The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1220–1246
https://doi.org/10.1214/21-AOAS1487
© Institute of Mathematical Statistics, 2021

STABILIZING VARIABLE SELECTION AND REGRESSION

BY NIKLAS PFISTER[1,*], EVAN G. WILLIAMS[2], JONAS PETERS[1,†], RUEDI AEBERSOLD[3] AND PETER BÜHLMANN[4]

[1]*Department of Mathematical Sciences, University of Copenhagen,* [*]*np@math.ku.dk;* [†]*jonas.peters@math.ku.dk*
[2]*Luxembourg Centre for Systems Biomedicine, University of Luxembourg, evan.williams@uni.lu*
[3]*Institute of Molecular Systems Biology, ETH Zürich, aebersold@imsb.biol.ethz.ch*
[4]*Seminar for Statistics, ETH Zürich, buehlmann@stat.math.ethz.ch*

We consider regression in which one predicts a response Y with a set of predictors X across different experiments or environments. This is a common setup in many data-driven scientific fields, and we argue that statistical inference can benefit from an analysis that takes into account the distributional changes across environments. In particular, it is useful to distinguish between stable and unstable predictors, that is, predictors which have a fixed or a changing functional dependence on the response, respectively. We introduce stabilized regression which explicitly enforces stability and thus improves generalization performance to previously unseen environments. Our work is motivated by an application in systems biology. Using multiomic data, we demonstrate how hypothesis generation about gene function can benefit from stabilized regression. We believe that a similar line of arguments for exploiting heterogeneity in data can be powerful for many other applications as well. We draw a theoretical connection between multi-environment regression and causal models which allows to graphically characterize stable vs. unstable functional dependence on the response. Formally, we introduce the notion of a stable blanket which is a subset of the predictors that lies between the direct causal predictors and the Markov blanket. We prove that this set is optimal in the sense that a regression based on these predictors minimizes the mean squared prediction error, given that the resulting regression generalizes to unseen new environments.

1. Introduction. Statistical models usually describe the observational distribution of a data generating process. In many applied problems this data generating process may change over time or across experiments. In such settings it is useful to get a mechanistic understanding of the underlying changes in the system, both to understand which parts of a system cause certain outcomes and to make reliable predictions under previously unseen conditions. One approach to rigorously model such changes are causal models (Pearl (2009), Imbens and Rubin (2015)) which allow for changes in the data generating process via the notion of interventions. As demonstrated in Section 3, this framework can be related to multi-environment regression, hence creating a link between the two areas of study: (i) learning a regression, which performs well under unseen intervention settings, and (ii) selecting variables, based on their behavior under different observed interventions. Although we use a causal framework for formulation, we do not necessarily address the ambitious task of inferring causality but rather aim for a notion of stability and invariance. The goal of this paper is to analyze the connection between (i) and (ii) and use it to develop a methodological framework for inference.

This study is motivated by an application in systems biology in which one performs an exploratory analysis to discover the impact of genetic and environmental variants on known metabolic pathways and phenotypes (Roy et al. (2019), Williams et al. (2020), Čuklina et al.

Received November 2019; revised January 2021.
Key words and phrases. Regression, variable selection, causality, multiomic data.

(2021)). More specifically, we consider multiomic data from the transcriptome and proteome of a mouse population of 57 different inbred strains that was split into two groups, fed either with a low fat diet or a high fat diet. Liver tissue from these cohorts was then collected at multiple timepoints across their natural lifespans, providing diet as an independent biological (environment) variable. Based on these data, the target of interest is to associate gene expression of mRNAs and proteins in central metabolic pathways and using the independent biological variables to infer causality. This provides two avenues of hypotheses generation: (1) identifying pathway-associated genes, which are not in the canonical lists, and (2) determining which genes are (causally) upstream and driving pathway activity across the population as a function of diet.

1.1. *Stabilized regression.* Consider the following multi-environment regression setting; let $X = (X^1, \ldots, X^d) \in \boldsymbol{\mathcal{X}}$ be a vector of predictor variables and $Y \in \mathbb{R}$ a response variable, both of which are observed in different (perturbation) environments $e \in \mathcal{E}$. We assume that in each environment $e \in \mathcal{E}$, the variables (Y_e, X_e) have joint distribution P_e. Assume further that we only observe data from a subset of the environments $\mathcal{E}^{\text{obs}} \subseteq \mathcal{E}$. For each observed environment there are i.i.d. data, yielding n observations across all observed environments. The data can thus be represented by an $(n \times d)$-matrix $\mathbf{X}$, an $(n \times 1)$-vector $\mathbf{Y}$ and an $(n \times 1)$-vector $\mathbf{E}$ indicating which experiment the data points come from. The special case of an underlying linear model is shown in Figure 1 (Left: Observed training data; Right: Unobserved test data) with data generated according to a Gaussian linear model consisting of shift environments (Example 2.1). The data have been fitted on the training environment using linear regression on all variables (red) and on only the direct causal variables of the response (blue), which might be unknown in practice, of course. Since the underlying data generation process changes across settings, the regression based on all predictors leads to a biased prediction in the unobserved test environment, while the regression based only on the direct causal variables allows to generalize to these settings. At the same time the fit of the model based solely on the direct causal variables has higher variance on both training and test environments, compared with the regression based on all predictors. The method we describe in this paper attempts (without knowing the underlying model) to be able to generalize to unseen settings

