Part III Essay
New advances in causal inference

Contents

Introduction

Constantly, we find ourselves trying to understand the cause and effect (or *causal*) relationships of what we observe in the world around us. This can range from questions at an individual or personal level ("Why is my stomach currently grumbling?"), to handling concerns on the scale of societies or the entirety of humanity or nature itself. Epidemiologists investigate causal relationships between health - for example, the likelihood of developing a certain type of cancer - and the factors which have an effect on this, may they be genetic, environmental or related to lifestyle. Sociologists attempt to discover the causes of order and disorder within a society. Geneticists are concerned with what genes, either by their addition, removal or allele type, give rise to certain characteristics or behaviours of a cell or living organism.

Before trying to answer these types of questions, we first need to clarify what we mean by "causality" and a "causal" relationship. *Causality* is the means by which one process (the *cause*) may be linked with another (the *effect*) through the passage of time. We say that the prior is *causal* for the latter if upon the action of the prior, through either some deterministic or probabilistic mechanism, the effect is more likely to occur than if we did not impose the cause. Therefore, *causal relationships* are simply the statements about causality we can make given a set of processes or events.

When handling questions about causality, we are concerned with what data we can use when trying to examine causal relationships, how it can be used to help answer questions relating to causality, and the form these questions take. Although the first and third questions are also of interest to the statistician, the question of "how" is the most important. Answering this also requires some work, as the classical tool of regression is no longer suitable here. The reason for this is best summed up by the paradigm "correlation does not imply causation". To explain further, the presence of correlation is a necessary condition for causality, but not sufficient. This is because generally there is some information about the causal effect of X on Y which is not identifiable from the regression function $\mathbb{E}[Y|X]$.

For these reasons, the development of new tools to help answer causal questions was necessary; nowadays we find ourselves with a variety of different methodologies at our disposal. As hinted to previously, these require incorporating some assumptions about the structure of causal relationships, and the questions we want to ask from them, in order to draw conclusions about them from the data. As there are a multiplicity of approaches, it is also worth considering the assumptions and implications of one approach over another.

We begin in Section 1 by detailing one of the more classical frameworks used to perform causal inference, that of structural equation modelling. We then focus on the method of invariant prediction as introduced by Peters, Bühlmann, and Meinshausen [2016]. This begins by discussing in Section 2 how this method can be used to perform causal inference. When there are a large number of covariates, the method proposed in Section 2 becomes computationally infeasible, and so we discuss how to pre-screen for potentially identifiable causal predictors in Section 3. We then discuss some of the links and differences between invariant prediction and structural equation modelling in Section 4. We end by applying some of these methods to simulated data in Section 5.

Section 1

Structural equation modelling

Here we give an overview of structural equation modelling, which uses directed graphs to represent the causal structure of random variables, containing information about the dependencies introduced as part of the data generating process. Formally, a *structural equation model* for a random vector $X \in \mathbb{R}^p$ is a system of equations

$$X_i = f_i\left(X_{S_i}, \epsilon_i\right) \text{ for } i = 1, \ldots, p \tag{1.1}$$

for some subsets $S_i \subseteq \{1, \ldots, p\} \setminus \{i\}$ and functions $f_i : \mathbb{R}^{|S_i|} \times \mathbb{R} \to \mathbb{R}$ such that

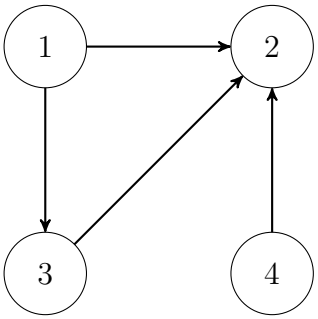
- the $\epsilon_1, \ldots, \epsilon_p$ are jointly independent (but can differ in distribution), and
- the graph $G = (V, E)$, with vertex set $V = \{1, \ldots, p\}$ and edge set E such that $\text{pa}(i) = S_i$ for $i = 1, \ldots, p$, is a directed acyclic graph.

An example is given in Figure 1.1. Appendix 1 gives a brief summary of the basic theory of directed acyclic graphs. We note that a structural equation model specifies the distribution of X, as given a topological ordering π of G, we can build X_i as a function of $\epsilon_{\pi^{-1}(1)}, \ldots, \epsilon_{\pi^{-1}(\pi(i))}$ for $i = 1, \ldots, p$.

Although a structural equation model determines the joint distribution of the X_i, it also provides more information about how the variables affect each other. Consider the following two (basic) structural equation models:

$$\begin{aligned} X_1 &= 2X_2 + \epsilon_1,\ \epsilon_1 \sim N(0,1) & \qquad Y_1 &= -2Y_2 + \epsilon_1',\ \epsilon_1' \sim N(0,1) \\ X_2 &= \epsilon_2 \sim N(0,1) & \qquad Y_2 &= \epsilon_2' \sim N(0,1) \end{aligned}$$

Although both (X_1, X_2) and (Y_1, Y_2) have the same distribution, we see that X_2 has a "positive" effect on X_1, whereas Y_2 has a "negative" effect on Y_1. If we were given only the joint distribution of X, we could not distinguish it from Y. Using a structural equation model is one method which allows us to do so.



$$
\begin{aligned}
X_1 &= 2 + \epsilon_1^2, & \epsilon_1 &\sim U[1,2] \\
X_2 &= X_1 - X_3^2 + X_4 + \epsilon_2, & \epsilon_2 &\sim N(0,1) \\
X_3 &= -X_1 \log(\epsilon_3), & \epsilon_3 &\sim U[0,1] \\
X_4 &= \epsilon_4, & \epsilon_4 &\sim N(0,2)
\end{aligned}
$$

Figure 1.1: An example of a structural equation model (under the assumption that the ϵ_i are independent) and its associated directed acyclic graph.

Furthermore, by using a structural equation model, we are assuming that there are no feedback loops between random variables in the model, as the associated graph G is a directed acyclic graph. Depending on the context, this may either a desirable or undesirable property. For example, in many biological processes, such as homoeostasis, feedback loops are present, meaning we cannot model the relationships between (say) hormonal levels, body temperature, blood glucose levels etc. using a structural equation model as defined here.

1.1 Interventions

An important feature of a structural equation model is that the defining equations allow us to see the behaviour and distribution of X after having "intervened" within the system. These can represent different experimental settings, such as modelling when we fix the concentration of a reagent in a chemical reaction, or indicating whether a patient is receiving a placebo or an actual treatment.

They can also represent different observational settings, where the distribution of a subset of the X_i may change when examining different populations. For example, suppose we are looking to see whether smoking tobacco is causal for the development of lung cancer. If it were, we would expect to see this relationship occur regardless of the proportion of smokers in a certain population[1]. By treating the observational setting of different populations as interventional settings - even though we could never create a randomized experiment to carry this out - we could potentially use

[1]Provided we did not intervene and give a proportion of the population lung cancer. This would be uninformative, infeasible and unlikely to be approved by any ethics board anywhere.

this to infer causal information[2].

With reference to the structural equation model in (1.1), which we refer to as $\mathcal{S}$, we now formally define what an intervention is (see e.g Peters [2015]). Letting $\mathcal{A} \subseteq \{1, \ldots, p\}$ be non-empty, we say that an *intervention* on the variables $X_{\mathcal{A}}$ in $\mathcal{S}$ is a structural equation model $\widetilde{\mathcal{S}}$ where

$$X_j = \begin{cases} f_j(X_{\mathrm{pa}(j)}, \epsilon_j) & \text{for } j \notin \mathcal{A} \ (f, \mathrm{pa}(j) \text{ and } \epsilon_j \text{ as according to } \mathcal{S}) \\ \widetilde{f}_j(X_{\widetilde{\mathrm{pa}}(j)}, \widetilde{\epsilon}_j) & \text{for } j \in \mathcal{A} \ (\text{for some } \widetilde{f}_j, \widetilde{\mathrm{pa}}(j) \text{ and } \widetilde{\epsilon}_j) \end{cases} \tag{1.2}$$

and we require that $\{\epsilon_j \mid j \notin S\}$ is independent of $\{\widetilde{\epsilon}_j \mid j \in S\}$. We frequently denote the interventional system by

$$X \mid \mathrm{do}\left(X_j = \widetilde{f}_j(X_{\widetilde{\mathrm{pa}}(j)}, \widetilde{\epsilon}_j) \text{ for } j \in S\right). \tag{1.3}$$

This is not to be confused with a conditional distribution, which is denoted by $X \mid Z$ (for example). Some special cases of interventions include:

- a *perfect* or *do-intervention* [Pearl, 2009] when $S = \{j\}$ and $X_j = a$ under $\widetilde{\mathcal{S}}$;
- a *structural intervention* [Eberhardt and Scheines, 2007] when $S = \{j\}$ and $\mathrm{pa}(j) = \widetilde{\mathrm{pa}}(j)$.

These specializations can be easily generalized to handle interventions on multiple random variables.

We can now determine what variables give rise to causal effects in a structural equation model by reference to interventions. We may expect that X has a causal effect on Y if Y is not independent of X for any distribution of X, where the freedom in definition of X allows us to identify that X causes Y rather than vice versa. However, this intuition is subtlety incorrect, as the following example illustrates. If we have a structural equation model

$$Y = X\mathbb{1}[X > 1] + \epsilon_1, \quad X = \epsilon_2, \quad \epsilon \sim N(0, I_2),$$

then we would agree that X has a causal effect on the value of Y. We also see

[2]Of course, how to do so is another question entirely. To begin, this requires being able to agree on what a "population" actually is, and to argue that the membership to a particular population does not have a causal effect in itself. Afterwards, there is then the matter of converting observational information to interventional information. Although these are interesting questions to consider, we do not do so any further.

that X and Y are not independent; however, this is not always preserved under interventions. For example, if we perform $\text{do}(X = \widetilde{\epsilon_1})$ where $\epsilon_1 \sim U[0,1]$, then under the interventional distribution Y and X are independent.

The correct intuition is instead that there exists a interventional distribution for X such that, in the intervened model, Y and X are not independent. Formally, we say that there is a *total causal effect* from X_j to X_k in a structural equation model if there exists $\widetilde{\epsilon}$ such that, under $X \,|\, \text{do}(X_j = \widetilde{\epsilon})$, X_j and X_k are not independent.

1.2 Conditional independence of random variables

The definition of a total causal effect inspires the question of how we can generally infer independence statements easily from a structural equation model. To begin, we say that P satisfies the *global Markov property* with respect to a directed acyclic graph G if for all $A, B, S \subseteq \{1, \ldots, p\}$ which are pairwise disjoint, whenever A and B are d-separated by S we also have that $X_A \perp\!\!\!\perp X_B | X_S$.

Supposing that P is absolutely continuous with respect to a product measure (say with density f), it is equivalent to the *Markov factorization property* [Lauritzen, 1996, Theorem 3.27], which says that for all $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$,

$$f(x_1, \ldots, x_p) = \prod_{i=1}^{p} f\left(x_i | x_{\text{pa}(i)}\right). \tag{1.4}$$

The following result shows that the latter holds for the distribution of a structural equation model.

Theorem 1.1. [Pearl, 2009, Theorem 1.4.1] *Let P be the law of a structural equation model with associated directed acyclic graph G, and suppose that P is absolutely continuous with respect to some measure with density f. Then P satisfies the Markov factorization property with respect to G.*

Proof. Let π be a topological ordering on G and $\tau = \pi^{-1}$. Then by conditioning on the variables in the order of the topological ordering, we get

$$f(x_1, \ldots, x_p) = f(x_{\tau(j+1)}, \ldots, x_{\tau(p)} \,|\, x_{\tau(1)}, \ldots, x_{\tau(j)}) \prod_{i=1}^{j} f(x_{\tau(i)} \,|\, x_{\tau(1)}, \ldots, x_{\tau(i-1)})$$

$$= \ldots = \prod_{i=1}^{p} f(x_{\tau(i)} \,|\, x_{\tau(1)}, \ldots, x_{\tau(i-1)}).$$

Out of variables with indices $\{\tau(1), \ldots, \tau(i-1)\}$, $X_{\tau(i)}$ depends only on those with indices in $\mathrm{pa}(\tau(i))$, which is completely contained within this set by definition of π. Therefore

$$f(x_1, \ldots, x_p) = \prod_{i=1}^{p} f(x_{\tau(i)} \,|\, x_{\mathrm{pa}(\tau(i))}) = \prod_{i=1}^{p} f(x_i \,|\, x_{\mathrm{pa}(i)}). \qquad \square$$

Therefore, under some mild assumptions, we know the directed acyclic graph associated with a structural equation model encodes positive information about conditional independence, and thus regular independence (when sets are d-separated by the empty set). The graph structure can also provide information about total causal effects, as according to the following proposition.

Proposition 1.2. [Peters, 2015] *Let X be generated by a structural equation model $\mathcal{S}$ with associated directed acyclic graph G. Then for $j, k \in \{1, \ldots, p\}$ with $j \neq k$, if there is not a directed path from j to k in G, then there is no total causal effect from X_j to X_k.*

Proof. Let $\widetilde{\epsilon}_j$ be arbitrary. Suppose the structural equation model $\widetilde{\mathcal{S}}$ produced after the intervention $\mathrm{do}(X_j = \widetilde{\epsilon}_j)$ on $\mathcal{S}$ has associated directed acyclic graph $\widetilde{G}$. Then $\widetilde{G}$ is simply G with the set of vertices $\{(i, j) \,|\, i \in \mathrm{pa}(j)\}$ removed. In particular, this means that j and k are d-separated by the empty set in $\widetilde{G}$ if and only if for any path from j to k, there exists a colliding node. As there is no directed path from j to k in G, the same holds in $\widetilde{G}$, and thus the path in $\widetilde{G}$ contains a collider. $\square$

Unfortunately, the converse is not true. For example, consider the following structural equation model:

$$\begin{aligned} X_1 &= X_2 - X_3 + \epsilon_1 \\ X_2 &= X_3 + \epsilon_2 \\ X_3 &= \epsilon_3. \end{aligned} \tag{1.5}$$

Even though $3 \to 1$, as we can write $X_1 = \epsilon_1 + \epsilon_2$, it follows that X_1 and X_3 are independent under $X \,|\, \mathrm{do}(X_3 = \widetilde{\epsilon}_3)$ for any distribution $\widetilde{\epsilon}_3$. This is interesting, as it means that the absence of a directed path means the absence of a causal effect, whereas the existence of a path says only that there *might* be an effect.

1.3 Causal identifiability

The inability to necessarily make positive total causal effect statements between variables in a structural equation model is an example of (what we call) *causal identifiability* issues. To give another example where problems may arise, suppose P is generated by an (unknown) structural equation model, and we want to estimate the underlying directed acyclic graph G from the data. However, P can potentially satisfy the Markov factorization property with respect to a large number of graphs, and so to perform meaningful inference we need to reduce this number.

We now assume that P is absolutely continuous with respect to some (product) measure, and refer to the two Markov properties as the same. We say that P satisfies *causal minimality* with respect to G if it is Markov with respect to $G = (V, E)$, but not to any proper subgraph $G' \subset G$ with the same vertex set V. This is reasonable to assume in practice as it is necessary for model identifiability; without causal minimality, we could not distinguish between models with $Y = 0 \cdot X + \epsilon_1$ and $Y = \epsilon_1$, for example. A causally minimal graph may also provide more information on (the absence of) total causal effects as compared to a supergraph also satisfying the Markov property.

However, this is not a strong enough assumption for identifiability purposes. Consider the structure equation models for X and $\widetilde{X}$ respectively, where $\epsilon \sim N(0, I_3)$ and $\widetilde{\epsilon} \sim N(0, \Lambda)$ where $\Lambda = \mathrm{diag}(2, 1/2, 1)$:

$$\begin{aligned} X_1 &= X_2 - X_3 + \epsilon_1 & \widetilde{X}_1 &= \widetilde{\epsilon}_1 \\ X_2 &= X_3 + \epsilon_2 & \widetilde{X}_2 &= \tfrac{1}{2}\widetilde{X}_1 + \widetilde{X}_3 + \widetilde{\epsilon}_2 \\ X_3 &= \epsilon_3 & \widetilde{X}_3 &= \widetilde{\epsilon}_3 \end{aligned}$$

Then both X and $\widetilde{X}$ are distributed as $N(0, \Sigma) = P$, where

$$\Sigma = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

and P satisfies causal minimality with respect to the associated directed acyclic graphs of both X and $\widetilde{X}$. Despite this, the latter model appears simpler; it does not have causal effects cancel out as X_2 and X_3 do in the former. We therefore may believe that the latter is more plausible or preferable for purposes of interpretation.

To solve these types of issues, we say that P is *faithful* with respect to a directed acyclic graph G if for all $A, B, S \subseteq \{1, \dots, p\}$ which are pairwise disjoint, A and B d-separated by S if and only if $X_A \perp\!\!\!\perp X_B \,|\, X_S$. Going back to the prior example, P is faithful with respect to the directed acyclic graph associated with $\widetilde{X}$, but not X. In contrast to causal minimality, this assumption is quite strong and harder to justify in practice. For linear structural equation models, this is frequently "justified" by saying that (informally) the "set of unfaithful distributions has Lebesgue measure zero" [Spirtes et al., 2001, Theorem 3.2][3]. Therefore if the law of a linear structural equation model is absolutely continuous with respect to Lebesgue measure, we (informally) have "faithfulness with probability one".

From a methodological viewpoint, we may instead interpret faithfulness as a version of Occam's razor, and so either justified or refuted as such. In our scenario, we want to select the most "simple" graph structure for which the Markov property holds. We have already argued why assuming causal minimality is a good starting criterion for "simplicity"; this is also a necessary condition for faithfulness. Indeed, if we remove an edge $j \to k$, this introduces a new positive conditional independence statement (that $X_j \perp\!\!\!\perp X_k \,|\, X_{\mathrm{pa}(k)}$, assuming that $\pi(j) < \pi(k)$ for a topological order π), which cannot happen as by faithfulness we already know them all. As faithful models are less likely to have the effects of variables cancel each other out (as this is controlled by d-separation), they are therefore a good candidate for "simple" models.

[3]This result is sometimes stated in passing without (or at least hiding) the linear requirement, but we are not able to find such a general result.

Section 2

An introduction to invariant prediction

We now discuss the method of *invariant prediction* presented by Peters, Bühlmann, and Meinshausen [2016] in detail, the main focus from herein. The method exploits the idea that if we condition on the variable of interest by all of its "direct causes", the resulting distribution will be invariant under different experimental settings which do not interfere with the variable of interest, and that this is not necessarily true if we ignore some of the direct causes. By testing for this, we can therefore attempt to infer some of the true causal variables, along with point estimates and confidence regions for the "effect size" of each.

Here, we develop a methodology to handle generalized linear models (in some sense), generalizing those developed in Peters et al. [2016] for linear models. We begin in Section 2.1 by introducing the invariant prediction assumption, establishing which causal variables we can then identify in Section 2.2. In Section 2.3, we formulate a hypothesis test which can tell us whether the baseline and direct effect sizes of our causal variables are invariant under different experimental conditions. We then use this in Section 2.4 to give a generic testing procedure, guaranteeing coverage of the "true causal predictors" up to a desired size. Finally in Section 2.5, we detail how these procedures can be implemented.

