}(\beta - \beta_{PA})$$
 (22)

where $\beta_{OLS} := G_+^{-1} Z_+$. We proceed to compute both $R_+(\beta)$ and $R_{\Delta}(\beta)$. Consider the definitions

$$G_{\Delta} = \mathbb{C}[X^{A}] - \mathbb{C}[X^{0}] \tag{23}$$

$$G_{+} = \mathbb{C}[X^{A}] + \mathbb{C}[X^{0}] \tag{24}$$

and the following expansion for $\tilde{A} \in \{A, 0\}$

$$\mathbb{C}[X^{\tilde{A}}] = u(\mathbb{C}[\epsilon] + \begin{bmatrix} 0 & \mathbb{O}^T \\ \mathbb{O} & \mathbb{C}[\tilde{A}] \end{bmatrix}) u^T \text{ where } u = [w \ M]$$
 from eq. (14)

$$= ww^T + MM^T + M\mathbb{C}[\tilde{A}]M^T$$
 since $\mathbb{C}[\epsilon] = I_{p \times p}$

$$= M \left(\beta_{\text{CH}} \beta_{\text{CH}}^T + I + \mathbb{C}[\tilde{A}]\right) M^T$$
 since $M \beta_{\text{CH}} = w$ by eq. (15)

hence by replacing $\mathbb{C}[X^A]$ and $\mathbb{C}[X^0]$ in equations (23) and (24), we get

$$G_{\Delta} = M \mathbb{C}[A]M^{T} = MM^{T}$$

$$G_{+} = M \left(2 \beta_{CH} \beta_{CH}^{T} + 3I\right)M^{T}$$
(25)

then plugging the last two equations into equations (22) and (21), we obtain

$$R_{+}(\beta) = 2 \|\beta_{CH}^{T} M^{T} (\beta - \beta_{OLS})\|_{2}^{2} + 3 \|M^{T} (\beta - \beta_{OLS})\|_{2}^{2}$$
(26)

$$R_{\Delta}(\beta) = \left\| M^{T}(\beta - \beta_{PA}) \right\|_{2}^{2} \tag{27}$$

We compute $M^T \beta_{\text{OLS}}$ to simplify the computation of $R_+(\beta)$

$$\beta_{\mathrm{OLS}} \coloneqq \mathrm{G}_{+}^{^{\text{-}1}} \, \mathrm{Z}_{+} = \mathrm{G}_{+}^{^{\text{-}1}} \left(\mathbb{C} \big[\mathrm{X}^{^{\mathrm{A}}}, \mathrm{Y}^{^{\mathrm{A}}} \big] + \mathbb{C} \big[\mathrm{X}^{^{\mathrm{0}}}, \mathrm{Y}^{^{\mathrm{0}}} \big] \right)$$

where for $\tilde{A} \in \{A, 0\}$

$$\mathbb{C}[X^{\tilde{A}}, Y^{\tilde{A}}] = \mathbb{C}[X^{\tilde{A}}, \beta_{PA}^{T} X^{\tilde{A}} + \epsilon_{Y}]
= \mathbb{C}[X^{\tilde{A}}] \beta_{PA} + \mathbb{C}[X^{\tilde{A}}, \epsilon_{Y}]
= \mathbb{C}[X^{\tilde{A}}] \beta_{PA} + u \mathbb{C}[\epsilon + \begin{bmatrix} 0 \\ \tilde{A} \end{bmatrix}, \epsilon_{Y}]
= \mathbb{C}[X^{\tilde{A}}] \beta_{PA} + w \mathbb{V}[\epsilon_{Y}]$$
 since $\mathbb{C}[\tilde{A}, \epsilon_{Y}] = 0$ and $\mathbb{C}[\epsilon_{X}, \epsilon_{Y}] = 0$
= $\mathbb{C}[X^{\tilde{A}}] \beta_{PA} + w$
= $\mathbb{C}[X^{\tilde{A}}] \beta_{PA} + M \beta_{CH}$ by eq. (15)

Thus,

$$\beta_{\rm OLS} \coloneqq {\rm G}_{+}^{\text{-}1} \, {\rm Z}_{+} = {\rm G}_{+}^{\text{-}1} \, \big({\rm G}_{+} \, \beta_{\rm PA} + 2 M \, \beta_{\rm CH} \big) = \beta_{\rm PA} + 2 \, {\rm G}_{+}^{\text{-}1} \, M \, \beta_{\rm CH}$$

and

$$M^{T} \beta_{\text{OLS}} = M^{T} \beta_{\text{PA}} + 2M^{T} G_{+}^{-1} M \beta_{\text{CH}}$$

$$= M^{T} \beta_{\text{PA}} + 2 \left(2 \beta_{\text{CH}} \beta_{\text{CH}}^{T} + 3I \right)^{-1} \beta_{\text{CH}} \qquad \text{by eq. (25)}$$

$$= M^{T} \beta_{\text{PA}} + \left(\beta_{\text{CH}} \beta_{\text{CH}}^{T} + \frac{3}{2}I \right)^{-1} \beta_{\text{CH}} \qquad (28)$$

In summary, an algorithm for computing $\sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta)$ is

Algorithm 1 Computation of population worst risk for β

$$1 M \leftarrow ((I - B_X) - \beta_{CH} \beta_{PA}^T)^{-1}$$
 > By eq. (16)

2
$$h \leftarrow M \beta_{PA}$$

3
$$m \leftarrow M^T \beta_{\text{OLS}} = h + \left(\beta_{\text{CH}} \beta_{\text{CH}}^T + \frac{3}{2}I\right)^{-1} \beta_{\text{CH}}$$
 > By eq. (28)

4
$$b \leftarrow M^T \beta$$

5
$$\delta \leftarrow M^T(\beta - \beta_{\text{OLS}}) = b - m$$

6 R₊(
$$\beta$$
) ← 2 $\|\beta_{\text{CH}}^{\text{T}} \delta\|_{2}^{2}$ + 3 $\|\delta\|_{2}^{2}$ ▷ By eq. (26)

$$7 R_{\Delta}(\beta) \leftarrow ||b - h||_2^2$$
 \triangleright By eq. (27)

8
$$\sup_{\tilde{A} \in C_{1+\tau}} R_{\tilde{A}}(\beta) \leftarrow \frac{1}{2} R_{+}(\beta) + \frac{1+2\tau}{2} R_{\Delta}(\beta)$$