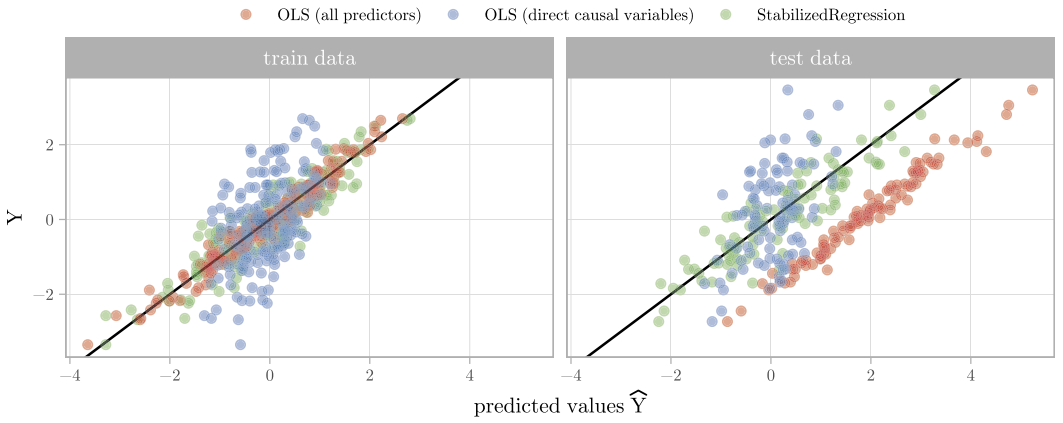


FIG. 1. *Illustrative example of three linear regression procedures applied to data generated according to Example 2.1 with two training and one testing environment. A good fit means that the dots are close to the identity line (given in black). Linear regression, based on all predictors (red), leads to biased results on the testing environment, while a linear regression, based only on direct causal variables of the response (blue), leads to unbiased estimation but with higher variance in both the testing and training environments. Stabilized regression (green) aims for the best fit which is also unbiased in the unobserved testing environment.*

without bias, while at the same time minimizing the prediction error. In Figure 1 we show the result of the proposed method in green.

Assuming an underlying causal structure, there is a key relation between a regression that is able to generalize and the variables that can be included into that regression. Details on this connection are given in Section 3. By looking at which sets of predictors lead to models that generalize and which do not, this gives us insights into the underlying causal mechanism.

Our method was developed as a tool for gene function discovery (e.g., Francesconi et al. (2008), Dutkowski et al. (2013)), where the goal is often two-fold: (i) Find novel gene relationships that can be associated to known pathways and (ii) Understand how these genes function within that pathway. For example, in the mouse data set mentioned above, one would like to both find genes that are related to a given pathway and understand whether their activity changes depending on diet or age. Often, such questions can be answered by understanding whether a functional dependence remains fixed or changes, depending on some exogenous environment variable. For an illustration of this problem based on the mouse data set, consider Figure 2. There, we consider protein expression levels of 3939 genes (based on $n = 315$ observations) and try to find functionally related genes to a known cholesterol biosynthesis gene (Hmgcs1). To do this, we set the response Y to be the protein expression levels of Hmgcs1 and then apply stabilized regression together with stability selection. The exact procedure is described in Section 6. In Figure 2 we plot the selection probabilities of genes (large probabilities imply we are certain about the finding) which either have an *unstable* or a *stable* functional relationship with Y across diets on the x-axis and y-axis, respectively. The genes have been annotated according to their relationship to the cholesterol biosynthesis pathway from the Reactome Pathway Knowledgebase (Fabregat et al. (2017)) which consists of 25 known canonical pathway genes of which 16 have been measured (including Hmgcs1). The result shows that stabilized regression is able to recover many relevant genes and also allows to separate findings into stable and unstable relationships. Details about the labeled genes and their relation with the cholesterol pathway are given in Supplementary Material A (Pfister et al. (2021)).

To achieve these goals, we propose a stabilizing procedure that can be combined with an arbitrary regression technique for each environment $e \in \mathcal{E}^{\text{obs}}$ individually. More specifically, for any subset $S \subseteq \{1, \dots, d\}$, let $\hat{f}^S$ be a regression estimate as a function of the predictors X^S. We then define the *stabilized regression* estimator to be a weighted average of the following form:

$$\hat{f}_{\text{SR}}(X) := \sum_{S \subseteq \{1,\dots,d\}} \widehat{w}_S \cdot \hat{f}^S(X^S), \tag{1.1}$$

where $\widehat{w}_S$ are normalized weights, that is, $\sum_S \widehat{w}_S = 1$. This type of model averaging appears often in the literature, and we discuss related approaches in Section 1.2. Commonly, the weights are chosen to optimize the predictive performance of the averaged model (e.g., by considering the residual sum of squares or various information criteria). We propose, however, that large weights should be given to models which are both stable and predictive. Here, stability means that the models do not vary much between the different environments. We provide a formal definition in Section 2, but other choices are possible, too, and may be of particular interest for complex data structures, such as dynamical data (Pfister, Bauer and Peters (2019)).