2.1 The invariant prediction assumption

Suppose we have different experimental settings denoted by $e \in \mathcal{E}$, where for each environment e we have (Y^e, X^e), with $X^e \in \mathbb{R}^p$ a (row) vector of predictor variables and $Y^e \in \mathbb{R}$ the response. We use the convention that if $S \subseteq \{1, \ldots, p\}$, then X_S is the (row) vector with entries X_j for $j \in S$.

We then say that the *invariant prediction assumption* is satisfied if there exists a subset $S^* \subseteq \{1, \ldots, p\}$ such that, for all $e \in \mathcal{E}$, X^e has an arbitrary distribution,

$$Y^e = h\left(X^e_{S^*}, \epsilon^e\right) \text{ where } \epsilon^e \sim F_\epsilon \text{ and } \epsilon^e \perp\!\!\!\perp X^e_{S^*}, \tag{2.1}$$

and both $h : \mathbb{R}^{|S^*|} \times \mathbb{R} \to \mathbb{R}$ and the error distribution F_ϵ do not depend on the experimental setting e. This framework is equivalent to the intuition that distributions conditional on causal effects should be invariant, in the following sense.

Proposition 2.1. [Peters et al., 2016, Section 6.1] *For a given subset $S \subseteq \{1, \ldots, p\}$, the following are equivalent:*

(i) There exists a function $h : \mathbb{R}^{|S|} \times \mathbb{R} \to \mathbb{R}$ and noise distribution for ϵ^e such that the invariant prediction assumption is satisfied.

(ii) For all $e, f \in \mathcal{E}$, $Y^e | X^e_S = x$ is equal in distribution to $Y^f | X^f_S = x$, for all $x \in \mathbb{R}^{|S|}$ such that both conditional distributions are well-defined.

Proof. The forward direction is immediate. For the reverse, let $\epsilon^e \sim U[0, 1]$ be independent of X^e_S, and $h(a, b) := F^-_{Y^e | X^e_S = a}(b)$ be the quantile function corresponding to F, the c.d.f of $Y^e | X^e_S = a$. This has no dependence on e by assumption. □

We now specialize to where the conditional distributions belong to an exponential dispersion family; see Appendix 2 for a review. Fix a particular exponential dispersion family $P_{\theta,\sigma}$, and write $Z \sim \text{ED}(\mu, \sigma)$ whenever $Z \sim P_{\theta(\mu),\sigma}$. We then say that the *invariant prediction assumption (for generalized linear models)* is satisfied if there exists a link function g, a (column) vector of coefficients $\gamma^* = \left(\gamma^*_1, \ldots, \gamma^*_p\right)^T \in \mathbb{R}^p$ with support $S^* := \text{supp}(\gamma) = \{k : \gamma^*_k \neq 0\} \subseteq \{1, \ldots, p\}$ and $\eta^* \in \mathbb{R}$ such that

(i) the X^e have an arbitrary distribution, and

(ii) for all $e \in \mathcal{E}$ and $x \in \mathbb{R}^{|S^*|}$ a row vector, $Y^e | X^e_{S^*} = x \sim ED(\mu_x, \sigma)$ whenever this is well-defined, where $g\left(\mu_x\right) = \eta^* + x\gamma^*$.

Here we use the abuse of notation $x\gamma^* = \sum_{i \in S^*} x_i \gamma^*_i$, as this is not a bona fide dot product; generally, if we say $x \in \mathbb{R}^{|S|}$ is a row vector and $\beta \in \mathbb{R}^p$ is a column vector, we will write $x\beta = \sum_{i \in S} x_i \beta_i$, i.e as if x were embedded in $\mathbb{R}^p$ with $x_{-S} \equiv 0$. Although these are not strictly generalized linear models in the sense of regression, we will use similar language (e.g canonical link functions) as the ideas are similar.

As a special case, the linear (Gaussian) model as examined by Peters et al. [2016]

lies within this framework. Indeed, if

$$Y^e = \eta^* + X^e\gamma^* + \epsilon^e \text{ where } \epsilon^e \sim N(0, \sigma^2) \text{ and } \epsilon^e \perp\!\!\!\perp X_{S^*}, \tag{2.2}$$

then when using the canonical link for the Normal distribution ($g(x) = x$) we have that $Y^e|X^e_{S^*} = x \sim N(\eta^* + x\gamma^*, \sigma^2)$. However, as now we can use the tools of exponential dispersion families and generalized linear models, we could also consider e.g logistic models where

$$Y^e|X_{S^*} = x \sim \text{Bernoulli}\left(\frac{e^{\eta^*+x\gamma^*}}{1+e^{\eta^*+x\gamma^*}}\right). \tag{2.3}$$

Given this framework, we now want to estimate (η^*, γ^*, S^*) and give confidence regions for the former two quantities. From herein, we may refer to the two definitions interchangeably by the *invariant prediction assumption*; the context will be sufficient to distinguish between them.

2.2 Plausible and identifiable causal predictors

The first obstacle in trying to estimate (η^*, γ^*, S^*) is that in general, we a priori have no reason to expect that there is a unique pair (η^*, γ^*, S^*) which allow the invariant prediction assumption to be satisfied. Therefore we define, for $\gamma \in \mathbb{R}^p$, $\eta \in \mathbb{R}$ and $S \subseteq \{1, \ldots, p\}$, the null hypothesis

$$H_{0,\gamma,\eta,S}(\mathcal{E}) : \begin{cases} \text{there exists } \sigma \in (0,\infty) \text{ and a link function } g \text{ s.t} \\ \text{for all } e \in \mathcal{E} \text{ and } x \in \mathbb{R}^{|S|} \text{ when this is well defined,} \\ Y^e|X^e_S = x \ \sim \text{ED}(\mu_x, \sigma) \text{ where } g(\mu_x) = \eta + x\gamma. \end{cases} \tag{2.4}$$

We then call the variables $S \subseteq \{1, \ldots, p\}$ *plausible causal predictors under* $\mathcal{E}$ if the null hypothesis

$$H_{0,S}(\mathcal{E}) : H_{0,\gamma,\eta,S} \text{ is true for some } \gamma \in \mathbb{R}^p \text{ and } \eta \in \mathbb{R} \tag{2.5}$$

is true. The *identifiable causal predictors under* $\mathcal{E}$ are then defined to be the following subset of the plausible causal predictors:

$$S(\mathcal{E}) := \bigcap_{S:H_{0,S}(\mathcal{E}) \text{ is true}} S. \tag{2.6}$$

If the intersection is empty, we let $S(\mathcal{E}) = \emptyset$. We may think of $S(\mathcal{E})$ as the maximal subset of $\{1, \ldots, p\}$ which is contained by all the plausible predictors; as we may not know the "true" predictors, these are the only variables we can hope to identify.

We now make a few remarks about the above definitions. Firstly, if the invariant prediction assumption is true, then by construction $S(\mathcal{E}) \subseteq S^*$, which will be useful later in guaranteeing coverage statements. Secondly, if $\mathcal{E}_1 \subseteq \mathcal{E}_2$ and the intersection over $\{S \mid H_{0,S}(\mathcal{E}_2) \text{ is true}\}$ is non-empty, then $S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2)$. This occurs simply as if for some S we know $H_{0,S}(\mathcal{E}_2)$ is true, then so is $H_{0,S}(\mathcal{E}_1)$. One way of interpreting this is to say that looking at more interventional environments allows us to gain more information about which variables are directly causal. However, this says nothing about whether there exists a finite set of environments $\mathcal{E}$ such that $S(\mathcal{E}) = S^*$; see Section 4.3 for a discussion of this issue when handling structural equation models.

Thirdly, it is important to remember that in $H_{0,S}(\mathcal{E})$ we specify the conditional distribution of $Y^e | X_S^e = x$, unlike in the general case where the corresponding null hypothesis is

$$H_{0,S}^{\text{gen}}(\mathcal{E}) : \begin{cases} \text{for all } e, f \in \mathcal{E} \text{ and } x \in \mathbb{R}^{|S|} \text{ such that these are well} \\ \text{defined, } Y^e | X_S^e = x \text{ equals } Y^f | X_S^f = x \text{ in distribution.} \end{cases} \tag{2.7}$$

In a similar fashion, we may define $S^{\text{gen}}(\mathcal{E})$ as in (2.6). As whenever $H_{0,S}(\mathcal{E})$ is true, so is $H_{0,S}^{\text{gen}}(\mathcal{E})$, we therefore have that $S^{\text{gen}}(\mathcal{E}) \subseteq S(\mathcal{E})$ if some $H_{0,S}$ is true. Now suppose we have S^* as according to the *general* invariant prediction such that $H_{0,S^*}(\mathcal{E})$ is *false*. If $H_{0,S}^{\text{gen}}(\mathcal{E})$ is false for all $S \neq S^*$, then $S(\mathcal{E}) = \emptyset \subseteq S^*$. However, if S^* is not unique in the sense that $H_{0,S}(\mathcal{E})$ (and so $H_{0,S}^{\text{gen}}(\mathcal{E})$) is true for some $S \neq S^*$, we may have that $S^{\text{gen}}(\mathcal{E}) \subseteq S^* \subseteq S(\mathcal{E})$. Therefore if our distributional assumptions are false, we cannot always guarantee that $S(\mathcal{E})$ contains only variables in S^*.

Finally, we consider the interpretations of our results if $|\mathcal{E}| = 1$. Working in the general case, $H_{0,S}^{\text{gen}}(\mathcal{E})$ is (vacuously) true for any $S \subseteq \{1, \ldots, p\}$ and so $S^{\text{gen}}(\mathcal{S}) = \emptyset$. Peters et al. [2016] suggest that this should be interpreted as a conservative principle which makes no claim as to which variables are causal. However, as $S^{\text{gen}}(\mathcal{E}) = \emptyset$ can occur when $H_{0,S}^{\text{gen}}(\mathcal{E})$ is either true or false for all $S \subseteq \{1, \ldots, p\}$, we should not try to interpret $S^{\text{gen}}(\mathcal{E})$ in this way and rather consider $\{S \mid H_{0,S}^{\text{gen}}(\mathcal{E}) \text{ is true}\}$ instead. Similarly, we should only be concerned when $\{S \mid H_{0,S}(\mathcal{E}) \text{ is true}\} = \emptyset$, as then the invariant prediction assumption (for generalized linear models) is false for all $S \subseteq \{1, \ldots, p\}$, rendering any further analysis moot.

We now return to considering the $\eta \in \mathbb{R}$ and $\gamma \in \mathbb{R}^p$ which allow for $H_{0,S}(\mathcal{E})$ to be true (if indeed it is), which we can interpret as the *baseline* and *direct effect* sizes we want to infer. We define, for $S \subseteq \{1, \ldots, p\}$,

$$\Gamma_S(\mathcal{E}) := \{(\gamma, \eta) \in \mathbb{R}^p \times \mathbb{R} \,|\, H_{0,\gamma,\eta,S} \text{ is true}\} \tag{2.8}$$

the set of *plausible causal coefficients for S under $\mathcal{E}$*, and consequently define the *global set of plausible causal coefficients under $\mathcal{E}$* by

$$\Gamma(\mathcal{E}) := \bigcup_{S \subseteq \{1,\ldots,p\}} \Gamma_S(\mathcal{E}). \tag{2.9}$$

Clearly, if $H_{0,S}(\mathcal{E})$ is false, then $\Gamma_S(\mathcal{E}) = \emptyset$. Similar to the inclusion rule for $S(\mathcal{E})$, and with the same interpretation, we also know that if $\mathcal{E}_1 \subseteq \mathcal{E}_2$, then $\Gamma(\mathcal{E}_1) \supseteq \Gamma(\mathcal{E}_2)$.

2.3 Simplifying the null hypothesis $H_{0,S}(\mathcal{E})$

We now reformulate the null hypothesis $H_{0,S}(\mathcal{E})$ as given in (2.5) to make it more amenable to testing. We first introduce the shorthand $\theta^e(\beta, \zeta) = \theta(\mu^e(\beta, \zeta)) = \theta\left(g^{-1}(\zeta + X^e\beta)\right)$. Then, for each $e \in \mathcal{E}$ and $S \subseteq \{1, \ldots, p\}$, if $f(y; \theta, \sigma)$ is the density of the fixed exponential dispersion family we are considering (see (A.4)), we define the *population regression coefficients*

$$\left(\beta^{\text{pred},e}(S), \zeta^{\text{pred},e}(S)\right) := \underset{\beta \in \mathbb{R}^p \,:\, \text{supp}(\beta) \subseteq S;\ \zeta \in \mathbb{R}}{\text{argmin}} \mathbb{E}\left[-\log f\left(Y^e; \theta^e(\beta, \zeta), \sigma^e\right)\right] \tag{2.10}$$

$$= \underset{\beta \in \mathbb{R}^p \,:\, \text{supp}(\beta) \subseteq S;\ \zeta \in \mathbb{R}}{\text{argmin}} \mathbb{E}\left[K\left(\theta^e(\beta, \zeta)\right) - Y^e\theta^e(\beta, \zeta)\right] \tag{2.11}$$

whenever such a minimizing pair exists. If such a pair is not unique, this is a set of minimizing pairs, otherwise it is an ordered pair. For example, if we specialise to Gaussian models, then (2.10) simplifies to give (c.f Peters et al. [2016, Equation 9])

$$\underset{\beta \in \mathbb{R}^p \,:\, \text{supp}(\beta) \subseteq S;\ \zeta \in \mathbb{R}}{\text{argmin}} \mathbb{E}\left[\left(Y^e - (\zeta + X^e\beta)\right)^2\right]. \tag{2.12}$$

If alternatively we are considering a logistic model (with canonical link) for the Y^e, we instead want to find

$$\underset{\beta \in \mathbb{R}^p \,:\, \text{supp}(\beta) \subseteq S;\ \zeta \in \mathbb{R}}{\text{argmin}} \mathbb{E}\left[\log\left(1 + e^{\zeta + X^e\beta}\right) - Y^e(\zeta + X^e\beta)\right]. \tag{2.13}$$

Under the invariant prediction assumption and provided $S^* = S$, we know that $(\gamma^*, \eta^*) \in (\beta^{\text{pred},e}(S), \zeta^{\text{pred},e}(S))$. Let $h(\beta, \zeta)$ denote the objective function in (2.10), ν the dominating measure for Y^e and P the distribution measure for X_S^e. Suppose that $\beta \in \mathbb{R}^p$ has $\text{supp}(\beta) \subseteq S$, so $\theta^e(\beta, \zeta)$ has no dependence on X^e_{-S}. We therefore define $\theta^e_x(\beta, \zeta) := \theta(g^{-1}(\zeta + x\beta))$ for $x \in \mathbb{R}^{|S|}$, recalling the notation $x\beta = \sum_{i \in S} x_i \beta_i$. The subscript x denotes that this is the value of $\theta^e(\beta, \zeta)$ when $X_S^e = x$. Then

$$\begin{aligned}
h(\beta, \zeta) - h(\gamma^*, \eta^*) &= \mathbb{E}\left[\log\left(\frac{f\left(Y^e; \theta^e(\beta, \zeta), \sigma^e\right)}{f\left(Y^e; \theta^e(\gamma^*, \eta^*), \sigma^e\right)}\right)\right] \\
&\geq \log\left(\mathbb{E}\left[\frac{f\left(Y^e; \theta^e(\beta, \zeta), \sigma^e\right)}{f\left(Y^e; \theta^e(\gamma^*, \eta^*), \sigma^e\right)}\right]\right) \text{ (by Jensen's inequality)} \\
&= \log\left(\int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{f\left(y; \theta^e_x(\beta, \zeta), \sigma^e\right)}{f\left(y; \theta^e_x(\gamma^*, \eta^*), \sigma^e\right)} f\left(y; \theta^e_x(\gamma^*, \eta^*), \sigma^e\right) \nu(dy) P(dx)\right) \\
&= \log\left(\int_{\mathcal{X}} \int_{\mathcal{Y}} f\left(y; \theta^e_x(\beta, \zeta), \sigma^e\right) \nu(dy) P(dx)\right) \\
&= \log\left(\int_{\mathcal{X}} P(dx)\right) = \log(1) = 0 \implies h(\beta, \zeta) \geq h(\gamma^*, \eta^*).
\end{aligned}$$

Therefore (γ^*, η^*) is a minimizing pair; if X^e is non-degenerate it is also unique, as the inequality is strict unless $\beta = \gamma^*$ and $\zeta = \eta^*$.

From now onwards, we suppose that X^e is non-degenerate; we briefly discuss the degenerate case in Appendix 5. Provided (2.10) has a unique minimizer, we then define the *population residual dispersion parameters* by

$$\sigma^e(S) := \frac{\mathbb{E}\left[(Y^e - \mu^{\text{pred},e}(S))^2\right]}{V\left(\mu^{\text{pred},e}(S)\right)} \tag{2.14}$$

where $\mu^{\text{pred},e}(S) := g^{-1}\left(\zeta^{\text{pred},e}(S) + X^e \beta^{\text{pred},e}(S)\right)$. This is motivated by recalling the formula (A.5) for the variance of a random variable belonging to an exponential dispersion family. If the invariant prediction assumption is true for some S^* with dispersion parameter σ^*, we then have that $\sigma^e(S^*) \equiv \sigma^*$ for all $e \in \mathcal{E}$.

By mimicking the above arguments, we can reformulate (2.5) as

$$H_{0,S}(\mathcal{E}) : \begin{cases} \text{there exists } (\beta, \zeta) \in \mathbb{R}^p \times \mathbb{R} \text{ and a link function } g \text{ s.t for all} \\ e \in \mathcal{E} \text{ and } x \in \mathbb{R}^{|S|}, Y^e | X_s^e = x \sim \text{ED}(\mu_x, \sigma) \text{ when this exists} \\ \text{with } g(\mu_x) = \zeta + x\beta,\ \beta^{\text{pred},e}(S) \equiv \beta \text{ and } \zeta^{\text{pred},e}(S) \equiv \zeta \end{cases} \tag{2.15}$$

if the dispersion parameter σ of the family is fixed or known, as for example in

logistic models where $\sigma \equiv 1$, or more generally as

$$H_{0,S}(\mathcal{E}): \begin{cases} \text{there exists } (\beta, \zeta, \sigma) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}_+ \text{ and a link function } g \text{ s.t} \\ \text{for all } e \in \mathcal{E} \text{ and } x \in \mathbb{R}^{|S|},\ Y^e | X_S^e = x \sim \mathrm{ED}(\mu_x, \sigma) \text{ when this exists,} \\ g(\mu_x) = \zeta + x\beta, \text{ and } (\beta^{\mathrm{pred},e}(S), \zeta^{\mathrm{pred},e}(S), \sigma^{\mathrm{pred},e}(S)) \equiv (\beta, \zeta, \sigma). \end{cases} \tag{2.16}$$

In either case, we can test these directly (see Section 2.5). Furthermore, it implies that

$$\Gamma_S(\mathcal{E}) = \begin{cases} \emptyset & \text{if } H_{0,S}(\mathcal{E}) \text{ is false} \\ \left(\beta^{\mathrm{pred},e}(S), \zeta^{\mathrm{pred},e}(S)\right) & \text{otherwise,} \end{cases} \tag{2.17}$$

which can then be used to help compute confidence intervals for the linear predictor coefficients of the causal random variables.