1.2. *Related work.* Predicting in new unobserved perturbed or changed environments is of huge importance in many applied areas and has been termed transfer learning or domain adaption in the machine learning and statistics community. While there are many different types of modeling frameworks for this problem, one well-established idea is to use

FIG. 2. *Stabilized regression (SR) applied to the* cholesterol biosynthesis *(CB) pathway. The data set consists of protein expression levels* ($n = 315$) *measured for* $d = 3939$ *genes,* 16 *of which are known to belong to CB (red gene names). We take protein expression levels of one known CB gene (Hmgcs1) as response* Y*. On the x- and y-axes we plot subsampling-based selection probabilities for two SR based variable selection procedures—y-axis: stable genes* $\mathrm{SB}_I(Y)$ *and x-axis: nonstable genes* $\mathrm{NSB}_I(Y)$*. (The precise definitions can be found in Section 3.) Many significant genes (green area) are canonical CB genes (red label) or part of an adjacent pathway (blue label). Annotated genes with a semievident relationship have yellow labels and with no clear relation black labels. The color coding of the nodes (interpolating between red and black) corresponds to the fraction of times the sign of the regression coefficient was negative/positive (red: negative sign; black: positive sign; grey: never selected).*

causal models (Pearl (2009)) and formalize the changes across environments by the notion of interventions. The key idea behind this approach is that causal models offer an intuitive way of modeling the conditional distribution of the response Y given its predictors X. More specifically, a causal model implies invariance of the conditional distribution under certain conditions which can be used to perform prediction in unseen environments. This is a fundamental concept in causality and has been referred to as invariance, autonomy or modularity (Aldrich (1989), Haavelmo (1944), Hoover (1990), Imbens and Rubin (2015), Wright (1921), Richardson and Robins (2013)). The invariance principle can be used to learn parts of a causal model from data and hence give a causal interpretation to some of the variables. This can be done by turning the invariance assumption around and inferring a causal model by finding

models which remain invariant. Using this idea to find direct causes of a response has been done in Peters, Bühlmann and Meinshausen (2016), Pfister, Bühlmann and Peters (2019) and Heinze-Deml, Peters and Meinshausen (2018). On the other hand, one can also use the invariance principle to improve prediction on unseen environments. Several existing methods learn models that explicitly enforce this assumption in order to generalize to new settings, as, for example, Schölkopf et al. (2012), Zhang et al. (2013), Rojas-Carulla et al. (2018) and Heinze-Deml and Meinshausen (2021) have. Others have tried to weaken the invariance assumption by only penalizing the noninvariance and hence trading-off generalization with in-sample prediction performance (e.g., Ganin et al. (2016), Pan et al. (2010), Rothenhäusler et al. (2021)). A general discussion about the relation of invariance and causality is given by Bühlmann (2020). Our proposed framework incorporates the idea of using invariance in order to improve generalization while at the same time aiming for a causal interpretation of the resulting variable selection.

From an algorithmic point of view, our proposed method is related to several averaging techniques from the literature. Averaging is a common regularization principle throughout statistics with many different types of applications in regression and variable selection. The idea of aggregating over several models is, for example, done in the generalized ensemble method due to Perrone and Cooper (1992), which gives explicit equations for optimal weights in terms of prediction MSE. Similar ideas, also exist in the Bayesian community, termed Bayesian model averaging (BMA) (Hoeting et al. (1999)). There, models are aggregated by optimizing the posterior approximation, based either on the Bayesian information criterion (BIC) (Schwarz (1978)) or on the Akaike information criterion (AIC) (leading to the so-called Akaike weights due to Burnham and Anderson (1998)). Our stabilized regression estimator in (1.1) averages over all subsets of predictors which is similar to how, for example, random forests (Breiman (2001)) are constructed. Other related approaches based on resampling subsets of predictors are due to Wang et al. (2011) and Cannings and Samworth (2017). Our method is, however, unique in combining this type of averaging with environment-wise stability or invariance.

Finally, the notion of stability has been widely used in several related contexts in statistics. As pointed out by, for example, Yu (2013) and Yu and Kumbier (2020), reproducible research relies on the statistical inference being stable across repetitions of the same procedure. This idea also underlies well-established resampling schemes, such as bagging by Breiman (1996) and stability selection by Meinshausen and Bühlmann (2010).

1.3. *Contributions*. We introduce a novel regression framework, based on averaging, that allows to incorporate environment-wise stability into arbitrary regression procedures. Under mild model assumptions our resulting regression estimates are shown to generalize to novel environmental conditions. The usefulness of our procedure is demonstrated for an application about gene detection from systems biology. For this application, besides using our novel stabilized regression, we propose an additional graphical tool which allows to visualize which genes are related to a response variable and whether this relationship is stable or unstable across environments. We believe this can aid practitioners to explore novel biological hypotheses. Finally, we introduce a theoretical framework for multi-environment regression and prove several results which relate it to structural causal models. Based on this correspondence, we introduce the stable blanket $\mathrm{SB}_I(Y)$, a subset of the Markov blanket, and discuss how this might help interpreting the output of different variable selection techniques. Our procedure is available in the R-package `StabilizedRegression` on CRAN.

1.4. *Outline*. In Section 2 we define our formal target of inference and describe the multi-environment regression setting. Then, in Section 3, we propose a causal model framework and

prove theoretical results relating the causal model perspective and multi-environment regression. Moreover, we introduce the concept of a stable blanket and discuss how this allows us to interpret different variable selection techniques. Most parts of this section can be skipped by the practical-minded reader. Our proposed algorithm is presented in Section 4 in which we also give details about practical issues in the implementation. In Section 5 we benchmark our method with commonly employed techniques based on two simulation experiments. Finally, in Section 6 we discuss the biological pathway analysis application in detail and explain how to construct visualizations as in Figure 2.

2. Multi-environment regression. Stabilized regression can be seen as a multi-environment regression technique for domain adaptation or transfer learning. The following summarizes the technical details of our multi-environment setup.

SETTING 1 (multi-environment regression). Let $\boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^d$ be a d-dimensional product of measurable spaces, let $X = (X^1, \ldots, X^d) \in \boldsymbol{\mathcal{X}}$ be a random vector of predictor variables, let $Y \in \mathbb{R}$ be a random response variable and let $\mathcal{E}^{\text{tot}}$ be a collection of perturbation environments such that, for each environment $e \in \mathcal{E}^{\text{tot}}$, the variables (Y_e, X_e) have joint distribution P_e. We assume that the distributions P_e are absolutely continuous with respect to a product measure which factorizes. Assume that we only observe data from a subset of the environments $\mathcal{E}^{\text{obs}} \subseteq \mathcal{E}^{\text{tot}}$.

Given this setting, our goal is to make predictions on a potentially unseen environment $e \in \mathcal{E}^{\text{tot}}$. For this to be meaningful, some assumption on the type of perturbations in $\mathcal{E}^{\text{tot}}$ is required. Motivated by previous work in causality (e.g., Peters, Bühlmann and Meinshausen (2016)), we assume that that there exists a subset $S \subseteq \{1, \ldots, d\}$ such that, for all environments $e, h \in \mathcal{E}^{\text{tot}}$ and all $x \in \boldsymbol{\mathcal{X}}$, it holds that

$$\mathbb{E}(Y_e \mid X_e^S = x^S) = \mathbb{E}(Y_h \mid X_h^S = x^S). \tag{2.1}$$

As we point out in Section 3, this assumption can be related to an underlying causal model. In that case, condition (2.1) coincides with parts of the causal system being fixed which is a fundamental concept referred to as invariance, autonomy or modularity.

An illustration of the multi-environment regression setting is given in Figure 3. Neglecting the environment structure, a classical approach to this problem is to use least squares to estimate a function $f : \boldsymbol{\mathcal{X}} \to \mathbb{R}$, which minimizes the (weighted) pooled squared loss

$$\sum_{e \in \mathcal{E}^{\text{obs}}} \frac{n_e}{n} \cdot \mathbb{E}\big((Y_e - f(X_e))^2\big), \tag{2.2}$$

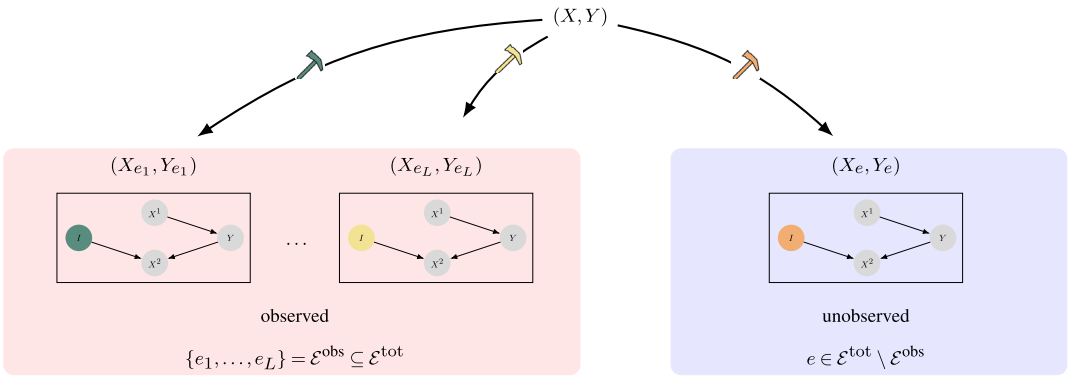


FIG. 3. *Illustration of multi-environment data generation setting. Only some environments are observed, but one would like to be able to make predictions on any further potentially unobserved environment.*

where n_e is the number of observations in environment e. Due to the heterogeneity, the optimizer on each individual environment, which is given by $f_e(x) = \mathbb{E}(Y_e \mid X_e = x)$, generally changes across environments. Therefore, it is not necessarily the case that the pooled optimizer generalizes to unseen settings $e \in \mathcal{E}^{\text{tot}} \setminus \mathcal{E}^{\text{obs}}$. Instead, we propose to explicitly use the assumed invariance in (2.1) and estimate a function $f : \boldsymbol{\mathcal{X}} \to \mathbb{R}$, which minimizes the pooled squared loss in (2.2) subject to the constraint that there exists a subset $S \subseteq \{1, \dots, d\}$, such that, for all $e \in \mathcal{E}^{\text{tot}}$ and all $x \in \boldsymbol{\mathcal{X}}$, it holds that