2.4 Estimating the identifiable causal predictors

We would now like to both infer $S(\mathcal{E})$ by observing (Y^e, X^e) for different environments $e \in \mathcal{E}$, and also determine confidence intervals for the baseline and direct effect sizes of these variables. Suppose that, for some desired α, we can test $H_{0,S}(\mathcal{E})$ to a (conservative) size α for all $S \subseteq \{1, \ldots, p\}$. A generic method for giving an estimate $\widehat{S}(\mathcal{E})$ of $S(\mathcal{E})$ and a confidence set $\widehat{\Gamma}(\mathcal{E})$ of $\Gamma(\mathcal{E})$ is then as follows:

(i) For each $S \subseteq \{1, \ldots, p\}$, we test whether $H_{0,S}(\mathcal{E})$ holds to a level α.

(ii) We then estimate (the indices of) the identifiable causal random variables by

$$\widehat{S}(\mathcal{E}) := \bigcap_{S \subseteq \{1,\ldots,p\} : H_{0,S}(\mathcal{E}) \text{ not rejected}} S \tag{2.18}$$

(iii) A confidence set for (γ^*, η^*) is then obtained by forming

$$\widehat{\Gamma}(\mathcal{E}) := \bigcup_{S \subseteq \{1,\ldots,p\}} \widehat{\Gamma}_S(\mathcal{E}), \tag{2.19}$$

where

$$\widehat{\Gamma}_S(\mathcal{E}) := \begin{cases} \emptyset & \text{if } H_{0,S}(\mathcal{E}) \text{ is rejected at a level } \alpha \\ \widehat{C}(S) & \text{otherwise,} \end{cases} \tag{2.20}$$

given some level $(1-\alpha)$-confidence set $\widehat{C}(S)$ for $\left(\beta^{\mathrm{pred}}(S), \zeta^{\mathrm{pred},e}(S)\right)$.

We now state and prove a result guaranteeing coverage of the true causal predictors S^* and coefficients (γ^*, η^*) when using the above method.

Theorem 2.2. [Peters et al., 2016, Theorem 1] *Suppose we have a valid test for $H_{0,S}(\mathcal{E})$ at level α for all sets $S \subseteq \{1, \ldots, p\}$, in the sense that for all $S \subseteq \{1, \ldots, p\}$,*

$$\sup_{\mathbb{P}\,:\, H_{0,S}(\mathcal{E}) \text{ is true}} \mathbb{P}\left(H_{0,S}(\mathcal{E}) \text{ is rejected}\right) \leq \alpha.$$

Further suppose that $\widehat{S}(\mathcal{E})$ and $\widehat{\Gamma}(\mathcal{E})$ are constructed according to (2.18) and (2.19) respectively. Let $\mathbb{P}$ be a distribution over (Y, X) and consider any (γ^*, η^*, S^*) such that the invariant prediction assumption holds. Then $\widehat{S}(\mathcal{E})$ satisfies

$$\mathbb{P}\left(\widehat{S}(\mathcal{E}) \subseteq S^*\right) \geq 1 - \alpha.$$

Moreover, if $\mathbb{P}\left((\gamma, \eta) \in \widehat{C}(S)\right) \geq 1 - \alpha$ for all (γ, η, S) such that the invariant prediction assumption is satisfied, then

$$\mathbb{P}\left((\gamma^*, \eta^*) \in \widehat{\Gamma}(\mathcal{E})\right) \geq 1 - 2\alpha.$$

Proof. As when $H_{0,S^*}(\mathcal{E})$ is not rejected, we know that $\widehat{S}(\mathcal{E}) \subseteq S^*$. Therefore the first coverage statement follows as

$$\mathbb{P}\left(\widehat{S}(\mathcal{E}) \subseteq S^*\right) \geq \mathbb{P}\left(H_{0,S^*}(\mathcal{E}) \text{ is rejected}\right) \geq 1 - \alpha.$$

The second coverage statement is then a consequence of Boole's inequality:

$$\begin{aligned}\mathbb{P}\left((\gamma^*, \eta^*) \notin \widehat{\Gamma}(\mathcal{E})\right) &\leq \mathbb{P}\left(H_{0,S^*}(\mathcal{E}) \text{ is rejected, or } (\gamma^*, \eta^*) \notin \widehat{C}(S^*)\right) \\ &\leq \alpha + \alpha = 2\alpha.\end{aligned}$$

Note that the extra loss of α for probability of coverage by the confidence set is necessary; it may be the case that $H_{0,S^*}(\mathcal{E})$ is not rejected, yet $(\gamma^*, \eta^*) \notin \widehat{C}(S^*)$. □

2.5 Constructing conservative tests for $H_{0,S}(\mathcal{E})$

Given the generic testing method developed in Section 2.4, all that remains is to construct hypothesis tests to be used in practice. We now assume that

- the data consists of n independent observations (y_i, x_i) where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ is a row vector,
- for each $e \in \mathcal{E}$, we make n_e (where $n_e > p+1$) i.i.d observations from (Y^e, X^e), so in particular $\sum_{e\in\mathcal{E}} n_e = n$, and
- the $n_e \times (p+1)$ design matrix $\mathbf{X}_e$ of the n_e observations from the experimental setting $e \in \mathcal{E}$, with rows $(1, x_i)$, has full rank p.

The tests we propose to handle generalized linear models will not have exact (conservative) size α, but will do so asymptotically as the sample size goes to infinity, and so the statements of Theorem 2.2 will also hold "asymptotically".

More formally, suppose that we perform experiments e in a finite number of environments $\mathcal{E}$, with each experiment corresponding to a probability space $(\Omega_e, \mathcal{F}_e, \mathbb{P}_e)$. For our asymptotic analyses, we then work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ corresponding to an infinite number of observations from each environment, where

$$\Omega := \prod_{e\in\mathcal{E}} \Omega_e^{\mathbb{N}}, \quad \mathcal{F} := \bigotimes_{e\in\mathcal{E}} \mathcal{F}_e^{\mathbb{N}}, \quad \mathbb{P} := \bigotimes_{e\in\mathcal{E}} \mathbb{P}_e^{\mathbb{N}}; \tag{2.21}$$

in other words, we take the product measure over $e \in \mathcal{E}$ of $(\Omega_e^{\mathbb{N}}, \mathcal{F}_e^{\mathbb{N}}, \mathbb{P}_e^{\mathbb{N}})$, the canonical model for i.i.d random variables drawn from the environment e. Theorem 2.2 then holds by replacing the appropriate distributions with the above measures, and saying that the coverage statements hold in the limit as the $n_e \to \infty$ for all $e \in \mathcal{E}$.

2.5.1 Testing $H_{0,S}(\mathcal{E})$ in linear Gaussian models

We now detail the testing procedure for the Gaussian case (2.2) suggested by Peters et al. [2016], although we propose an elliptical confidence region which handles the intercept term. Here the coverage statements guaranteed by Theorem 2.2 will be exact, unlike for the testing procedure in Section 2.5.2 which could also be applied in this scenario. We begin with some notation. Fixing $S \subseteq \{1, \ldots, p\}$ and $e \in \mathcal{E}$, we let:

- $I_e \subseteq \{1, \ldots, n\}$ be the labels of the n_e observations corresponding to the environment $e \in \mathcal{E}$, and $I_{-e} := \{1, \ldots, n\} \setminus I_e$ for the remaining $n_{-e} := n - n_e$;
- $\mathbf{X}_{e,S}$ be the $n_e \times (1 + |S|)$ design matrix with rows $(1, (x_i)_S)$ for $i \in I_e$, corresponding to samples in I_e and random variables in S, and similarly $\mathbf{X}_{-e,S}$ for the samples in I_{-e};

- $\widehat{\beta}^{\text{pred}}(S)$ and $\widehat{\zeta}^{\text{pred}}(S)$ be the MLE's of β and ζ under the null hypothesis $H_{0,S}(\mathcal{E})$ (note the lack of e in the superscript); and finally
- $\widehat{Y}_e$ be the prediction vector for the observations $Y_e = (y_i : i \in I_e)$ using the MLE estimator computed on samples in I_{-e}, and $D = Y_e - \widehat{Y}^e$ be the vector of differences between the observed and predicted values.

The method then works as follows:

Step 1: Fix $S \subseteq \{1, \dots, p\}$. Then we reject $H_{0,S}(\mathcal{E})$ if, for any $e \in \mathcal{E}$,

$$\frac{D^T \Sigma_D^{-1} D}{\widetilde{\sigma}^2 n_e} > F_{n_e, n_{-e}-|S|-1}\left(\frac{\alpha}{|\mathcal{E}|}\right), \tag{2.22}$$

where $\widetilde{\sigma}^2$ is the (unbiased) estimate of the variance using samples in I_{-e}, the covariance matrix Σ_D and its inverse (after using the Woodbury matrix identity; see e.g Hager [1989]) are given by

$$\begin{aligned}
\Sigma_D &= 1_{n_e} + \mathbf{X}_{e,S}\left(\mathbf{X}_{-e,S}^T \mathbf{X}_{-e,S}\right)^{-1} \mathbf{X}_{e,S}^T, \\
\Sigma_D^{-1} &= 1_{n_e} - \mathbf{X}_{e,S}\left(\mathbf{X}_{-e,S}^T \mathbf{X}_{-e,S} + \mathbf{X}_{e,S}^T \mathbf{X}_{e,S}\right)^{-1} \mathbf{X}_{e,S}^T,
\end{aligned}$$

and $F_{r,s}(\widetilde{\alpha})$ is the $100(1-\widetilde{\alpha})\%$ quantile of the $\mathrm{F}_{r,s}$ distribution.

This corresponds to performing the Chow test [Chow, 1960] $|\mathcal{E}|$ times and then using the Bonferroni correction to account for the multiple tests. From a computational perspective, using Σ_D^{-1} directly means that only one matrix inversion is required rather than two, which is desirable both in terms of speed and stability. Alternatively, as $\Sigma_D^{-1} D$ is the solution to $\Sigma_D x = D$ for $x \in \mathbb{R}^{n_e}$, we can use LU decomposition (for example) to solve this without inversion.

Step 2: Repeat Step 1 for all $S \subseteq \{1, \dots, p\}$, starting with $S = \emptyset$. If we

(i) do not reject $H_{0,\emptyset}(\mathcal{E})$, or

(ii) do not reject $H_{0,S_1}(\mathcal{E})$ and $H_{0,S_2}(\mathcal{E})$ for some S_1, S_2 such that $S_1 \cap S_2 = \emptyset$,

then we stop and set $\widehat{S}(\mathcal{E}) = \emptyset$. Otherwise we set $\widehat{S}(\mathcal{E})$ as in (2.18).

Step 3: If we reject $H_{0,S}(\mathcal{E})$ set $\widehat{\Gamma}_S(\mathcal{E}) = \emptyset$. Otherwise, it contains $(\beta, \zeta) \in \mathbb{R}^p \times \mathbb{R}$ if and only if $\text{supp}(\beta) \subseteq S$ and

$$\left\| \mathbf{X}_S \begin{pmatrix} \zeta - \widehat{\zeta}^{\text{pred}}(S) \\ \beta - \widehat{\beta}^{\text{pred}}(S) \end{pmatrix} \right\|_2^2 \leq (|S|+1)\widetilde{\sigma}^2 F_{|S|+1, n-|S|-1}(\alpha), \tag{2.23}$$

where $\mathbf{X}_S$ is the $n \times (1 + |S|)$ design matrix for all the observations for variables in S and $\widetilde{\sigma}^2$ is the (unbiased) estimator of the variance using all the samples. We then form $\widehat{\Gamma}(\mathcal{E})$ as in (2.19).

To justify the data pooling used here to give the confidence region, recall that under the invariant prediction assumption, $Y^e|X^e_{S^*} = x$ does not depend on $e \in \mathcal{E}$ (Theorem 2.1). Therefore, even though the $(Y^e, X^e_{S^*})$ may vary in distribution across different $e \in \mathcal{E}$, as the confidence region depends only on the behaviour of the regression function, the results obtained by pooling are valid.

2.5.2 Testing $H_{0,S}(\mathcal{E})$ for generalized linear models

We now give a method for testing the null hypothesis $H_{0,S}(\mathcal{E})$ when the conditional distributions $Y^e|X^e = x$ behave as *any* generalized linear model, rather than just a Gaussian one. We implicitly assume any regularity conditions necessary to obtain the desired asymptotic coverage (see e.g. Jørgensen [1987]). Fixing $S \subseteq \{1, \dots, p\}$ and $e \in \mathcal{E}$, we require some more notation to that from Section 2.5.1; we let

- $\widehat{\beta}^{\text{pred},e}(S)$ and $\widehat{\zeta}^{\text{pred,e}}(S)$ be the MLE's for $\beta^{\text{pred},e}(S)$ and $\zeta^{\text{pred},e}(S)$ under the individual model for an environment $e \in \mathcal{E}$;
- $D := D(y; \widehat{\beta}^{\text{pred}}(S), \widehat{\zeta}^{\text{pred}}(S))$ be the deviance under the null hypothesis for all n observations y; and finally
- $D^e := D^e(y^e; \widehat{\beta}^{\text{pred},e}(S), \widehat{\zeta}^{\text{pred},e}(S))$ be the deviance for the model corresponding to the environment $e \in \mathcal{E}$ with observations Y_e only.

The testing procedure then works as follows:

Step 1: Fix $S \subseteq \{1, \dots, p\}$. If the dispersion parameter σ is known, we reject $H_{0,S}(\mathcal{E})$ if the test statistic

$$\frac{D - \sum_{e \in \mathcal{E}} D^e}{\sigma} > \chi^2_{(|S|+1)(|\mathcal{E}|-1)}(\alpha), \tag{2.24}$$

where $\chi^2_\nu(\alpha)$ is the upper α-quantile of a χ^2_ν distribution [Liao, 2002]. If it is unknown, then we instead reject $H_{0,S}(\mathcal{E})$ if

$$\frac{\frac{1}{(|S|+1)(|\mathcal{E}|-1)}\left(D - \sum_{e \in \mathcal{E}} D^e\right)}{\widehat{\sigma}(S)} > F_{(|S|+1)(|\mathcal{E}|-1),\, n-|\mathcal{E}|(|S|+1)}(\alpha) \tag{2.25}$$

where

$$\widehat{\sigma}(S) := \frac{1}{n - |\mathcal{E}|(|S|+1)} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \text{ given}$$
$$\widehat{\mu}_i := g^{-1}\left(\widehat{\zeta}^{\text{pred},e}(S) + x_i \widehat{\beta}^{\text{pred},e}(S)\right) \text{ for } i \in I_e,$$

is a consistent estimator of σ.

To explain where these tests arise from, note that under either (2.15) or (2.16), we can embed our models in a framework where our distributions belong to the a exponential dispersion family with the dispersion parameter fixed. Supposing $\mathcal{E} = \{e_1, \ldots, e_m\}$, the corresponding design matrix is diag $(\mathbf{X}_{e_1,S}, \ldots, \mathbf{X}_{e_m,S})$, and we seek to estimate $\left(\zeta^{\text{pred},e_1}(S), \beta^{\text{pred},e_1}(S)^T, \ldots, \zeta^{\text{pred},e_m}(S), \beta^{\text{pred},e_m}(S)^T\right)^T$. The test then corresponds to performing the (generalized) likelihood ratio test of

$$H_0 : \left(\beta^{\text{pred},e}(S), \zeta^{\text{pred},e}(S)\right) \text{ is constant across } e \in \mathcal{E}, \text{ against}$$
$$H_1 : \left(\beta^{\text{pred},e}(S), \zeta^{\text{pred},e}(S)\right) \neq \left(\beta^{\text{pred},f}(S), \zeta^{\text{pred},f}(S)\right) \text{ for some } e, f \in \mathcal{E}.$$

In a similar vein to how the Chow test was used in Section 2.5.1, we could instead test for whether the regression coefficients are different across $e \in \mathcal{E}$ and $\mathcal{E} \setminus \{e\}$ for each e. Let D^{-e} be the deviance for the model as under $H_{0,S}(\mathcal{E} \setminus \{e\})$, and $\widehat{\beta}^{\text{pred},-e}(S)$ and $\widehat{\zeta}^{\text{pred},-e}(S)$ be the respective MLE's of $\beta^{\text{pred},-e}(S)$ and $\zeta^{\text{pred},-e}(S)$. Then provided the dispersion parameter is known, we reject $H_{0,S}(\mathcal{E})$ if for any $e \in \mathcal{E}$,

$$\frac{D - (D^e + D^{-e})}{\sigma} > \chi^2_{|S|+1}\left(\frac{\alpha}{|\mathcal{E}|}\right). \tag{2.26}$$

Otherwise, we reject $H_{0,S}(\mathcal{E})$ if, for any $e \in \mathcal{E}$,

$$\frac{\frac{1}{|S|+1}\left(D - (D^e + D^{-e})\right)}{\widehat{\sigma}(S)} > F_{|S|+1,\, n-2(|S|+1)}\left(\frac{\alpha}{|\mathcal{E}|}\right), \tag{2.27}$$

where

$$\widehat{\sigma}(S) := \frac{1}{n - 2(|S|+1)} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \text{ given}$$
$$\widehat{\mu}_i := g^{-1}\left(\widehat{\zeta}^{\text{pred},e}(S) + x_i \widehat{\beta}^{\text{pred},e}(S)\right) \text{ if } i \in I_e, \text{ otherwise}$$
$$:= g^{-1}\left(\widehat{\zeta}^{\text{pred},-e}(S) + x_i \widehat{\beta}^{\text{pred},-e}(S)\right).$$

Although both tests proposed for $H_{0,S}(\mathcal{E})$ will have the same asymptotic size, for finite samples we would expect that they are powered to handle different types of changes across experimental settings. As the first contains the sum of the deviances across all $e \in \mathcal{E}$, we would expect it to perform best when there are small changes in the regression coefficients across multiple environments. Conversely, we would expect the second to perform better when there is a big change in one environment. The finite sample performance using both tests for $H_{0,S}(\mathcal{E})$ is investigated in Section 5.

Step 2: As in Section 2.5.1, we repeat Step 1 for all $S \subseteq \{1, \ldots, p\}$, starting with $S = \emptyset$. Again, if we

(i) do not reject $H_{0,\emptyset}(\mathcal{E})$, or

(ii) do not reject $H_{0,S_1}(\mathcal{E})$ and $H_{0,S_2}(\mathcal{E})$ for some S_1, S_2 such that $S_1 \cap S_2 = \emptyset$,

then we stop and set $\widehat{S}(\mathcal{E}) = \emptyset$. Otherwise we set $\widehat{S}(\mathcal{E})$ as in (2.18).

Step 3: If we reject $H_{0,S}(\mathcal{E})$, set $\widehat{\Gamma}_S(\mathcal{E}) = \emptyset$. Otherwise, it contains $(\beta, \zeta) \in \mathbb{R}^p \times \mathbb{R}$ if and only if $\text{supp}(\beta) \subseteq S$ and

$$\left\| \mathbf{W}^{1/2} \mathbf{X}_S \begin{pmatrix} \zeta - \widehat{\zeta}^{\text{pred}}(S) \\ \beta - \widehat{\beta}^{\text{pred}}(S) \end{pmatrix} \right\|_2^2 \leq \widehat{\sigma}(S) \chi^2_{|S|+1}(\alpha) \tag{2.28}$$

(c.f (2.23)), where $\mathbf{W}$ is a $n \times n$ diagonal matrix with entries

$$W_{ii} = \frac{1}{V(\widehat{\mu}_i)(g'(\widehat{\mu}_i))^2}. \tag{2.29}$$

The pooling can be justified with the same argument as before; similarly, we then form $\widehat{\Gamma}(\mathcal{E})$ as in (2.19).

Section 3

Invariant prediction for large p

In Section 2.4, we specified a method which requires testing $H_{0,S}(\mathcal{E})$ over all subsets S of $\{1, \dots, p\}$. Unfortunately, this is computationally infeasible for large p. As seen in Section 2.5, a few early stopping criteria can be used in practice to possibly save time in the case $\widehat{S}(\mathcal{E}) = \emptyset$. However, if we ever hope to a detect a non-empty set of identifiable causal predictors, we must test all $S \subseteq \{1, \dots, p\}$ as we a-priori do not know the true causal predictors.

We therefore discuss ways of performing invariant prediction when p is large, even when $p \gg n$. One way of handling this is to suppose that the number of causal variables q is actually far smaller than p, say for reasons of interpretation. We can then restrict $S(\mathcal{E})$ and $\widehat{S}(\mathcal{E})$ to intersections over $S \subseteq \{1, \dots, p\}$ of size $|S| \leq q$. This means that we need only test at most $\binom{p}{q} = O(p^q)$ subsets of $\{1, \dots, p\}$ rather than 2^p, and so the procedure gains a large computational speed-up. However, even for moderately size q, this may still be too large a number of subsets to test on.

To try and gain a further increase in speed, we can do so at some cost to the coverage probability of our estimator $\widehat{S}(\mathcal{E})$. Suppose we can find $B \subseteq \{1, \dots, p\}$ such that both $|B| \leq q$ and $\mathbb{P}(S^* \subseteq B) \geq 1 - \alpha$. Then by searching over subsets of B, $\widehat{S}(\mathcal{E}) \subseteq S^*$ holds with probability at least $1 - 2\alpha$ (or asymptotically so), in a manner analogous to Theorem 2.2.

To achieve this, we begin in Section 3.1 by proving a result which links the support set of an estimator $\widehat{\beta}$ for γ^* to its ℓ_1 error. In Section 3.2, we then propose a modified version of the Lasso [Tibshirani, 1994] to use, whose ℓ_1 error we investigate in Section 3.3. We end in Section 3.4 by showing that, under several conditions, we can guarantee $S^* \subseteq B$ with high probability. As one example, suppose $\min_{e \in \mathcal{E}} n_e \gg \log p$, our variables belong to a structural equation model with a sparse graph and $|S^*| = O(\log(p))$. Then provided B is of a similar order of magnitude, we only need to test $O(p)$ subsets of $\{1, \dots, p\}$, and yet have $\widehat{S}(\mathcal{E}) \subseteq S^*$ with high probability.

3.1 Linking coverage of S^* to ℓ_1 prediction error

We now discuss how to ensure that $S^* \subseteq B$ with high probability. To begin, suppose that the invariant prediction assumption holds, and that B arises as the support set of some estimator $\widehat{\beta}$ for γ^*. The following lemma (a minor refinement of Lemma 3 from Bunea [2008]) links the coverage of S^* by B to the ℓ_1 norm of the difference between $\widehat{\beta}$ and γ^* when restricted to entries in S^*.

Lemma 3.1. *In the scenario described above, we have that*

$$\mathbb{P}\left(S^* \not\subseteq B\right) \leq \mathbb{P}\left(\|\widehat{\beta} - \gamma^*\|_{1,S^*} \geq \min_{i \in S^*} |\gamma_i^*|\right),$$

where we define for $\emptyset \neq S \subseteq \{1, \ldots, p\}$ *the semi-norm* $\|x\|_{1,S} := \sum_{i \in S} |x_i|$.

Proof. This inequality is simply a consequence of how B and S^* are support sets for $\widehat{\beta}$ and γ^* respectively. Indeed,

$$\begin{aligned}
\mathbb{P}\left(S^* \not\subseteq B\right) &\leq \mathbb{P}\left(j \notin B \text{ for some } j \in S^*\right) \\
&\leq \mathbb{P}\left(\widehat{\beta}_j = 0 \text{ and } \gamma_j^* \neq 0 \text{ for some } j \in S^*\right) \\
&\leq \mathbb{P}\left(|\widehat{\beta}_j - \gamma_j^*| = |\gamma_j^*| \text{ for some } j \in S^*\right) \\
&\leq \mathbb{P}\left(\|\widehat{\beta} - \gamma^*\|_{1,S^*} \geq \min_{i \in S^*} |\gamma_i^*|\right), \qquad (\dagger)
\end{aligned}$$

where (†) follows as $\|\widehat{\beta} - \gamma^*\|_{1,S^*} \geq |\widehat{\beta}_j - \gamma_j^*|$ for $j \in S^*$ and $|\gamma_j^*| \geq \min_{i \in S^*} |\gamma_i^*|$. □

As a result, if we assume that $\min_{i \in S^*} |\gamma_i^*| > \delta$ for some δ *not depending on the data* (referred to as a *beta-min* condition in the literature), and we can guarantee that $\|\widehat{\beta} - \gamma^*\|_{1,S^*} \leq \delta$ with probability $1 - \alpha$, then $\mathbb{P}(S^* \subseteq B) \geq 1 - \alpha$ as desired.

One basic way to use this is to look at the (Y^e, X^e) separately for $e \in \mathcal{E}$, each giving an estimator $\widehat{\beta}^e$ and corresponding support set $B^e \subseteq \{1, \ldots, p\}$. The set $\cap_{e \in \mathcal{E}} B^e$ will then contain S^* with high probability. More precisely, suppose for each $e \in \mathcal{E}$ that, for all $\alpha \in (0, 1)$, there exists $\delta(e, \alpha)$ such that,

$$\min_{i \in S^*} |\gamma_i^*| > \delta(e, \alpha) \text{ and } \mathbb{P}\left(\|\widehat{\beta}^e - \gamma^*\|_{1,S^*} \geq \delta(e, \alpha)\right) \leq \alpha \qquad (3.1)$$

so by Lemma 3.1, $\mathbb{P}(S^* \subseteq B^e) \geq 1 - \alpha$. Fixing α, if we then also have that

$$\min_{i \in S^*} |\gamma_i^*| > \max_{e \in \mathcal{E}} \delta\left(e, \tfrac{\alpha}{|\mathcal{E}|}\right) \text{ and } \mathbb{P}\left(\|\widehat{\beta}^e - \gamma^*\|_{1,S^*} \geq \delta\left(e, \tfrac{\alpha}{|\mathcal{E}|}\right)\right) \leq \frac{\alpha}{|\mathcal{E}|} \text{ for all } e \in \mathcal{E}, \tag{3.2}$$

we may deduce that $\mathbb{P}(S^* \subseteq \cap_{e \in \mathcal{E}} B^e) \geq 1 - \alpha$. However, if (2.1) is true, we would expect that the B^e are similar across $e \in \mathcal{E}$, and so $\cap_{e \in \mathcal{E}} B^e$ should not be much smaller than any of the B^e. Moreover, as this method does not directly try and take into account that the $\widehat{\beta}^e$ should be identical across different $e \in \mathcal{E}$, we might expect it to be sub-optimal in some regards.

3.2 A generalization of the Lasso

We therefore focus on handling a estimator for γ^* which attempts to use all the information across different $e \in \mathcal{E}$. Here we specialize to the linear Gaussian model as in (2.2): we assume there exists $\eta^* \in \mathbb{R}$ and $\gamma^* \in \mathbb{R}^p$ with $S^* := \text{supp}(\gamma^*)$ such that

$$Y^e = \eta^* + X^e \gamma^* + \epsilon^e \text{ where } \epsilon^e \perp\!\!\!\perp X^e_{S^*} \text{ and } \epsilon^e \sim N(0, \sigma^2) \tag{3.3}$$

for all $e \in \mathcal{E}$. Note that is *not* necessarily true that $\epsilon^e \perp\!\!\!\perp X^e$, *only* that $\epsilon^e \perp\!\!\!\perp X^e_{S^*}$. We take independent samples (y_i, x_i) for $i = 1, \ldots, n$, where $I_e \subseteq \{1, \ldots, n\}$ is the index set of samples taken from (Y^e, X^e). We then denote $Y_e = (y_i)_{i \in I_e}$ and $\mathbf{X}_e$ for the design matrix (*without intercept*) whose rows consist of the x_i for $i \in I_e$. Without loss of generality, we may remove the intercept from (3.3) so that

$$Y^e = X^e \gamma^* + \epsilon^e \text{ where } \mathbb{E}[Y^e] = \mathbb{E}[X^e] = 0, \tag{3.4}$$

and that Y_e and the columns of $\mathbf{X}_e$ are all mean-centred. This means that our samples are such that

$$Y_e = \mathbf{X}_e \gamma^* + \epsilon_e - \bar{\epsilon_e} 1_{n_e} \text{ for all } e \in \mathcal{E}, \tag{3.5}$$

where $\epsilon_e = (\epsilon_i)_{i \in I_e}$, $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ for some $\sigma^2 > 0$ and $\bar{\epsilon_e} = \sum_{i \in I_e} \epsilon_i / n_e$.

Given this framework, we then define for $\lambda \geq 0$ the *invariant Lasso* by

$$\widehat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \sum_{e \in \mathcal{E}} \frac{1}{2n_e} ||Y_e - \mathbf{X}_e \beta||_2^2 + \lambda \|\beta\|_1 \right\}. \tag{3.6}$$

We use $\in$ as a solution may not exist uniquely, although one does exist (see Appendix 3). The non-uniqueness means that we must be careful when considering statements such as $S^* \subseteq B$, as the support set of invariant Lasso solutions may not be unique. Note that if $|\mathcal{E}| = 1$, (3.6) corresponds to an ordinary Lasso estimator for γ^*. Generally, (3.6) may be thought as a re-weighted version of the ordinary Lasso, which avoids biasing against environmental settings $e \in \mathcal{E}$ with samples sizes n_e which are small when compared to other settings.

Suppose we were to pool variables in $\mathcal{F} \subseteq \mathcal{E}$ together to form a single environment, which we also denote as $\mathcal{F}$. The ordinary Lasso then corresponds to the case when $\mathcal{F} = \mathcal{E}$. Now, note that in practice we do not sample from $(Y^{\mathcal{F}}, X^{\mathcal{F}})$ and actually sample from the (Y^e, X^e) individually. Denoting $n_{\mathcal{F}} = \sum_{e \in \mathcal{F}} n_e$ and $I_{\mathcal{F}} = \cup_{e \in \mathcal{F}} I_e$, we know that

$$\frac{1}{n_{\mathcal{F}}} \sum_{i \in I_{\mathcal{F}}} (y_i - x_i \beta)^2 = \sum_{e \in \mathcal{F}} \frac{n_e}{n_{\mathcal{F}}} \left(\frac{1}{n_e} \sum_{i \in I_e} (y_i - x_i \beta)^2 \right). \tag{3.7}$$

Therefore, if we pool together settings with unequal samples size, in (3.6) we are penalizing weighted sums of the mean squared error of y_i. This means that environments with smaller numbers of samples contribute less to the objective function, even if from an information-theoretic perspective they could tell us more about γ^* than other environments with the same sample size.

Returning to properties of the invariant Lasso, as the optimization problem in (3.6) is convex, we know that β is a solution if and only if the KKT conditions

$$\sum_{e \in \mathcal{E}} \frac{1}{n_e} \mathbf{X}_e^T (Y_e - \mathbf{X}_e \beta) = \lambda \nu \text{ for some } \nu \text{ such that } \|\nu\|_\infty \leq 1 \\ \text{and, if } A = \text{supp}(\beta), \text{ then } \nu_A = \text{sgn}(\beta_A), \tag{3.8}$$

are satisfied (see e.g Boyd and Vandenberghe [2004]). Furthermore, the fitted values $\mathbf{X}_e \widehat{\beta}_\lambda$ are unique for all $e \in \mathcal{E}$ given any invariant Lasso solution. The value of $||\widehat{\beta}_\lambda||_1$ is also unique; see Appendix 3 for further details on these two points. As a consequence, the following analogue of the *equicorrelation set*

$$\widehat{E}_\lambda := \left\{ k \in \{1, \ldots, p\} : \left| \sum_{e \in \mathcal{E}} \frac{1}{n_e} \mathbf{X}_e^T (Y_e - \mathbf{X}_e \widehat{\beta}_\lambda) \right|_k = \lambda \right\} \tag{3.9}$$

is well defined for all invariant Lasso solutions $\widehat{\beta}_\lambda$.

By the KKT conditions (3.8), this contains the set of non-zero indices for every invariant Lasso solution; however, it is not necessarily equal to the support set of every solution, which may vary. On a similar note, as we analyse the coverage probability of $S^* \subseteq B$ by examining the ℓ_1-norm of $\widehat{\beta}_\lambda - \gamma^*$ on S^*, we need to ensure that such a bound holds independently of the solution.

3.3 A compatibility condition

To get such a bound, we require some conditions on the design matrices $\mathbf{X}_e$. We say the *compatibility condition*[1] is satisfied if there exists a constant $\phi > 0$ such that

$$\mathbb{P}\left(\sum_{e\in\mathcal{E}} \frac{1}{n_e} \|\mathbf{X}_e\beta\|_2^2 \geq \frac{\phi^2}{|S^*|} \|\beta\|_{1,S^*}^2\right) = 1 \tag{3.10}$$

for all $\beta \in \mathbb{R}^p$ such that $\beta_{S^*} \neq 0$ and $\|\beta\|_{1,-S^*} \leq 3\|\beta\|_{1,S^*}$, where we denote $-S^* := \{1,\dots,p\} \setminus S^*$. Intuitively, this says that provided "enough of the mass" of $\|\beta\|_1$ lies within S^*, then $\sum_{e\in\mathcal{E}} \frac{1}{n_e}\|\mathbf{X}_e\beta\|_2^2 \asymp \|\beta\|_2^2$ almost surely. We now prove a result similar to e.g Theorem 2.2 of Bunea [2008], or Theorem 6.4 of Bühlmann and van de Geer [2011], which gives an upper bound on $\|\widehat{\beta}_\lambda - \gamma^*\|_{1,S^*}$ which holds with high probability provided the X^e are bounded[2] and (3.10) holds.

Theorem 3.2. *Suppose we have a series of linear Gaussian models as in* (3.4)*, with independent observations* (y_i, x_i) *as in* (3.5)*. Further suppose that the compatibility condition* (3.10) *holds with constant* $\phi > 0$*, and that there exists* $M > 0$ *such that* $|X_i^e| \leq M$ *almost surely for all* $e \in \mathcal{E}$ *and* $i = 1,\dots,p$*. Fix* $\lambda > 0$*. Then for all invariant Lasso solutions* $\widehat{\beta}_\lambda$ *to* (3.6)*,*

$$\|\gamma^* - \widehat{\beta}_\lambda\|_{1,S^*} \leq \frac{3\lambda|S^*|}{2\phi^2}$$

occurs with probability at least

$$1 - \exp\left(-\frac{\lambda^2 n_{min}}{8|\mathcal{E}|M^2\sigma^2}\right)\left[|S^*| + 2(p-|S^*|)\left\{1 + \exp\left(\frac{-\lambda^2 n_{min}^2}{8|\mathcal{E}|^2M^2\sigma^2 n_{max}^2}\right)\right\}^n\right]$$

where $n_{min} := \min_{e\in\mathcal{E}} n_e$*,* $n_{max} := \max_{e\in\mathcal{E}} n_e$ *and* $n = \sum_{e\in\mathcal{E}} n_e$*.*

[1] This is a generalization of an "almost sure" version of the compatibility condition for the ordinary Lasso when the design matrices $\mathbf{X}_e$ are fixed.

[2] This assumption is common for theoretical results about the (ordinary) Lasso when the design matrix is not fixed; see e.g Bunea [2008] or Chatterjee [2013].

Proof. Fix a solution $\widehat{\beta}_\lambda$ of (3.6). We begin by deriving the following analogue of the "basic inequality" for the ordinary Lasso:

$$\sum_{e\in\mathcal{E}} \frac{1}{n_e}\|\mathbf{X}_e(\gamma^* - \widehat{\beta}_\lambda)\|_2^2 \leq \sum_{e\in\mathcal{E}} \frac{1}{n_e}(\epsilon^e)^T\mathbf{X}_e(\widehat{\beta}_\lambda - \gamma^*) + \lambda\left\|\gamma^*\right\|_1 - \lambda\|\widehat{\beta}_\lambda\|_1. \qquad (\dagger)$$

Note that the uniqueness of fitted values across $e \in \mathcal{E}$ implies that this holds for all invariant Lasso solutions. Now, by multiplying both sides of the KKT condition (3.8) by $(\gamma^* - \widehat{\beta}_\lambda)^T$, we find that

$$\sum_{e\in\mathcal{E}} \frac{1}{n_e}(\mathbf{X}_e(\gamma^* - \widehat{\beta}_\lambda))^T(Y_e - \mathbf{X}_e\widehat{\beta}_\lambda) = \lambda\nu(\gamma^* - \widehat{\beta}_\lambda)^T$$

where ν has $\|\nu\|_\infty \leq 1$ and, writing $A := \mathrm{supp}(\widehat{\beta}_\lambda)$, $\nu_A = \mathrm{sgn}(\widehat{\beta}_{\lambda,A})$. Then by using

(i) $Y^e - \mathbf{X}_e\widehat{\beta}_\lambda = \mathbf{X}_e(\gamma^* - \widehat{\beta}_\lambda) + \epsilon_e - \bar{\epsilon_e}1_{n_e}$,

(ii) $(\bar{\epsilon_e}1_{n_e})^T\mathbf{X}_e = 0$ as $\mathbf{X}_e$ has mean centred columns,

(iii) $\nu\widehat{\beta}_\lambda^T = \|\widehat{\beta}_\lambda\|_1$ by definition of ν,

(iv) $|\nu(\gamma^*)^T| \leq \|\nu\|_\infty \|\gamma^*\|_1 = \|\gamma^*\|_1$ by Hölder's inequality

and rearranging, we obtain (†).