$$f(x) = \mathbb{E}\big(Y_e \mid X_e^S = x^S\big). \tag{2.3}$$

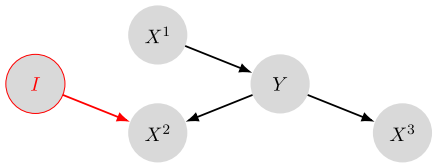
Define the constraint set $\mathcal{C} = \{f : \boldsymbol{\mathcal{X}} \to \mathbb{R} \mid f \text{ satisfies (2.3)}\}$ which is nonempty by the assumption in (2.1). Therefore, we have the following well-defined optimization problem:

$$\text{minimize} \quad \sum_{e \in \mathcal{E}^{\text{obs}}} \frac{n_e}{n} \cdot \mathbb{E}\big(\big(Y_e - f(X_e)\big)^2\big) \quad \text{subject to } f \in \mathcal{C}. \tag{2.4}$$

The standard approach to this problem is to solve the optimization directly by optimizing over all function $f \in \mathcal{C}$. We suggest a different approach. The optimization problem in (2.4) is equivalent to searching over all subset $S \subseteq \{1, \dots, d\}$ which satisfy (2.3) and for which the conditional mean based on the predictors X^S has minimal loss in (2.2). The solution to the optimization is then simply the conditional mean based on X^S. Such a set is not necessarily unique which is why our proposed method in Section 4 averages over an estimate for all these sets. The advantage of this approach is that, in particular in the finite sample case, the averaging technique leads to improved performance. This can be seen in Sections 5 and 6 in comparison with the instrumental variable procedure that in the linear case directly optimizes (2.4). The following toy example illustrates the difference between the unconstrained optimization in (2.2) and constrained optimization in (2.4).

EXAMPLE 2.1 (toy model). Consider a variable I which generates the environments or perturbations. Let the variables (I, X, Y) satisfy the following structural causal model (Definition 3.1):

$$\mathcal{S}^* \quad \begin{cases} I := \varepsilon_I \\ X^1 := \varepsilon_{X^1} \\ Y := X^1 + \varepsilon_Y \\ X^2 := Y + I + \varepsilon_{X^2} \\ X^3 := Y + \varepsilon_{X^3} \end{cases}$$



with ε_Y, ε_{X^1}, ε_{X^2} and ε_{X^3} independently $\mathcal{N}(0, 1)$-distributed and $\varepsilon_I = c(e)$ for a constant $c(e) \in \mathbb{R}$, depending on the environment $e \in \mathcal{E}^{\text{obs}}$. Variable I is unobserved and describes the changes across environments (see Section 3). Consider two cases, where: (i) only the variables (Y, X^1, X^2) and (ii) only the variables (Y, X^1, X^2, X^3) are observed. Given case (i) and assuming a mixture model across the observed environments $\mathcal{E}^{\text{obs}}$ (with equal probabilities across all environments) allows us to compare optimization of (2.2) solved by a pooled least squares estimator with optimization (2.4) by a simple calculation. The standard ordinary least squares (OLS) estimator in the population case is given by

$$\beta^{\text{OLS}} = \begin{pmatrix} \text{Var}(X^1) & \text{Cov}(X^1, X^2) \\ \text{Cov}(X^1, X^2) & \text{Var}(X^2) \end{pmatrix}^{-1} \begin{pmatrix} \text{Cov}(X^1, Y) \\ \text{Cov}(X^2, Y) \end{pmatrix} = \begin{pmatrix} \dfrac{1 + \text{Var}(I)}{2 + \text{Var}(I)} \\ \dfrac{1}{2 + \text{Var}(I)} \end{pmatrix},$$

where by slight abuse of notation $\text{Var}(I)$ refers to the variation of $c(e)$ across environments. Hence, the coefficient of X^2 is nonzero in this case implying that predictions can become bad

on environments where I takes large values. Since the constraint in (2.3) is satisfied for both $S = \varnothing$ and $S = \{1\}$, the optimizer of (2.4) is given by $f(x) = \mathbb{E}(Y \mid X^1 = x^1) = x^1$, and the optimal regression parameter is given by $\beta^* = (1, 0)^\top$. This regression coefficient is ideal in the sense that it contains all the information about Y that can be explained independent of the value of I. If the observed perturbations have a large spread, that is, $\frac{1}{|\mathcal{E}^{\text{obs}}|}\sum_{e\in\mathcal{E}^{\text{obs}}} c(e)^2$ is large, then the OLS regression parameter β^{OLS} approximates the constrained regression parameter β^* (see Corollary 3.7). Strong heterogeneity in the data, therefore, improves the generalization performance of a standard pooled regression.

Consider now case (ii) in which we additionally observe variable X^3. While X^2 was harmful for the generalization performance, X^3 is in general beneficial (see Figure 1). In particular, the regression parameter for the regression of Y on (X^1, X^2, X^3) with the constraint in (2.3) has the form $\beta^* = (\beta_1^*, 0, \beta_2^*)$, where the two parameters are, in general, nonzero and depend on the underlying system. Similar to case (i), it can be shown that the standard OLS parameter again converges to this constrained estimator if the interventions are sufficiently strong. A formal result describing when the pooled OLS converges to the constrained optimizer in the case of linear systems is given in Section 3.4. In many applications, however, there might be insufficient heterogeneity for the OLS, and the difference between solutions to (2.2) and (2.4) might be substantial. Therefore, whenever the training environments consist of weaker interventions than the testing environment, one can benefit from explicitly incorporating stability into the estimation (also shown in Figure 1).