We now work on the event Ω where $\left\|\sum_{e\in\mathcal{E}} \mathbf{X}_e^T\epsilon_e/n_e\right\|_\infty \leq \lambda/2$. By Holder's inequality, we obtain that

$$\left|\sum_{e\in\mathcal{E}} \frac{1}{n_e}(\epsilon_e)^T\mathbf{X}_e(\widehat{\beta}_\lambda - \gamma^*)\right| \leq \frac{\lambda}{2}\|\widehat{\beta}_\lambda - \gamma^*\|_1$$

and therefore by (†) that

$$C := \sum_{e\in\mathcal{E}} \frac{1}{\lambda n_e}\|\mathbf{X}_e(\gamma^* - \widehat{\beta}_\lambda)\|_2^2 \leq \frac{1}{2}\|\widehat{\beta}_\lambda - \gamma^*\|_1 + \|\gamma^*\|_1 - \|\widehat{\beta}_\lambda\|_1.$$

Using that $\|v\|_1 = \|v\|_{1,A} + \|v\|_{1,-A}$ for $A \subseteq \{1,\ldots,p\}$ and $\gamma^*_{-S^*} = 0$, it follows that

$$C + \frac{1}{2}\|\widehat{\beta}_\lambda - \gamma^*\|_{1,-S^*} \leq \frac{3}{2}\|\widehat{\beta}_\lambda - \gamma^*\|_{1,S^*}.$$

Finally, we obtain on Ω intersected by an event of probability one that

$$\|\gamma^* - \widehat{\beta}_\lambda\|_{1,S^*} \leq \frac{3\lambda|S^*|}{2\phi^2}.$$

by using the compatibility condition (3.10), multiplying both sides by λ and rearranging. To find a lower bound on $\mathbb{P}(\Omega)$, we note that

$$\mathbb{P}(\Omega^c) = \mathbb{P}\left(|Z_j| > \frac{\lambda}{2} \text{ for some } j \in \{1, \dots, p\}\right) \leq \sum_{j=1}^{p} \mathbb{P}\left(|Z_j| > \frac{\lambda}{2}\right)$$

by Boole's inequality, where we define

$$Z_j := \left(\sum_{e \in \mathcal{E}} \frac{1}{n_e} \mathbf{X}_e^T \epsilon^e\right)_j = \sum_{e \in \mathcal{E}} \frac{1}{n_e} \sum_{i \in I_e} x_{ij} \epsilon_i.$$

The result then follows by using Lemma 3.3 below. □

Lemma 3.3. *Under the conditions stated in Theorem 3.2, we have that*

$$\mathbb{P}\left(|Z_j| > t\right) \leq \exp\left(-\frac{t^2 n_{min}}{2|\mathcal{E}|M^2\sigma^2}\right)$$

if $j \in S^*$ *or* $X_j^e \perp\!\!\!\perp \epsilon$ *for all* $e \in \mathcal{E}$*; otherwise*

$$\mathbb{P}\left(|Z_j| > t\right) \leq 2\exp\left(\frac{-t^2 n_{min}}{2|\mathcal{E}|M^2\sigma^2}\right)\left\{1 + \exp\left(\frac{-t^2 n_{min}^2}{2|\mathcal{E}|^2 M^2 \sigma^2 n_{max}^2}\right)\right\}^n.$$

Proof. See Appendix 4. □

As we can obtain tighter bounds on $\mathbb{P}(|Z_j| > t)$ whenever $X_j^e \perp\!\!\!\perp \epsilon$ for all $e \in \mathcal{E}$, if we let $\mathcal{I}$ be the set of indices for which this occurs (so $S^* \subseteq \mathcal{I}$), the lower bound in 3.2 can be improved to

$$1 - \exp\left(-\frac{\lambda^2 n_{\min}}{8|\mathcal{E}|M^2\sigma^2}\right)\left[|\mathcal{I}| + 2(p - |\mathcal{I}|)\left\{1 + \exp\left(\frac{-\lambda^2 n_{\min}^2}{8|\mathcal{E}|^2 M^2 \sigma^2 n_{\max}^2}\right)\right\}^n\right]. \quad (3.11)$$

If we can guarantee that $|\mathcal{I}|$ is large, then the $2(n - |\mathcal{I}|)\{\cdots\}^n$ term above is mostly negligible, ensuring that $S^* \subseteq \widehat{B}_\lambda$ occurs with far higher probability (for the same λ) than if $|\mathcal{I}|$ were small. One example where this can occur is when Y and the X_i form a structural equation model. By iterating the governing equations of the model, X_j can be written as a function of only the noise terms ϵ_i for $i \in \text{an}(j)$. Moreover, this property is preserved by interventions which e.g do not change the structure of the underlying directed acyclic graph. Therefore, if Y does not belong to a large number of ancestor sets, either because of e.g sparsity or as Y appears late in some topological ordering, we can expect $|\mathcal{I}|$ to be large.

3.4 Guaranteeing $S^* \subseteq \widehat{B}_\lambda$ with high probability

Using the above results so far, we obtain the following:

Corollary 3.4. *Suppose we have a series of linear Gaussian models as in* (3.4) *and independent observations* (y_i, x_i) *as in* (3.5). *Fix* $0 < \alpha < 1$ *and let* $n_{min} := \min_{e\in\mathcal{E}} n_e$, $n_{max} := \max_{e\in\mathcal{E}} n_e$. *Further suppose that (i) the compatibility condition* (3.10) *is true with constant* $\phi > 0$, *(ii) we choose* λ *greater than*

$$2M\sigma \min_{1\leq C\leq 2^n} \left(\frac{|\mathcal{E}|n_{max}}{n_{min}} \sqrt{-2\log(C^{1/n}-1)} \right) \vee \sqrt{\frac{|\mathcal{E}|}{n_{min}} 2\log\left(\frac{|\mathcal{I}| + 2C(p-|\mathcal{I}|)}{\alpha} \right)}$$

and (iii) the "beta-min" condition

$$\min_{i\in S^*} |\gamma_i^*| > \frac{3\lambda|S^*|}{2\phi^2}$$

holds. Then if B_λ *is the support set of any invariant Lasso solution* $\widehat{\beta}_\lambda$ *to* (3.6), *we have* $S^* \subseteq B_\lambda$ *with probability at least* $1-\alpha$.

Proof. Condition (i) allows us to use Theorem 3.2, so we know that $S^* \not\subseteq B_\lambda$ occurs with probability no more than

$$\exp\left(-\frac{\lambda^2 n_{\min}}{8|\mathcal{E}|M^2\sigma^2}\right)\left[|\mathcal{I}| + 2(p-|\mathcal{I}|)\left\{1+\exp\left(\frac{-\lambda^2 n_{\min}^2}{8|\mathcal{E}|^2M^2\sigma^2 n_{\max}^2}\right)\right\}^n\right].$$

We now want λ such that this quantity is less than α. We begin by controlling the $\{\cdots\}^n$ term. Let $1 \leq C \leq 2^n$ be a constant such that

$$\left\{1+\exp\left(\frac{-\lambda^2 n_{\min}^2}{8|\mathcal{E}|^2M^2\sigma^2 n_{\max}^2}\right)\right\}^n \leq C \iff \lambda \geq 2M\sigma\frac{|\mathcal{E}|n_{\max}}{n_{\min}}\sqrt{-2\log(C^{1/n}-1)}$$

Given this, we can then choose λ which satisfies the above and

$$\exp\left(-\frac{\lambda^2 n_{\min}}{8|\mathcal{E}|M^2\sigma^2}\right)[|\mathcal{I}| + 2C(p-|\mathcal{I}|)] \leq \alpha$$
$$\iff \lambda \geq 2M\sigma\sqrt{\frac{|\mathcal{E}|}{n_{\min}}2\log\left(\frac{|\mathcal{I}|+2C(p-|\mathcal{I}|)}{\alpha}\right)}.$$

To get the exact statement of condition (ii), we optimise these bounds over C. Using condition (iii) with Lemma 3.1 then gives the desired conclusion. □

Rather than detail an optimal choice of C in full generality, we end by briefly describing the behaviour when $|\mathcal{I}| = \eta p$ for some η close to one. Suppose that $p \ll 2^n$, and η is such that $1-\eta = p2^{-(n+1)}$. In this case, we choose $C = 2^n$ to give a lower bound for λ in Corollary 3.4. As a consequence, if we write $\min_{i\in S^*} |\gamma_i^*| = D$, then provided we have bounds

$$2\sigma M\sqrt{\frac{2|\mathcal{E}|(2\log p - \log\alpha)}{n_{\min}}} \lesssim \lambda \leq \frac{2D\phi^2}{3|S^*|}, \tag{3.12}$$

the support set of any invariant Lasso solution will contain S^* with probability of (approximately) $1-\alpha$. As the $\log\alpha$ term is negligible for sufficiently large p, rearranging this inequality gives

$$|S^*| \lesssim \frac{D\phi^2}{3\sigma M}\sqrt{\frac{n_{\min}}{4|\mathcal{E}|\log p}}. \tag{3.13}$$

In particular, this tells us that Corollary 3.4 likely can be applied in the regime where $n_{\min} \gg \log p$, noting that this condition is mostly consistent with the requirement that $p \ll 2^n$.

An interesting feature of the requirements on λ in this case is that they feature only $n_{\min}$, and not also the ratio $n_{\max}/n_{\min}$ as in the general case. This is a consequence of the number of j where $X_j^e \not\perp\!\!\!\perp \epsilon$ for some $e \in \mathcal{E}$ is very small. Therefore, even if we sample largely from the $e \in \mathcal{E}$ where this occurs often, there is little impact on the overall behaviour of the estimator. However, if $\mathcal{I}$ is not large, we need the ratio $n_{\max}/n_{\min}$ to be small, as otherwise it may be possible to take a large number of samples from $e \in \mathcal{E}$ where $X_j^e \not\perp\!\!\!\perp \epsilon$ for a large number of j. In this case, the estimate $\widehat{B}_\lambda$ is unlikely to contain S^*.

Section 4

Links with structural equation modelling

Having previously discussed both structural equation modelling (Section 1) and invariant prediction (Sections 2 and 3), we now discuss some links and differences between the two. In Section 4.1 we discuss how, given a variable Y in a structural equation model, its set of parents $\text{pa}(Y)$ satisfies the invariant prediction assumption whenever $\mathcal{E}$ consists of interventions not acting on the variable of interest. We then discuss what other subsets of variables satisfy this property in Section 4.2.

As we will find that, provided the structural equation model satisfies causal minimality, $\text{pa}(Y)$ is the unique minimal set which satisfies this property, we then seek to answer when we can ensure that $S(\mathcal{E}) = \text{pa}(Y)$. In Section 4.3, we show that for a certain types of structural equation model, this can be achieved whenever $\mathcal{E}$ contains certain types of interventions which act on a specific single (but a-priori unknown) variable within the model. We then end in Section 4.4 by briefly discussing a few differences between the two frameworks.

4.1 Invariant prediction and autonomy

Recall that one of the important features of a structural equation model is that it informs us what happens when we intervene on variables within it. From an inference perspective, this is important as causal relationships are invariant under different interventional settings - a property sometimes referred to as *autonomy*.

To be precise, let f and $\widetilde{f}$ be densities for $X \in \mathbb{R}^p$ under the law of a structural equation model prior and post intervention respectively. Then one can show that

$$f(x_k|x_{\text{pa}(k)}) = \widetilde{f}(x_k|x_{\text{pa}(k)}) \tag{4.1}$$

provided that the intervention does not involve X_k. We now illustrate a similar principle which shows that, under the same conditions, the invariant prediction assumption is satisfied when S^* is the set of parents of Y.

Theorem 4.1. *Suppose* $(X_0 = Y, X_1, \ldots, X_p) \in \mathbb{R}^{p+1}$ *is generated by a structural equation model, so that*

$$X_i = f_i(X_{pa(i)}, \epsilon_i) \text{ for } i = 0, \ldots, p,$$

for some functions f_i *and noise distributions* ϵ_i *(which are jointly independent). Suppose* $\mathcal{E}$ *is a (non-empty) set of interventional settings, as defined in* (1.2)*, where we denote variables under the interventional setting* $e \in \mathcal{E}$ *by a superscript* e*. Further suppose that all interventions act only on variables in* $\{X_1, \ldots, X_p\}$ *and never on* Y*. Then the invariant prediction assumption* (2.1) *is true with* $S^* = pa(0)$*,* $F_\epsilon \overset{d}{=} \epsilon_0$ *and* $h = f_0$*.*

Proof. As interventions are never performed on Y, the structural equation model tells us that $Y^e = f_0(X^e_{\mathrm{pa}(0)}, \epsilon^e_0)$ for all $e \in \mathcal{E}$. By definition of an intervention, we know that for all $e, f \in \mathcal{E}$ that (i) $\epsilon^e_0 \overset{d}{=} \epsilon^f_0$ and (ii) $\epsilon^e_0 \perp\!\!\!\perp \{\epsilon^e_i \,|\, i \in \mathrm{an}(0)\}$. By (ii) we may deduce that $\epsilon^e_0 \perp\!\!\!\perp X^e_{\mathrm{pa}(0)}$. This follows by recursively building

$$\epsilon^e_0 \perp\!\!\!\perp \{X_i \,|\, i \in \mathrm{an}(0), \pi(i) \leq j\},$$

where π is a topological ordering on the corresponding directed acyclic graph of the structural equation model, until $j = \max\{\pi(i) \,|\, i \in \mathrm{pa}(0)\}$. Then as (i) guarantees that ϵ^e_0 has the same distribution across all $e \in \mathcal{E}$, the invariant prediction assumption is satisfied under the desired conditions. □

One scenario where this is useful is when $Y|X_{\mathrm{pa}(Y)} = x \sim \mathrm{ED}(\mu_x, \sigma)$, where $g(\mu_x) = \eta^* + x\gamma^*$ and η^*, γ^* are known, as then the methodology developed in Section 2 can be used as a way of falsifying the structural equation model. Firstly, this implies that $H_{0,\mathrm{pa}(Y)}(\mathcal{E})$ is true, which we can test using the methods developed in Section 2.5. Secondly, as this only tests whether there exist η^*, γ^* such that the above relationship holds, we can check whether the confidence region produced actually contains the "true" values (η^*, γ^*) as according to the structural equation model. More generally, as the only requirement on $\mathcal{E}$ is that it contains interventions, invariant prediction may be used naturally alongside structural equation modelling.

4.2 The global invariant prediction assumption

Although Proposition 4.1 guarantees that the invariant prediction assumption holds for $S^* = \text{pa}(Y)$ under mild conditions on the interventions performed, the question arises of whether this is the *unique* such subset of $\{1, \dots, p\}$. In other words, is the invariant prediction assumption also true for another $S^* \neq \text{pa}(0)$, given any $\mathcal{E}$ which satisfies the conditions of Proposition 4.1? We refer to this property as saying that the *global invariant prediction assumption* is satisfied.

There are some trivial counterexamples to this. For example, if we have a structural equation model where

$$Y = X_3^2 + X_1 X_2 + 0 \cdot X_4 + \epsilon,$$

then although $\text{pa}(0) = \{1, 2, 3, 4\}$, it is clear that $S^* = \{1, 2, 3\}$ would suffice for the global invariant prediction assumption to hold. Here the model fails to satisfy causal minimality; in fact, this type of behaviour can only occur when causal minimality is not satisfied.

Proposition 4.2. *Let $\mathcal{S}$ be a structural equation model for $X = (Y = X_0, X_1, \dots, X_p)$ and P be the distribution of X. Suppose that P is absolutely continuous with respect to a product measure, and satisfies causal minimality with respect to the associated directed acyclic graph G. Then the global invariant prediction assumption cannot be true for any $T^* \subset S^* = pa(0)$.*

Proof. Suppose the global invariant prediction assumption is true for $T^* \subset \text{pa}(0)$, so that for any set of interventions $\mathcal{E}$ which only act on $\{X_1, \dots, X_p\}$, we know that

$$Y^e = g(X^e_{T^*}, \upsilon^e) \text{ where } \upsilon^e \perp\!\!\!\perp X^e_{T^*} \text{ and } \upsilon^e \sim G_\upsilon$$

for some function g and noise distribution G_υ. Now let $\mathcal{E} = \{e\}$, where e corresponds to performing no intervention, and G' be the subgraph of G where $\text{pa}(0) = T^*$ in G'. In particular, this means that G' is a directed acyclic graph and X is generated by a structural equation model with distribution P and associated directed acyclic graph G'. Then by Theorem 1.1, P satisfies the Markov property with respect to G', which contradicts causal minimality. □

Despite this result, there are still a large number of possible S^* to rule out. As

the following result shows, the global invariant prediction assumption is true for many S^* containing pa(0), but in a "redundant" way.

Proposition 4.3. *Let $\mathcal{S}$ be a structural equation model for $X = (Y = X_0, X_1, \ldots, X_p)$. Suppose that the distribution P of X is absolutely continuous with respect to a product measure, say with density f. Then the global invariant prediction assumption is satisfied by any $pa(0) \cup T$, where $T \subseteq nd(0)$.*

Proof. Without loss of generality we may assume $T \subseteq \mathrm{nd}(0) \setminus \mathrm{pa}(0)$; fix such a T. This result then follows from the fact that $X_0 \perp\!\!\!\perp X_{\mathrm{nd}(0)\setminus\mathrm{pa}(0)} \,|\, X_{\mathrm{pa}(0)}$. Given this, it implies that $X_0 \perp\!\!\!\perp X_T \,|\, X_{\mathrm{pa}(0)}$, and therefore that

$$f_{Y|X_{\mathrm{pa}(0)\cup T}}(y \,|\, x_{\mathrm{pa}(0)\cup T}) = f_{Y|X_{\mathrm{pa}(0)}}(y \,|\, x_{\mathrm{pa}(0)}).$$

By Theorem 4.1, we know that the right hand side is invariant (up to almost sure equivalence) under any set of interventions $\mathcal{E}$ which do not act on Y. This therefore holds for the left hand side, which then gives the desired result.

In order to prove that $X_0 \perp\!\!\!\perp X_{\mathrm{nd}(0)\setminus\mathrm{pa}(0)} \,|\, X_{\mathrm{pa}(0)}$, due to the conditions we impose on P, it is enough to show that 0 and $\mathrm{nd}(0) \setminus \mathrm{pa}(0)$ are d-separated by pa(0). Take a path from 0 to any node in $\mathrm{nd}(0) \setminus \mathrm{pa}(0)$. There are then two possibilities:

(i) If the path starts as $0 \leftarrow j$, then the path is blocked by pa(0) as $j \in \mathrm{pa}(0)$ and j is not a collider.

(ii) If the path starts as $0 \rightarrow j$, then the path must contain a collider, as else we would have a directed path from 0 to a node in $\mathrm{nd}(0) \setminus \mathrm{pa}(0)$, which is absurd. If we then take the first colliding node, it nor any of its descendants can lie in pa(0) as else we create a cycle, meaning that the path is blocked by pa(0). □

This suggests that the correct question to ask is whether there exists a *unique minimal* subset S^* (with respect to inclusion) which satisfies the global invariant prediction assumption. As $\mathcal{S}(\mathcal{E})$ depends only on the intersection of these minimal subsets, a positive answer to this tells us that it is sensible to ask whether we can find finite $\mathcal{E}$ such that $S(\mathcal{E}) = S^*$. The following result shows that the answer to this question is yes; provided causal minimality holds, this set is therefore pa(0).

Theorem 4.4. *Let $\mathcal{S}$ be a structural equation model for $X = (Y = X_0, X_1, \ldots, X_p)$. Then there exists a unique minimal set S such that the global invariant prediction assumption is satisfied by S.*

Proof. By Theorem 4.1 we know that a minimal set exists. For sake of contradiction, suppose we have two distinct minimal (so non-empty) sets S and T. Let $V = S \cap T$. Then for any set $\mathcal{E}$ of interventions which do not intervene on Y, we have

$$\begin{aligned} Y^e &= f(X_V^e, X_{S\setminus V}^e, \epsilon^e), \text{ where } \epsilon^e \perp\!\!\!\perp X_S^e \text{ and } \epsilon^e \sim F_\epsilon \\ &= g(X_V^e, X_{T\setminus V}^e, \upsilon^e), \text{ where } \upsilon^e \perp\!\!\!\perp X_T^e \text{ and } \upsilon^e \sim G_\upsilon \end{aligned}$$

for all $e \in \mathcal{E}$, some choice of functions f, g and some noise distributions F_ϵ, G_υ. Now consider the interventional settings

$$e \to \text{do}\left(X_V^e = x_V, X_{S\setminus V}^e = x_{S\setminus V}, X_{T\setminus V}^e = x_{T\setminus V}\right).$$

By fixing x_V and varying $x_{S\setminus V}$, it follows that $f(x_V, x_{S\setminus V}, \epsilon) = f(x_V, \epsilon)$. But then

$$Y^e = f(X_V^e, \epsilon^e) \text{ where } \epsilon^e \perp\!\!\!\perp X_V^e, \epsilon^e \sim F_\epsilon$$

for all $e \in \mathcal{E}$, so V satisfies the global invariant prediction assumption. However, this contradicts the minimality of S and T. □

4.3 When does $S(\mathcal{E}) = S^*$?

The results in Section 4.2 tell us that is sensible to ask whether there exists a *finite* (and ideally practically realisable) set of interventional settings $\mathcal{E}$ for which $S(\mathcal{E}) = S^*$. To illustrate one case of when this is so, we generalize a result of Peters et al. [2016, Theorem 2] on guaranteeing $S(\mathcal{E}) = S^*$ in linear Gaussian structural equation models.

We do so by extending the result to handle *additive structural equation models*. Supposing $X = (Y = X_0, X_1, \ldots, X_p)$, this is a structural equation model $\mathcal{S}$ (with associated directed acyclic graph G) such that

$$X_i = \sum_{j=0}^{p} f_{i,j}(X_j) + \epsilon_i \tag{4.2}$$

for some functions $f_{i,j} : \mathbb{R} \to \mathbb{R}$, belonging to some fixed function class $\mathcal{C}$, such that $j \to k$ in G if and only if $f_{k,j} \not\equiv 0$. We further assume that

$$\mathcal{C} \subseteq \mathcal{C}' := \left\{ f : \mathbb{R} \to \mathbb{R} \,|\, f(0) = 0, \mathbb{E}[|f(X_i)|^2] < \infty \text{ for all } i = 0, \ldots, p \right\}, \tag{4.3}$$

to handle integrability and identifiability issues (the latter introduced by adding and subtracting constants to functions). Provided $f \in \mathcal{C}'$, we know that f is non-constant if and only if $f \not\equiv 0$. This means that pa(0) is the unique minimal set satisfying the global invariant prediction, by arguing analogously to Theorem 4.4. We then define

$$H_{0,S}(\mathcal{E}) : \begin{cases} \text{there exist } \widetilde{f}_j \in \mathcal{C} \text{ with } f_j \equiv 0 \text{ for } j \notin S \text{ and } F_\epsilon \text{ s.t for} \\ \text{all } e \in \mathcal{E}, Y^e = \sum_{j=1}^p \widetilde{f}_j(X_j^e) + \epsilon^e \text{ where } \epsilon^e \perp\!\!\!\perp X_S^e,\ \epsilon^e \sim F_\epsilon \end{cases} \tag{4.4}$$

in a similar fashion to (2.7), and $S(\mathcal{E})$ analogously to (2.6).

One type of intervention we will consider is the do-intervention, as introduced in Section 1.1, allowing it to act on multiple variables $\mathcal{A} \subseteq \{1, \ldots, p\}$. We also consider *noise interventions* [Peters et al., 2016, Section 4.2] on variables in $\mathcal{A}$, which have the form

$$e \to \text{do}\left(X_j = \sum_{k=0}^{p} f_{j,k}(X_k) + \widetilde{\epsilon}_j \text{ for } j \in \mathcal{A} \right), \tag{4.5}$$

where either $\widetilde{\epsilon}_j = A_j^e \epsilon_j$ or $\epsilon_j + C_j^e$ for some random A_j^e or C_j^e, which can be constant almost surely. We also require that they are independent of each other across $j \in \mathcal{A}$, and each are generally independent of every other variable in the model except the $\{X_k \,|\, k \in \text{de}(j)\}$ for $j \in \mathcal{A}$. We now have the following result:

Theorem 4.5. *Suppose $\mathcal{S}$ is an additive structural equation model as in* (4.2) *which generates X. Further suppose that the noise distributions ϵ_i are such that $0 < Var(\epsilon_i) < \infty$ for $i = 0, \ldots, p$. Let $\mathcal{E}$ be a set of interventional settings, which includes $1 \in \mathcal{E}$ corresponding to performing no intervention at all. For $e \in \mathcal{E}$, denote the interventional system by a superscript e. Then:*

(i) Suppose $\mathcal{E}$ contains $e \in \mathcal{E}$ corresponding to a do intervention on $\mathcal{A}^e = \{j\}$ for all $j = 1, \ldots, p$. Then there exist do interventions such that $S(\mathcal{E}) = pa(0)$, in the sense that

$$\left\{ a \in \mathbb{R}^p \,:\, S(\mathcal{E}) = pa(0) \text{ when } \bigcup_{j=1}^{p} \{e_j \to do(X_j = a_j)\} \subseteq \mathcal{E} \right\} \neq \emptyset.$$

(ii) Suppose $\mathcal{E}$ contains $e \in \mathcal{E}$ corresponding to a noise intervention on $\mathcal{A}^e = \{j\}$ for all $j = 1, \ldots, p$. Then provided $\mathbb{E}\left[(A_j^e)^2\right] \neq 1$ or $\mathbb{E}\left[(C_j^e)^2\right] \neq 0$, depending on whether the intervention scales or shifts the noise respectively, for all $j = 1, \ldots, p$, we have that $S(\mathcal{E}) = pa(0)$.

Proof. We suppose that $S(\mathcal{E}) \neq \text{pa}(0)$ to deduce a contradiction. As $S(\mathcal{E}) \subseteq \text{pa}(0)$ by Theorem 4.1, this implies that there exists S such that $\text{pa}(Y) \not\subseteq S$ and $H_{0,S}(\mathcal{E})$ is true. In particular, this means that there exists functions $\widetilde{f}_j \in \mathcal{C}$, with $f_j \equiv 0$ for $j \notin S$ and $f_j \not\equiv 0$ and non-constant otherwise, such that

$$R^e(S) := Y^e - \sum_{j=1}^{p} \widetilde{f}_j(X_j^e) \overset{\text{d}}{=} R^f(S) \text{ for all } e, f \in \mathcal{E}.$$

By the defining equations of the structural equation model, we can write

$$R^e(S) = Y^e - \sum_{j=1}^{p} \widetilde{f}_j(X_j^e) = \sum_{j=1}^{p} g_j(X_j^e) + \epsilon_0^e$$

for some functions $g_j := f_{0,j} - \widetilde{f}_j \in \mathcal{C}'$. In particular, we note that there exists j such that $g_j \not\equiv 0$, and therefore non-constant also (as $g_j(0) = 0$).

Let $N = \{j \in \{1, \dots, p\} \mid g_j \not\equiv 0\}$, and select $k \in N$ such that it is not an ancestor of any $j \in N \setminus \{k\}$; we can find such a k as G is acyclic. Suppose there exists $e \in \mathcal{E}$ such that e is a noise or do-intervention on $\mathcal{A}^e = \{k\}$. In the case of this being a do-intervention, we have that

$$R^1(S) = g_k(X_k^1) + \sum_{j=0, j\neq k}^{p} g_j(X_j^1) + \epsilon_0^1, \quad R^e(S) \overset{\text{d}}{=} g_k(a_k^e) + \sum_{j=0, j\neq k}^{p} g_j(X_j^1) + \epsilon_0^1.$$

If these are equal in distribution, then $\mathbb{E}[g_k(X_k^1)] = g_k(a_k^e)$. This holds even when we vary a_k^e, so as g_k is non-constant, we can find $a_k^e \in \mathbb{R}$ which will give rise to a contradiction.

If we have a noise intervention, then by iteratively using the defining equations of the structural equation model, we get that

$$R^1(S) = \epsilon_k^1 + \widetilde{g}\left(\epsilon_0^1, \dots, \epsilon_{k-1}^1, \epsilon_{k+1}^1, \dots, \epsilon_p^1\right)$$

for some non-trivial function $\widetilde{g}$. Similarly we find that

$$\begin{aligned} R^e(S) &\overset{\text{d}}{=} A_k^e \epsilon_k^1 + \widetilde{g}\left(\epsilon_0^1, \dots, \epsilon_{k-1}^1, \epsilon_{k+1}^1, \dots, \epsilon_p^1\right) \text{ or} \\ &\overset{\text{d}}{=} C_k^e + \epsilon_k^1 + \widetilde{g}\left(\epsilon_0^1, \dots, \epsilon_{k-1}^1, \epsilon_{k+1}^1, \dots, \epsilon_p^1\right), \end{aligned}$$

depending on whether we scale or shift the noise distribution respectively. In the

former case, by the joint independence of the ϵ_i^1 and the independence of A_k^e from these, we see that

$$\mathbb{E}[R^1(S)^2] - \mathbb{E}[R^e(S)^2] = \left(1 - \mathbb{E}\left[(A_k^e)^2\right]\right) \mathbb{E}\left[(\epsilon_k^1)^2\right] \neq 0$$

provided $\mathbb{E}\left[(A_k^e)^2\right] \neq 1$, which gives rise to a contradiction. In the latter case of a shift, we may deduce in a similar fashion that $\mathbb{E}[R^1(S)^2] - \mathbb{E}[R^e(S)^2] = \mathbb{E}\left[(C_k^e)^2\right] \neq 0$, which again creates a contradiction. □

We note that, in practice, the possible values of a_k^e which allow for $S(\mathcal{E}) = \text{pa}(0)$ may be very broad. For example, if $\mathcal{C} = C(\mathbb{R}) \cap \mathcal{C}'$ and X_k^1 has a density with respect to Lebesgue measure, then $\mathbb{E}[g_k(X_k^1)] = g(c)$ for some $c \in \mathbb{R}$ by the integral version of the mean value theorem. In this case it is therefore enough for a_k^e to not lie in the same level set as c. If we specialise further to $\mathcal{C}$ being the intersection of $\mathcal{C}'$ with the set of polynomial functions, then all but finitely many $a_k^e \in \mathbb{R}$ would suffice.

4.4 Differences with invariant prediction

In the context of structural equation models, invariant prediction is strongly linked to the autonomy property. Despite this, they differ in philosophy about what causal information they are concerned about. In a structural equation model, we are interested about the causality structures between all random variables of interest, versed via the language of d-separation and conditional independence. In contrast, with invariant prediction we are only concerned with finding variables which have a "non-zero direct effect" on one particular variable of interest.

Recall that in a structural equation model, we say that a variable X_j has a total causal effect on X_k if there exists a distribution ϵ_j such that $X_j \not\perp\!\!\!\perp X_k$ under $X \,|\, \text{do}(X_j = \epsilon_j)$. We have already seen that, although it is necessary for j to be an ancestor of k, it is not sufficient even if it is a parent (e.g (1.5)). Supposing X_k is the variable of interest, we know that the unique minimal set satisfying the global invariant prediction assumption is the parent set of k. As these variables are the ones with a non-zero direct effect, this notion is therefore separate from that of variables with a total causal effect; they describe different aspects of causality. Invariant prediction should therefore be considered as a way of testing for the universality of the data generating process, for a particular variable of interest, under any intervention performed on variables other than itself.

Section 5

Simulation study of invariant prediction methods

Having detailed how we can carry out invariant prediction for generalized linear models in Section 2.5.2, we now investigate the coverage probability $\mathbb{P}(\widehat{S}(\mathcal{E}) \subseteq S^*)$ and equality probability $\mathbb{P}(\widehat{S}(\mathcal{E}) = S^*)$ for both the methods we proposed via simulation. For the former, although we can guarantee $\mathbb{P}(\widehat{S}(\mathcal{E}) \subseteq S^*) \geq 1-\alpha$ asymptotically, we want to see if we can expect similar coverage for finite samples. As we ideally want to completely identify S^*, we also want to know the probability that $\widehat{S}(\mathcal{E}) = S^*$ occurs, as we currently have no general theoretical results for such a quantity.

We investigate the behaviour of these two quantities both as the number of samples n_e per environment increase, and when we take successively increasing sets of environments $\mathcal{E}_1 \subset \mathcal{E}_2 \subset \ldots$ to examine. To do so, we perform invariant prediction on a binary variable in various simulated structural equation models. The form these models take, in addition to the interventions we perform on the model in order to try and infer S^*, is described in Section 5.1. We then display visually the results of these simulations, and discuss their implications, in Section 5.2.

5.1 Simulation settings

In order to study the coverage and equality probabilities, we look at the results of performing various interventions on 25 separate simulated structural equation models. For each model, we begin by simulating a directed acyclic graph with p vertices and average degree k, such that

(i) $p = 6$ or 7; $k = 1$ or 2 (13 models), or

(ii) $p = 8$ or 9; $k = 2$ or 3 (12 models).

The number of vertices and average degree are selected independently and randomly with equal probabilities. Labelling the variables as $Y = X_1, X_2, \ldots, X_p$, we redraw graphs until $S^* := \text{pa}(1) \neq \emptyset$, in order to allow us to investigate the coverage and equality probabilities when these are not necessarily equal.

To generate the directed acyclic graph, we use the method described by Peters et al. [2016, Appendix G]. Letting p be the number of nodes and k the average degree, to generate a directed acyclic graph with these properties, we begin by generating a random permutation π of $\{1, \ldots, p\}$. Treating this as a topological order, we then connect $j \to k$ with probability $k/(p-1)$ provided $\pi(j) < \pi(k)$. Calling the generated graph G, on average we would expect it to contain

$$\sum_{i=1}^{p-1} \frac{k}{p-1}(p-i) = \frac{k}{p-1} \cdot \frac{p(p-1)}{2} = \frac{kp}{2} \tag{5.1}$$

(directed) edges in total. Therefore, viewing G as an undirected graph, the average degree is then

$$\frac{2 \cdot (\text{average size of } |E|)}{|V|} = \frac{2}{p} \cdot \frac{kp}{2} = k \text{ as desired.} \tag{5.2}$$

Given a directed acyclic graph, we construct a structural equation model by setting for $i = 1, \ldots, p$

$$X_i = \begin{cases} 1\left[\frac{\exp\left(\eta_i + \sum_{j=1}^p \gamma_{i,j} X_j\right)}{1+\exp\left(\eta_i + \sum_{j=1}^p \gamma_{i,j} X_j\right)} > \epsilon_i\right], \text{ where } \epsilon_i \sim U[0,1] & \text{if } i = 1, \\ \eta_i + \sum_{j=1}^p \gamma_{i,j} X_j + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma_i^2) & \text{if } i \neq 1, \end{cases} \tag{5.3}$$

where $\gamma_{i,j} \neq 0$ if and only if $j \to i$, and the ϵ_i are jointly independent. The remaining parameters are then all independently simulated as follows:

(i) the non-zero $\gamma_{i,j}$ are assigned a random sign and magnitude uniformly from $\{0.5, 0.6, \ldots, 2\}$,

(ii) the σ_i^2 are randomly sampled uniformly from $\{0.7, 0.8, 0.9, 1.0, 1.1\}$, and

(iii) the η_i are randomly sampled uniformly from -0.5 to 0.5.

We use e_1 to denote the interventional or observational setting corresponding to sampling from the above model. We then consider the following interventional settings, which act on nodes only in $\{2, \ldots, p\}$:

- e_2 - for two nodes randomly selected in advance, we perform independent noise interventions (recall (4.5)) on each, where we shift ϵ_i by $A \sim U[0.5, 1.5]$;
- e_3 - for two nodes randomly selected in advance, we perform interventions of the form $\text{do}(X_i = \widetilde{\epsilon}_i)$ where $\widetilde{\epsilon}_i \sim N(0, 0.7)$;
- e_4 - for three nodes randomly selected in advance, we perform interventions of the form $\text{do}\,(X_i = \widetilde{\epsilon}_i)$ where $\widetilde{\epsilon}_i \sim N(-0.5, 0.8)$;
- e_5 - for three nodes randomly selected in advance, we perform interventions of the form

$$\text{do}\left(X_i = 0.5 + \eta_i + \sum_{j:\pi(j)<\pi(i)} (\gamma_{i,j} + A_{i,j})X_j\right)$$

  where the $A_{i,j}$ are independent and identically distributed $U[0.3, 0.8]$ random variables. We note that this is a bona-fide intervention, as no cycles are introduced by adding variables j with $\pi(j) < \pi(i)$ to the parent set of i; if j were an ancestor of i, we would have that $\pi(i) < \pi(j)$ which is absurd.

We then form $\mathcal{E}_j = \{e_i : i \leq j\}$, so $\mathcal{E}_2 \subset \ldots \subset \mathcal{E}_5$. We add interventions in increasing order of the severity of changes to the model they make, in order to try and examine the strength of interventions required to identify $S^* = \text{pa}(1)$ (the minimal unique set satisfying the global invariant prediction assumption), as we cannot use e.g Theorem 4.5 for the structural equation model constructed here.

For each structural equation model and framework of interventions (which we now refer to as a *scenario*), we attempt to identify $S^* = \text{pa}(1)$ using the two different tests for $H_{0,S}(\mathcal{E})$ provided in Section 2.5.2. To do so, we take n_e samples from each environment $e_1, \ldots, e_5$, and then form $\widehat{S}(\mathcal{E}_j)^A$ and $\widehat{S}(\mathcal{E}_j)^B$ for each $j = 2, \ldots, 5$, constructing $\widehat{S}(\mathcal{E}_j)^{A/B}$ by using the F tests[1] (2.25) or (2.27) respectively to test $H_{0,S}(\mathcal{E})$ to a size 0.1 . We form these estimators for sample sizes $n_e = 100, 200, \ldots, 1000$, and count whether $\widehat{S}(\mathcal{E}_j)^{A/B}$ is equal to pa(1), a subset of it, or neither. We then repeat this process 100 times[2] per scenario, in order to give estimates of the coverage and equality probabilities for various $\mathcal{E}_j$ and sample sizes n_e.

[1] Although this is not necessary as the dispersion parameter is known, we found that these versions give better equality properties for small samples than the corresponding estimators using the χ^2 versions (2.24) or (2.26) for testing $H_{0,S}(\mathcal{E})$.

[2] This number was chosen as it was sufficiently low to allow for simulations to be carried out in a reasonable time, yet high enough so that the standard error in any coverage or equality probability estimate is at most 0.05, which will suffice for our purposes of investigating (mostly) qualitative behaviour.

5.2 Results and discussion

The results of the simulations described in Section 5.1 are displayed in Figures 5.1 and 5.2. The former displays estimates for the coverage and equality probabilities, $\mathbb{P}(\widehat{S}(\mathcal{E}_j)^{A/B} \subseteq S^*)$ and $\mathbb{P}(\widehat{S}(\mathcal{E}_j)^{A/B} = S^*)$, for fixed values of n_e, examining how they vary as the number of environmental settings used increases. In contrast, the latter displays information about how the equality probabilities vary with the total sample size across all the different environmental settings and keeping the $\mathcal{E}_j$ fixed.

We begin by mentioning that the estimated coverage probability, under any combination of scenario, testing method, sample size and number of environments, is consistent with $\mathbb{P}(\widehat{S}(\mathcal{E}_j)^{A/B} \subseteq S^*) \geq 0.9$, as observed in Figures 5.1a, 5.1c and 5.1e. Recalling that the standard error of each estimate is no greater than 0.05, we see that nearly all the probability estimates are either above 0.9, with the remainder being within one standard error of this threshold. Furthermore, there is no apparent change in behaviour when using either testing method.

Asymptotically, this behaviour is to be expected by analogy to Theorem 2.2, because both tests have asymptotic size 0.1. In contrast, for small sample sizes this likely occurs due to both tests for $H_{0,S}(\mathcal{E})$ having small power, meaning a large number of $H_{0,S}(\mathcal{E})$ will be (incorrectly) not rejected. Consequently, it is likely that $\widehat{S}(\mathcal{E}) = \emptyset \subseteq S^*$, regardless of whether $H_{0,S^*}(\mathcal{E})$ is rejected or not.