The pooled squared loss (2.2) and the constraint (2.3) combine two aspects: (i) Predictive performance of the model given by the optimization objective and (ii) stability across perturbations enforced by the constraint in (2.3). These concepts are formalized in the following definitions.

DEFINITION 2.2 (generalizable sets). A set $S \subseteq \{1, \dots, d\}$ is called generalizable with respect to $\mathcal{E} \subseteq \mathcal{E}^{\text{tot}}$ if, for all $e, h \in \mathcal{E}$ and for all $x \in \boldsymbol{\mathcal{X}}$, it holds that

$$\mathbb{E}\big(Y_e | X_e^S = x^S\big) = \mathbb{E}\big(Y_h | X_h^S = x^S\big).$$

We denote by $\mathbb{G}_{\mathcal{E}}$ the collection of all generalizable sets.

Any generalizable set will, by definition, have the property that a regression based on the predictors in that set should have similar predictive performance across all environments $e \in \mathcal{E}$. In practice, it is, however, also important that the predictive performance is not only equal across different environments but is equally good in all environments.

DEFINITION 2.3 (generalizable and regression optimal sets). A set $S \subseteq \{1, \dots, d\}$ is called generalizable and regression optimal with respect to $\mathcal{E} \subseteq \mathcal{E}^{\text{tot}}$ if it is generalizable in the sense that $S \in \mathbb{G}_{\mathcal{E}}$ and if it satisfies

$$S \in \operatorname*{arg\,min}_{\bar{S}\in\mathbb{G}_{\mathcal{E}}} \sum_{e\in\mathcal{E}} \frac{n_e}{n} \cdot \mathbb{E}\big[\big(Y_e - \mathbb{E}\big(Y_e | X_e^{\bar{S}}\big)\big)^2\big].$$

The collection of all generalizable and regression optimal sets (with respect to $\mathcal{E}$) is denoted by $\mathbb{O}_{\mathcal{E}}$.

In general, the sizes of $\mathbb{G}_{\mathcal{E}}$ and $\mathbb{O}_{\mathcal{E}}$ decrease when more environments are added to $\mathcal{E}$. In Section 3.3 we discuss when the observed environments $\mathcal{E}^{\text{obs}}$ are sufficient for generalization to all potential environments $\mathcal{E}^{\text{tot}}$, that is, when $\mathbb{G}_{\mathcal{E}^{\text{obs}}} = \mathbb{G}_{\mathcal{E}^{\text{tot}}}$ and $\mathbb{O}_{\mathcal{E}^{\text{obs}}} = \mathbb{O}_{\mathcal{E}^{\text{tot}}}$ hold. In Section 4 we will introduce an algorithm that approximates a solution to the constrained optimization (2.4) by explicitly estimating the generalizable and regression optimal sets.

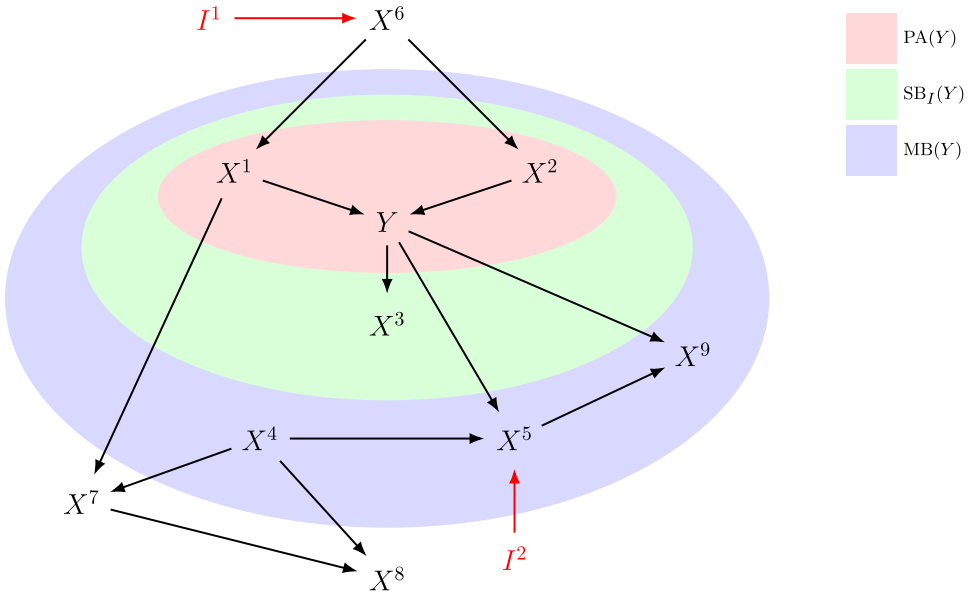


FIG. 4. *Graphical illustration of variable selection. The goal is to find predictors* $X = (X^1, \dots, X^9)$ *that are functionally related to the response* Y. *Here, variables* $I = (I^1, I^2)$ *are unobserved intervention variables. The colored areas represent different targets of inference*: *Markov blanket, stable blanket and parents* (*causal variables*). *If the goal is to get as close as possible to the parents, the stable blanket can improve on the Markov blanket if there are sufficiently many informative interventions.*