The behaviour of the equality probability, like that of the coverage probability, is governed by the size and power of our tests for $H_{0,S}(\mathcal{E})$. However, the former is also governed by whether we have enough environments to identify S^*, unlike the latter. In order to obtain a high equality probability, we require

(i) the sample size to be sufficiently large for $H_{0,S}(\mathcal{E})$ to be powerful enough such that $\widehat{S}(\mathcal{E}) = S(\mathcal{E})$ with high probability, and then

(ii) $\mathcal{E}$ to be such that $S(\mathcal{E}) = S^*$.

As for any fixed size α, the tests we propose for $H_{0,S}(\mathcal{E})$ have asymptotic power one (under mild regularity conditions), the first item should not be an issue. This is confirmed by Figure 5.2, where the average trend lines for each fixed $\mathcal{E}_j$ show an improvement in the equality probability with total sample size. Examining some of the individual paths suggests that the equality probabilities can approach one even for relatively small sample sizes (less than 500 samples in total).

The trend lines in Figure 5.2 also show that, when *averaging across all scenarios*, the equality probability increases alongside the $\mathcal{E}_j$, even when the sample size is fixed. There are two potential (not necessarily mutually exclusive) explanations for this behaviour:

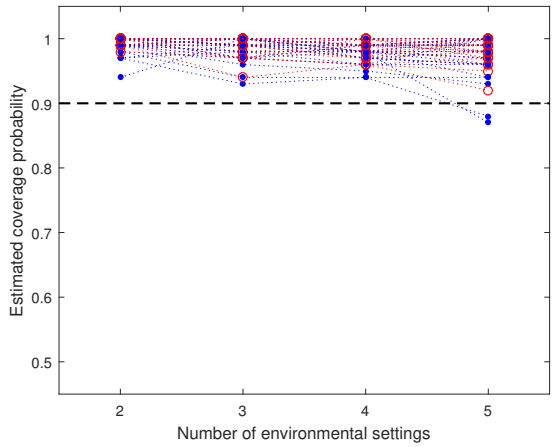
(i) *Identifiability of* S^* - As $\mathcal{E}_j$ increases, we have $\mathcal{E}_j \subset \mathcal{E}_{j+1}$ such that $S(\mathcal{E}_j) \subset S^*$ yet $S(\mathcal{E}_{j+1}) = S^*$, so only with $\mathcal{E}_{j+1}$ can we now identify S^*.

(ii) *Power of* $H_{0,S}(\mathcal{E})$ *increases with* $\mathcal{E}$ - Even if we have $\mathcal{E}_j \subset \mathcal{E}_{j+1}$ such that both $S(\mathcal{E}_j) = S(\mathcal{E}_{j+1}) = S^*$, a test for $H_{0,S}(\mathcal{E}_{j+1})$ is more powerful (when the total sample size is fixed) than that of $H_{0,S}(\mathcal{E}_j)$ for information theoretic reasons.

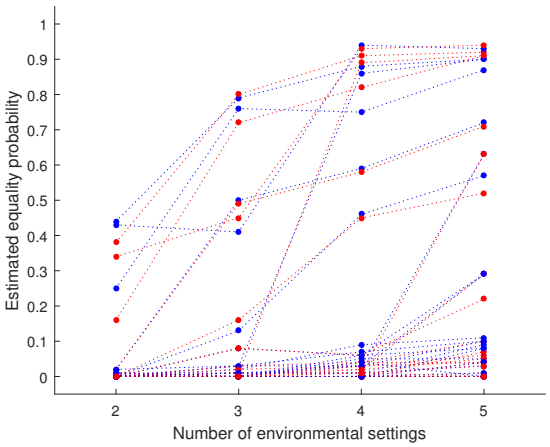
Figure 5.3 suggests that both can occur. For the first, we see in Figure 5.3a that the estimated equality probability is zero, regardless of the sample size, and so it is likely that $S(\mathcal{E}_2) \neq S^*$. However, for subsequently larger $\mathcal{E}_j$ the equality probability becomes non-zero, suggesting that now $S(\mathcal{E}_j) = S^*$. For the second item, we now look to Figure 5.3b, where for the same total sample size we say that the equality probability under $\mathcal{E}_3$ is far greater than that under $\mathcal{E}_2$, although both are non-zero. This therefore suggests that the tests for $H_{0,S}(\mathcal{E}_j)$

Investigating which $\mathcal{E}_j$ identify S^* in more depth, Figures 5.1b, 5.1d and 5.1f show that for each $\mathcal{E}_j$, there are scenarios in which these environments allow S^* to be identified identify, and scenarios for which none of the $\mathcal{E}_j$ appear to be able to identify S^*. Without an analogue to Theorem 4.5, it is hard to identify exactly why this occurs. However, as even the action of noise interactions on two nodes is enough to allow for identifiability in some models, this suggests that Theorem 4.5 may be generalized to the structural equation models we use here, although we do not pursue this any further.

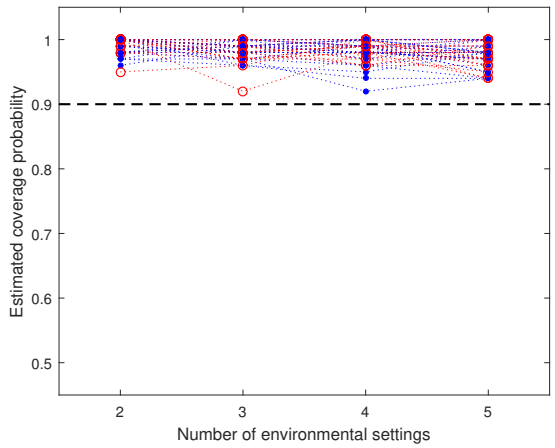
We end by noting that, as with the coverage probabilities, the choice of method for testing $H_{0,S}(\mathcal{E})$ has no discernible impact on the equality probability. This may be a consequence of the particular choice of interventions used, or simplify that the two methods have similar behaviours. In any case, we therefore prefer to use $\widehat{S}(\mathcal{E}_j)^A$ over $\widehat{S}(\mathcal{E}_j)^B$ for the interventions used here; this is because in practice we found it to quicker computationally.



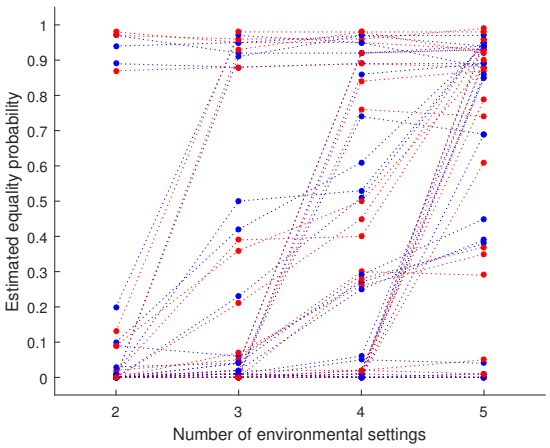
(a) Coverage probability when $n_e = 100$



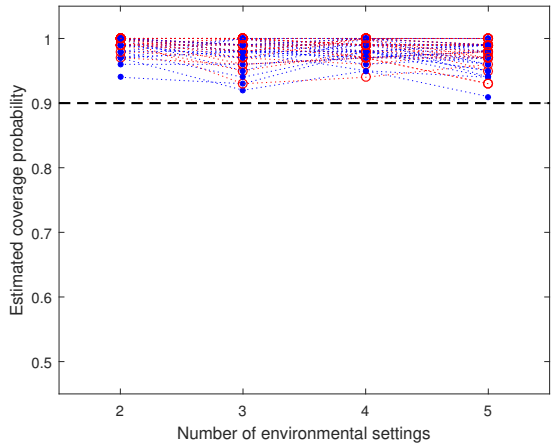
(b) Equality probability when $n_e = 100$



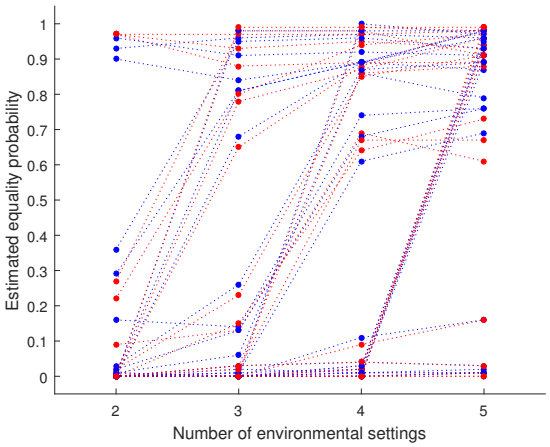
(c) Coverage probability when $n_e = 500$



(d) Equality probability when $n_e = 500$

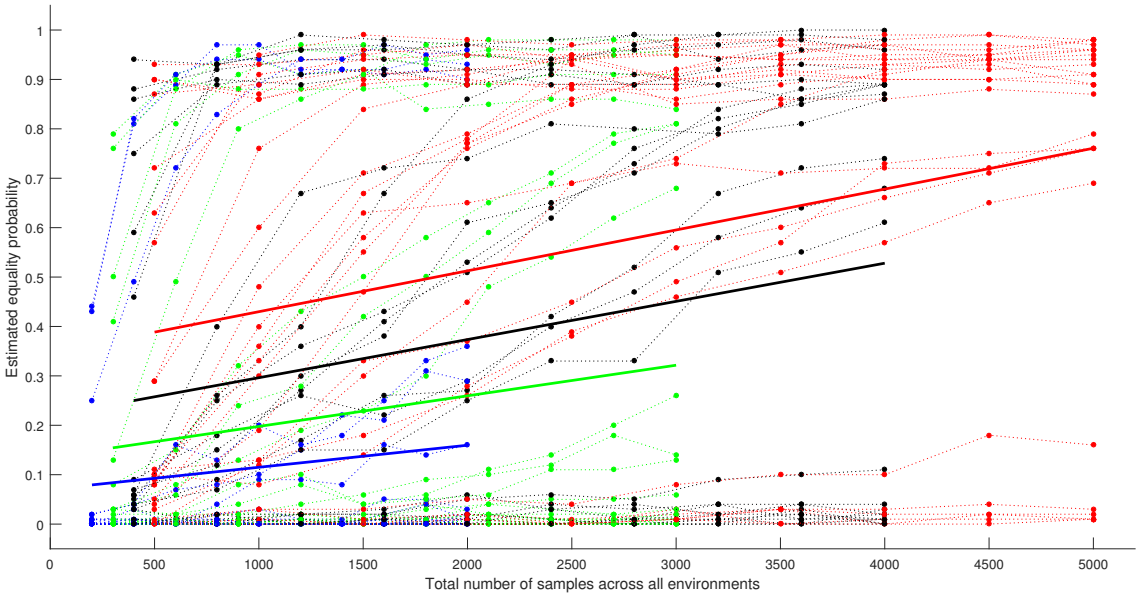


(e) Coverage probability when $n_e = 1000$

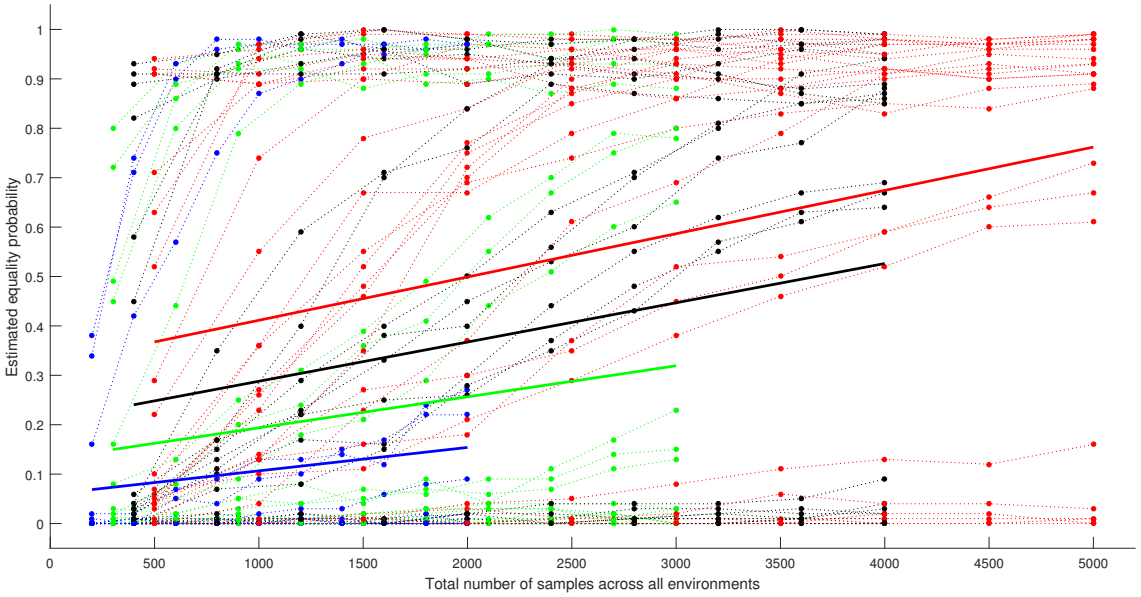


(f) Equality probability when $n_e = 1000$

Figure 5.1: Graphs illustrating the estimated coverage and equality probabilities (from 100 simulations per scenario, 25 scenarios in total) of $\widehat{S}(\mathcal{E}_j)^A$ (blue dots) and $\widehat{S}(\mathcal{E}_j)^B$ (red dots), for different sizes of n_e and $|\mathcal{E}_j|$. Observations common to the same scenario are linked by dotted lines.

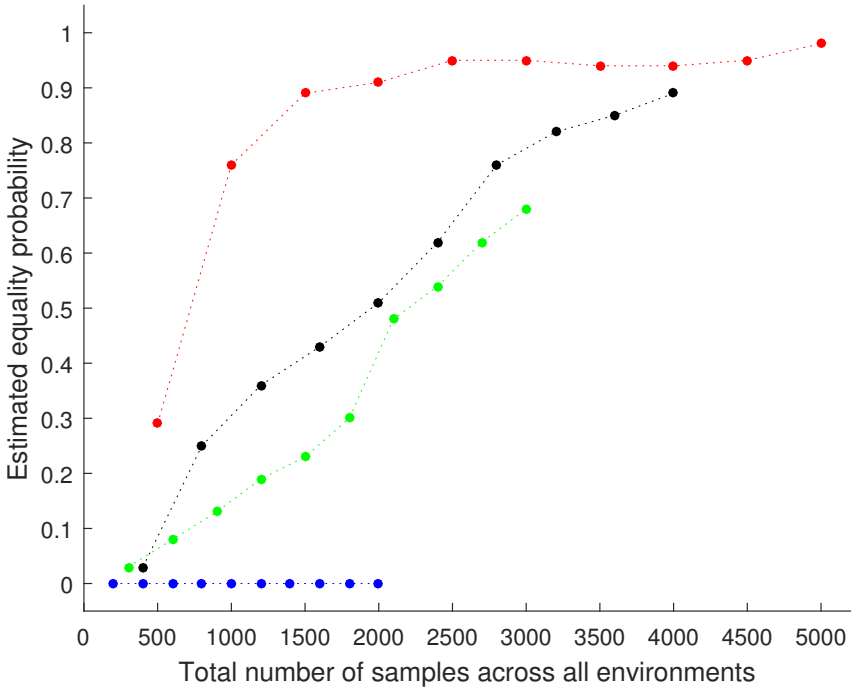


(a) Estimated equality probabilities for $\widehat{S}(\mathcal{E}_j)^A$

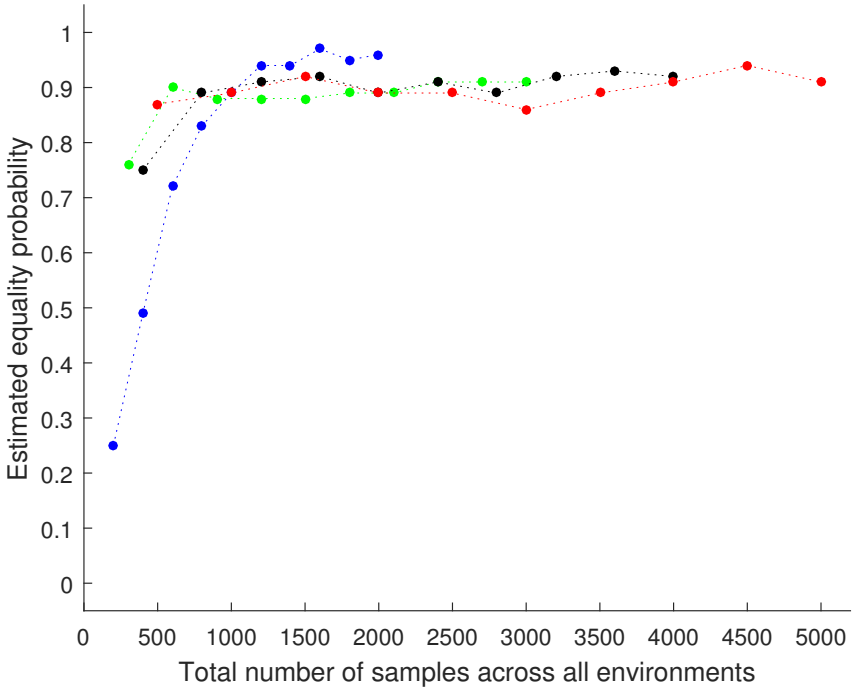


(b) Estimated equality probabilities for $\widehat{S}(\mathcal{E}_j)^B$

Figure 5.2: Graphs plotting the estimated equality probabilities (from 100 simulations per scenario, 25 scenarios in total) for $\widehat{S}(\mathcal{E}_j)^A$ and $\widehat{S}(\mathcal{E}_j)^B$ against the total sample size across all environments used, for $\mathcal{E}_2, \ldots, \mathcal{E}_5$ in blue, green, black and red respectively. The average linear trend, for each $\mathcal{E}_j$, of the equality probability against the total sample size is plotted using the same colour as in the individual case. Observations common to the same scenario and set of environmental settings $\mathcal{E}_j$ are linked by dotted lines.



(a) Estimated equality probabilities for $\widehat{S}(\mathcal{E}_j)^A$ under (S1)



(b) Estimated equality probabilities for $\widehat{S}(\mathcal{E}_j)^A$ under (S2)

Figure 5.3: Graphs plotting the estimated equality probabilities (from 100 simulations per scenario) for $\widehat{S}(\mathcal{E}_j)^A$ from two separate scenarios - labelled (S1) and (S2) - against the total sample size across all environments used, for $\mathcal{E}_2, \ldots, \mathcal{E}_5$ in blue, green, black and red respectively. Observations common to the same set of environmental settings $\mathcal{E}_j$ are linked by dotted lines.

Bibliography

Bela Bollobas. *Modern Graph Theory (Graduate Texts in Mathematics)*. Springer, 2002. ISBN 0387984887.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 0521833787.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications (Springer Series in Statistics)*. Springer-Verlag, 2011. ISBN 3642201911.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.*, 2:1153–1194, 2008. ISSN 1935-7524. doi: 10.1214/08-EJS287.

Sourav Chatterjee. Assumptionless consistency of the Lasso. 2013. arXiv:1303.5817 [math.ST].