3. Stable blankets. As in the previous sections, assume Y is a response variable and X is a set of predictors. In this section we want to consider the problem of finding a subset of predictors that are functionally related to Y. An example is given by the causal graphical model illustrated in Figure 4 (see Section 3.1 for details). A simple but common approach is to select predictors by pairwise association with Y. Often this selects many predictors, for example, in Figure 4 it might result in all of the predictors being selected. A fine-tuned approach is to predict Y from X and select the predictors that were most important in the prediction model. In the notation of Section 2, such a set of predictors would be regression optimal. Given further assumptions, it can be shown that, in some cases, such approaches actually recover the smallest set of informative predictors which is known as the Markov blanket of Y in the graphical model literature and denoted by $\mathrm{MB}(Y)$. As shown in Figure 4, $\mathrm{MB}(Y)$ consists of variables that are functionally *closer* to Y.

There is an important causal distinction between different predictors in $\mathrm{MB}(Y)$: Interventions that do not directly target Y will never affect the functional relation between Y and the causal parents of Y (denoted by $\mathrm{PA}(Y)$), but they can change the relation of Y with other variables in $\mathrm{MB}(Y)$ (e.g., X^5 in Figure 4). This motivates the idea of additionally distinguishing between stable and nonstable predictors. In analogy to the Markov blanket, this leads to the definition of the stable blanket of Y, which we denote by $\mathrm{SB}_I(Y)$ and which is equal to the smallest set of predictors that contains all information about Y that is unaffected by interventions. In Section 3.3 we will discuss how $\mathrm{SB}_I(Y)$ and $\mathrm{MB}(Y)$ are related to generalizable and regression optimal sets. A useful property of the stable blanket is that it always satisfies the hierarchy $\mathrm{PA}(Y) \subseteq \mathrm{SB}_I(Y) \subseteq \mathrm{MB}(Y)$. We can, therefore, use $\mathrm{SB}_I(Y)$ as a proxy for upstream predictors of Y. Similarly, the nonstable blanket $\mathrm{NSB}_I(Y) := \mathrm{MB}(Y) \setminus \mathrm{SB}_I(Y)$ can be used as a proxy for down-stream predictors of Y.

3.1. *A causal model perspective.* In this section we additionally assume an underlying causal model which allow us to specify graphical conditions for describing generalizable

sets. This characterization is not important from a methodological viewpoint but helps from a causal modeling perspective and can give some useful insights for interpreting the results of variable selection. It uses some terminology and concepts from the causal literature. The practically-oriented reader may skip this subsection.

We choose to work with structural causal models (SCMs) (e.g., Pearl (2009), Peters, Janzing and Schölkopf (2017)), sometimes also referred to as structural equation models (SEMs).

DEFINITION 3.1 (structural causal model). A structural causal model (SCM), over random variables $W = (W^1, \dots W^p)$, is a collection of p assignments,

$$\mathcal{S} \quad \begin{cases} W^1 := f^1\big(W^{\mathrm{PA}(W^1)}, \varepsilon^1\big) \\ \vdots \\ W^p := f^p\big(W^{\mathrm{PA}(W^d)}, \varepsilon^p\big), \end{cases} \tag{3.1}$$

where $\varepsilon^1, \dots, \varepsilon^p$ are independent noise variables. For all $k \in \{1, \dots, p\}$, $\mathrm{PA}(W^k) \subseteq \{1, \dots, p\} \setminus \{k\}$ is called the set of direct (causal) parents of W^k. Moreover, the assignments in (3.1) are assumed to be uniquely solvable, which is always true if the induced graph is acyclic, for example. An SCM induces a distribution over the variables W as well as a graph over the vertices $(W^1, \dots, W^p)$, denoted by $\mathcal{G}(\mathcal{S})$, by adding directed edges from $\mathrm{PA}(W^k)$ to W^k for all $k \in \{1, \dots, p\}$.

For any SCM $\mathcal{S}$ over $W = (W^1, \dots, W^p)$, an intervention on a variable W^j corresponds to a new SCM $\tilde{S}$ for which only the structural assignment of W^j has been replaced. We only consider interventions for which the new SCM remains solvable. When talking about graphs, we use the notion of d-separation (e.g., Pearl (2009)), which we denote by $\perp\!\!\!\perp_{\mathcal{G}}$ to distinguish it from conditional independence. We summarize the causal model setting below.

SETTING 2 (underlying causal model). Let $X \in \boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^d$ be predictor variables, $Y \in \mathbb{R}$ a response variable and $I = (I^1, \dots, I^m) \in \boldsymbol{\mathcal{I}} = \mathcal{I}^1 \times \cdots \times \mathcal{I}^m$ intervention variables which are assumed to be unobserved and are used to formalize interventions. Assume there exists a fixed SCM $\mathcal{S}^*$ over (I, X, Y) such that $\mathcal{G}(\mathcal{S}^*)$ is a directed acyclic graph (DAG) and for which the intervention variables I are source nodes and do not appear in the structural assignment of Y. An intervention environment e corresponds to an intervention SCM $\mathcal{S}_e$ over (I_e, X_e, Y_e) in which only the equations with I_e on the right-hand side change and the graph structure stays fixed (i.e., $\mathcal{G}(\mathcal{S}_e) = \mathcal{G}(\mathcal{S}^*)$). Let $\mathcal{E}^{\mathrm{tot}}$ be the set of all such intervention environments, and let $\mathcal{E}^{\mathrm{obs}} \subseteq \mathcal{E}^{\mathrm{tot}}$ be a finite set of observed environments. Lastly, assume the distribution of (I_e, X_e, Y_e) is absolutely continuous with respect to a product measure that factorizes.