Gregory C Chow. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3):591–605, 1960. ISSN 00129682, 14680262. doi: 10.2307/1910133.

Frederick Eberhardt and Richard Scheines. Interventions and Causal Inference. In *Philos. Sci.*, volume 74, pages 981–995, 2007.

William W Hager. Updating the Inverse of a Matrix. *SIAM Rev.*, 31(2):221–239, 1989. ISSN 00361445.

Bent Jørgensen. Exponential Dispersion Models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 49(2):127–162, 1987.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1st edition, 1996. ISBN 978-0198522195.

Tim Futing Liao. *Statistical Group Comparison*. Wiley, 2002. ISBN 0471386464.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009. ISBN 0-521-77362-8.

Jonas Peters. Lecture notes on "Causality", 2015. URL `http://people.tuebingen.mpg.de/jpeters/scriptChapter1-4.pdf`. (Last accessed 08/04/2016).

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Ser. B Stat. Methodol. (to appear)*, 2016.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, Second Edition (Adaptive Computation and Machine Learning)*. A Bradford Book, 2001. ISBN 0262194406.

Robert Tibshirani. Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. B*, 58(1):267–288, 1994. ISSN 00359246. doi: 10.2307/2346178.

Appendices

1 Theory of directed acyclic graphs

Although most of the graph theoretic terminology we use is elementary (see e.g Bollobas [2002] for a review), the use of directed acyclic graphs and their basic properties may be unfamiliar and so we give a brief overview.

Beginning with some useful definitions, a *directed acyclic graph* is as its namesake suggests - a directed graph which contains no (directed) cycles. We now let G be a directed acyclic graph. For a path $j = j_1, j_2, \ldots, j_m = k$ in G, so j_i and j_{i+1} are adjacent but not necessarily such that $j_i \to j_{i+1}$, we say that j_l is a *collider* (relative to the path) if $j_{l-1} \to j_l \leftarrow j_{l+1}$. We then say that such a path is *blocked by a set* S where $j_1, j_m \notin S$ if there exists j_l such that either

(i) $j_l \in S$ yet j_l is not a collider, or

(ii) j_l is a collider yet $(\{j_l\} \cup \mathrm{de}(j_l)) \cap S = \emptyset$, i.e j_l and none of its descendants are contained in S.

Given disjoint subsets A, B, S of the edge set E, we then say that A and B are *d-separated* by S if every path from A to B is blocked by the set S.

We now denote the vertex set of G by $V = \{1, \ldots, p\}$. The following result now guarantees the existence of a *topological ordering*, which is a permutation π of V such that $\pi(j) < \pi(k)$ whenever k is a descendent of j. This is a useful way of storing information about the dependence structure between nodes in the graph G.

Proposition A.1. *There exists a topological ordering π on G.*

Proof. We prove this by induction on the number of vertices p. If $p = 1$, we are done immediately. If $p > 1$, we first argue that there exists a source node (a node with no parents). Indeed, if we pick a node (say j) in the graph, then either $\mathrm{pa}(j)$ is empty, in which case we're done, or it is not empty. In the latter case, we can then visit a parent node and repeat this process. As G is acyclic, we can never revisit a

previous node; as G is finite, this process will eventually terminate.

Now let k be a source node, and G' the subgraph induced by removing the node k from G. Then as G' is also a directed acyclic graph, by the induction hypothesis, there exists a topological ordering π on G'. Writing π as a permutation of $\{1, \ldots, p\} \setminus \{k\}$, we can then define a topological ordering on G by:

$$\widetilde{\pi}(j) := \begin{cases} 1 & \text{if } j = k \\ \pi(j) + 1 & \text{if } \pi(j) < k \\ \pi(j) & \text{if } \pi(j) > k. \end{cases} \quad \square$$

2 Review of exponential dispersion families

We now give a brief review of exponential dispersion families, mostly for the sake of establishing notation and stating a few basic properties.

Let $\mathcal{P} = \{P_{\theta,\sigma} \,|\, \theta \in \Theta \subseteq \mathbb{R}, \sigma \in \Phi \subseteq (0, \infty)\}$ be a family of distributions indexed by θ and σ. We say that $\mathcal{P}$ is an *exponential dispersion family* if the $P_{\theta,\sigma}$ are absolutely continuous with respect to some measure (say ν) with density

$$f(y; \theta, \sigma) = a(y, \sigma) \exp\left(\frac{1}{\sigma}(y\theta - K(\theta))\right) \text{ for } y \in \mathcal{Y} \subseteq \mathbb{R}, \tag{A.4}$$

provided that the densities are non-degenerate, $a(y, \sigma)$ is a known positive function, and Θ is an open interval. We call σ the *dispersion parameter*, which can be either known or unknown.

If $Y \sim P_{\theta,\sigma}$, then it is well known that

$$\mu := \mathbb{E}_{\theta,\sigma}[Y] = K'(\theta), \qquad \mathrm{Var}_{\theta,\sigma}(Y) = \sigma K''(\theta), \tag{A.5}$$

and so as Y is non-degenerate, K' is invertible. We use this to give a bijection between θ (the *natural parameter*) and μ (the *mean parameter*), which in an abuse of notation is usually denoted via $\theta = \theta(\mu)$ and $\mu = \mu(\theta)$. We then define the *mean space* $\mathcal{M} := \{\mu(\theta) \,|\, \theta \in \Theta\}$ and *variance function* $V : \mathcal{M} \to (0, \infty)$ given by $V(\mu) = K''(\theta(\mu))$. The notation $Y \sim \mathrm{ED}(\mu, \phi)$ is used to denote the distribution of Y in terms of μ and ϕ, i.e we write $Y \sim \mathrm{ED}(\mu, \phi)$ if $Y \sim P_{\theta(\mu),\phi}$.

3 Properties of the invariant Lasso

We now prove some properties of the invariant Lasso which were stated in Section 3.2. We first show that there exists a solution to (3.6). Denote $Q_\lambda(\beta)$ for the objective function and let $C := \sum_{e\in\mathcal{E}} \|Y_e\|_2^2 / 2n_e$. Then as

$$\inf_{\beta\in\mathbb{R}^p\,:\,\lambda\|\beta\|_1\leq C} Q_\lambda(\beta) \leq Q_\lambda(0) = C < \inf_{\beta\in\mathbb{R}^p\,:\,\lambda\|\beta\|_1> C} Q_\lambda(\beta), \tag{A.6}$$

it suffices to minimize the continuous function $Q_\lambda(\beta)$ over the closed and bounded (thus compact) set $\{\beta \in \mathbb{R}^p \mid \|\beta\|_1 \leq C/\lambda\}$, and so we know a solution exists.

Although this solution may not be unique, we can show that both the fitted values $\mathbf{X}_e\widehat{\beta}_\lambda$ for each $e \in \mathcal{E}$ and $||\widehat{\beta}_\lambda||_1$ are unique. Suppose that β_1 and β_2 are both solutions to (3.6), and let $t \in (0,1)$. Then as $\|\cdot\|_2^2$ is strictly convex,

$$\begin{aligned}\sum_{e\in\mathcal{E}} \frac{1}{n_e} ||Y_e - \mathbf{X}_e(t\beta_1 + (1-t)\beta_2)||_2^2 \\ \leq t\sum_{e\in\mathcal{E}} \frac{1}{n_e} ||Y_e - \mathbf{X}_e\beta_1||_2^2 + (1-t)\sum_{e\in\mathcal{E}} \frac{1}{n_e} ||Y_e - \mathbf{X}_e\beta_2||_2^2\end{aligned} \tag{A.7}$$

with equality if and only if $\mathbf{X}_e\beta_1 = \mathbf{X}_e\beta_2$ for all $e \in \mathcal{E}$. As $\|\cdot\|_1$ is convex, we have

$$\|t\beta_1 + (1-t)\beta_2\|_1 \leq t\|\beta_1\|_1 + (1-t)\|\beta_2\|_1. \tag{A.8}$$

Adding these together gives $Q_\lambda(t\beta_1 + (1-t)\beta_2) \leq tQ_\lambda(\beta_1) + (1-t)Q_\lambda(\beta_2)$, which is an equality by convexity of the problem. Therefore both (A.7) and (A.8) are equalities, so the fitted values are unique for each $e \in \mathcal{E}$. As $Q_\lambda(\beta)$ is fixed across solutions to (3.6), this then implies the uniqueness of $||\widehat{\beta}_\lambda||_1$.

4 Proof of Lemma 3.3

We first consider the case when $j \in S^*$, although we only require that $\epsilon \perp\!\!\!\perp X_j^e$ for all $e \in \mathcal{E}$. Let $\mathcal{G}$ be the sigma-algebra generated by the independent samples $(x_i)_{i=1,\ldots,n}$, and denote $\mathbb{P}^{\mathcal{G}}(A) := \mathbb{E}[1_A|\mathcal{G}]$. Now, conditional on $\mathcal{G}$

$$Z_j = \sum_{e\in\mathcal{E}} \frac{1}{n_e} \sum_{i\in I_e} x_{ij}\epsilon_i \sim N\left(0, \sigma^2 \sum_{e\in\mathcal{E}} \sum_{i\in I_e} \frac{x_{ij}^2}{n_e^2}\right),$$

both as the $x_{ij}\epsilon_i$ are independent across $i = 1, \ldots, n$, and as x_{ij} is independent of ϵ_i for $i = 1, \ldots, n$ because $j \in S^*$. Therefore we obtain that

$$\mathbb{P}^{\mathcal{G}}\left(|Z_j| > t\right) \leq \exp\left(\frac{-t^2}{2\sigma^2}\left(\sum_{e\in\mathcal{E}}\sum_{i\in I_e}\frac{x_{ij}^2}{n_e^2}\right)^{-1}\right),$$

by using the tail bound $\mathbb{P}(Z > t) \leq \exp(-t^2/2\sigma^2)/2$ when $Z \sim N(0, \sigma^2)$. Then as the $|X_{i,j}| \leq M$ almost surely and $\sum_{e\in\mathcal{E}} n_e^{-1} \leq |\mathcal{E}|/n_{\min}$, we get that

$$\mathbb{P}^{\mathcal{G}}\left(|Z_j| > t\right) \leq \exp\left(-\frac{t^2 n_{\min}}{2|\mathcal{E}|M^2\sigma^2}\right).$$

holds on an event of probability one. Taking expectations then gives the result as the right hand side is non-random.

Now suppose that $j \notin S^*$ and $\epsilon \not\perp\!\!\!\perp X_j^e$ for some $e \in \mathcal{E}$. Despite the lack of independence, we will produce a tail bound which is the same as the bound in the $j \in S^*$ case, up to multiplication by a (potentially rapidly decaying) function depending on t. We proceed by producing a Chernoff bound (see e.g Bühlmann and van de Geer [2011, Chapter 14]) on Z_j. Letting $s \geq 0$, by applying Markov's inequality to $\mathbb{P}\left(e^{sZ_j} > e^{st}\right)$, we get

$$\begin{aligned}
\mathbb{P}(Z_j > t) &\leq e^{-st}\mathbb{E}\left[e^{sZ_j}\right] = \mathbb{E}\left[\exp\left(\sum_{e\in\mathcal{E}}\sum_{i\in I_e}\frac{1}{n_e}x_{ij}\epsilon_i\right)\right] \\
&= e^{-st}\prod_{e\in\mathcal{E}}\prod_{i\in I_e}\mathbb{E}\left[e^{sx_{ij}\epsilon_i/n_e}\right] \quad \text{(as the } x_{ij}\epsilon_i \text{ are independent)} \\
&\leq e^{-st}\prod_{e\in\mathcal{E}}\prod_{i\in I_e}\mathbb{E}\left[e^{sM|\epsilon_i|/n_e}\right] \quad \text{(as } x_{ij}\epsilon_i \leq M|\epsilon_i|\text{)}.
\end{aligned}$$

To proceed any further, we need to obtain the moment generating function of $|Y|$ where $Y \sim N(0, \sigma^2)$:

$$\begin{aligned}
\mathbb{E}\left[e^{s|Y|}\right] &= \int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{s|y|}e^{-y^2/2\sigma^2}\,\mathrm{d}y = 2\int_0^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{sy}e^{-y^2/2\sigma^2}\,\mathrm{d}y \quad \text{(by symmetry)} \\
&= 2e^{s^2\sigma^2/2}\int_0^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-(y-s\sigma^2)^2/2\sigma^2}\,\mathrm{d}y \quad \text{(by completing the square)} \\
&= 2e^{s^2\sigma^2/2}\int_{-s\sigma}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,\mathrm{d}y \quad \text{(by substituting } y' = (y - s\sigma^2)/\sigma\text{)} \\
&= e^{s^2\sigma^2/2}\left(1 + 2\int_{s\sigma}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,\mathrm{d}y\right).
\end{aligned}$$

Using the upper tail bound as before for a standard Normal distribution, we can then obtain the upper bound

$$\mathbb{E}\left[e^{s|Y|}\right] \leq e^{s^2\sigma^2/2}(1 + e^{-s^2\sigma^2/2}).$$

Returning back to our bound on $\mathbb{P}(Z_j > t)$, this therefore implies that

$$\begin{aligned}
\mathbb{P}(Z_j > t) &\leq e^{-st} \prod_{e\in\mathcal{E}} \prod_{i\in I_e} \left\{ e^{s^2M^2\sigma^2/2n_e^2}(1 + e^{-s^2\sigma^2M^2/2n_e^2}) \right\} \\
&= e^{-st} \prod_{e\in\mathcal{E}} \left\{ e^{s^2M^2\sigma^2/2n_e^2}(1 + e^{-s^2\sigma^2M^2/2n_e^2}) \right\}^{n_e} \\
&= \exp\left(-st + \frac{s^2M^2\sigma^2}{2} \sum_{e\in\mathcal{E}} \frac{1}{n_e} \right) \prod_{e\in\mathcal{E}} \left(1 + e^{-s^2\sigma^2M^2/2n_e^2}) \right)^{n_e}.
\end{aligned}$$

Optimizing this over s is likely intractable algebraically. However, as we would expect the $\prod_{e\in\mathcal{E}}(\cdots)$ term to be small for sufficiently large s, and s should scale with t which we would like to be large, we simply choose s which optimizes the left hand side of the product. We therefore choose

$$s = \frac{t}{M^2\sigma^2 \sum_{e\in\mathcal{E}} n_e^{-1}}$$

to obtain the bound, where we write $n_{\mathcal{E}} := \sum_{e\in\mathcal{E}}, n_e^{-1}$

$$\begin{aligned}
\mathbb{P}(Z_j > t) &\leq \exp\left(\frac{-t^2}{2M^2\sigma^2 n_{\mathcal{E}}} \right) \prod_{e\in\mathcal{E}} \left(1 + \exp\left(\frac{-t^2}{2M^2\sigma^2 n_e^2 n_{\mathcal{E}}^2} \right) \right)^{n_e} \\
&\leq \exp\left(\frac{-t^2 n_{\min}}{2|\mathcal{E}|M^2\sigma^2} \right) \left\{ 1 + \exp\left(\frac{-t^2 n_{\min}^2}{2|\mathcal{E}|^2 M^2\sigma^2 n_{\max}^2} \right) \right\}^n
\end{aligned}$$

where to obtain the last bound we have used that $n_{\mathcal{E}} \leq |\mathcal{E}|/n_{\min}$ and $n_e n_{\mathcal{E}} \leq |\mathcal{E}| n_{\max}/n_{\min}$. As the same argument follows through for $-Z_j$, we therefore obtain

$$\mathbb{P}(|Z_j| > t) \leq 2\exp\left(\frac{-t^2 n_{\min}}{2|\mathcal{E}|M^2\sigma^2} \right) \left\{ 1 + \exp\left(\frac{-t^2 n_{\min}^2}{2|\mathcal{E}|^2 M^2\sigma^2 n_{\max}^2} \right) \right\}^n.$$

5 Invariant prediction when X^e is degenerate

Here we briefly discuss how invariant prediction can possibly be performed when an X^e is degenerate for some $e \in \mathcal{E}$. This is problematic, as it is necessary for X^e to

be non-degenerate in order for the population regression coefficients $\widehat{\beta}^{\text{pred},e}(S)$ and $\widehat{\zeta}^{\text{pred},e}(S)$ to be unique. It also may mean that the design matrices $\mathbf{X}_e$ do not have full (column) rank. Moreover, degenerate X^e can arise naturally, such as when a do intervention is performed on some variables within a structural equation model. As highlighted by Theorem 4.5, such interventions may be useful for identifiability purposes, making this an issue of practical relevance.

One method of circumventing this is to pool some of the settings together to give a non-degenerate X^e. However, this adversely effects the interpretation of our results and could give a smaller set of plausible causal variables. Alternatively, suppose we know the structure of $\left(\widehat{\beta}^{\text{pred},e}(S), \widehat{\zeta}^{\text{pred},e}(S)\right)$ when X^e is degenerate, and that X^f is non-degenerate for $f \neq e$. Then to test whether $H_{0,S}(\mathcal{E})$ is true, we can first test whether the unique population regression coefficients across $\mathcal{E} \setminus \{e\}$ are identical, before using this information to allow the regression coefficients to be identifiable under $e \in \mathcal{E}$.

We illustrate this idea with a basic example. Let $X = (Y = X_0, X_1, \ldots, X_p)$ be generated by a structural equation model, with Y depending linearly on some of the X_i as in (2.2). Let $\mathcal{E} = \mathcal{E}' \cup \{f\}$, where f corresponds to $\text{do}(X_p = a)$ for some $a \in \mathbb{R}$. Otherwise, we suppose that X^e is non-degenerate for all $e \in \mathcal{E}'$, and that X^f_{-p} is also non-degenerate. Given this, it follows that there exists $\widetilde{\beta} \in \mathbb{R}^{p-1}$ and a constant c such that

$$\left(\widehat{\beta}^{\text{pred},f}(S), \widehat{\zeta}^{\text{pred},f}(S)\right) = \{(\beta, \zeta) \in \mathbb{R}^p \times \mathbb{R} \,|\, \beta_{-p} = \widetilde{\beta},\, \zeta + a\beta_p = c\}. \tag{A.9}$$

We know that otherwise $\left(\widehat{\beta}^{\text{pred},e}(S), \widehat{\zeta}^{\text{pred},e}(S)\right)$ is unique for all $e \in \mathcal{E}'$. Now, if $H_{0,S}(\mathcal{E})$ is true, we know that $\widehat{\beta}^{\text{pred},e}(S)_{-p} = \widetilde{\beta}$, $\widehat{\zeta}^{\text{pred},e}(S) = \widehat{\zeta}^{\text{pred},f}(S)$ and thus $\widehat{\zeta}^{\text{pred},e}(S) + a\widehat{\beta}^{\text{pred},f}(S)_p = c$. This allows us to uniquely determine the population regression coefficients under $f \in \mathcal{E}$.

We can then use the following testing procedure to incorporate information about $f \in \mathcal{E}$, which has size α in the sense of Theorem 2.2:

Step 1: First test whether $H_{0,S}(\mathcal{E}')$ is true to a size $\alpha|\mathcal{E}'|/|\mathcal{E}|$, using e.g the method given in Section 2.5.1.

Step 2: If this is true, we subtract the estimated intercept term from all the y_i. We then perform a Chow test to size $\alpha/|\mathcal{E}|$ for the equality of coefficients between the settings $f \in \mathcal{E}$ and $\mathcal{E}' \subset \mathcal{E}$ without including the intercept term in the model.