The intervention variables I are introduced as auxiliary variables to specify the intervention locations, similar to augmented DAGs or influence diagrams in the literature (e.g., Dawid (2002)). For the sake of simplicity, we assume the existence of an entire joint distribution of (I, X, Y). This allows interpreting interventions as conditioning statements on the intervention variables I, since they are source nodes. In general, the intervention variables I, however, do not need to be stochastic and can also be modeled as deterministic using a more complicated notion of conditional independence as in Constantinou and Dawid (2017).

Based on this setting, we can define intervention stable sets. Intuitively, a set S is called intervention stable if the corresponding predictors explain all of the intervention variability in the response variable.

DEFINITION 3.2 (intervention stable sets). Given Setting 2, a set $S \subseteq \{1, \dots, d\}$ is called intervention stable if, for all $\ell \in \{1, \dots, m\}$, the d-separation $I^\ell \perp\!\!\!\perp_{\mathcal{G}} Y \mid X^S$ holds in $\mathcal{G}(S^*)$.

As an example, consider the predictor set $S = \{1, 2, 9\}$ in Figure 4. The variable X^9 opens a path from I^2 to Y which means the set is not intervention stable. In contrast, the parent set $S = \{1, 2\}$ is intervention stable. More generally, since the graph $\mathcal{G}(S^*)$ remains fixed across interventions, it immediately follows that $\mathrm{PA}(Y)$ is always an intervention stable set. Together with the following proposition (which proves that any intervention stable set is generalizable), this implies that the invariance assumption in (2.1) is satisfied.

PROPOSITION 3.3 (intervention stable sets are generalizable). *Assume Setting 2; then, for all intervention stable sets $S \subseteq \{1, \dots, d\}$ it holds that $S \in \mathbb{G}_{\mathcal{E}^{\mathrm{tot}}}$.*

A proof is given in Supplementary Material A (Pfister et al. (2021)). Based on this proposition, it is possible to find generalizable sets using only the graphical structure. However, not all generalizable sets are intervention stable. More details on this relation are given in Section 3.3.

In graphical models the Markov blanket of Y, denoted by $\mathrm{MB}(Y)$, is defined as the smallest set $S \subseteq \{1, \dots, d\}$ that satisfies

$$\forall j \in \{1, \dots, d\} \setminus S: \quad X^j \perp\!\!\!\perp_{\mathcal{G}} Y \mid X^S. \tag{3.2}$$

The Markov blanket specifies the smallest set of variables that separates the response Y from all other variables and hence allows a precise notion of predictiveness. The following definition combines this notion with intervention stability.[1]

DEFINITION 3.4 (stable blanket). Assume Setting 2, and define the following set of variables:

$$N^{\mathrm{int}} := \{1, \dots, d\} \setminus \big\{ j \in \{1, \dots, d\} \mid \exists k \in \mathrm{CH}^{\mathrm{int}}(Y) : j \in \mathrm{DE}(X^k) \big\},$$

where $\mathrm{CH}^{\mathrm{int}}(Y)$ are all children of Y that are directly intervened on and $\mathrm{DE}(X^k)$ are all descendants of X^k, including X^k itself. Then, the stable blanket, denoted by $\mathrm{SB}_I(Y)$, is defined as the smallest set $S \subseteq N^{\mathrm{int}}$ that satisfies

$$\forall j \in N^{\mathrm{int}} \setminus S: \quad X^j \perp\!\!\!\perp_{\mathcal{G}} Y \mid X^S. \tag{3.3}$$

In words, the set N^{int} consists of all variables that are neither children of Y, which have been intervened, nor descendants of such children. In the example in Figure 4, it consists of all variables, except X^5 and X^9, that is, $N^{\mathrm{int}} = \{1, 2, 3, 4, 6, 7, 8\}$. It is helpful to compare (3.2) and (3.3) to see the parallels and differences between the Markov blanket and the stable blanket. A further characterization of the stable blanket is given in the following theorem which also proves that it is generalizable and regression optimal.

THEOREM 3.5 (stable blankets are generalizable and regression optimal). *Assume Setting 2; then, the stable blanket consists of all children of Y that are not in N^{int}, the parents of such children and the parents of Y. Furthermore, it holds that $\mathrm{SB}_I(Y) \in \mathbb{O}_{\mathcal{E}^{\mathrm{tot}}}$.*

A proof is given in Supplementary Material A (Pfister et al. (2021)). It is illustrative to think about the set $\mathrm{SB}_I(Y)$ in relation to the parent set $\mathrm{PA}(Y)$ and the Markov blanket $\mathrm{MB}(Y)$. By Theorem 3.5, it will lie somewhere between these two sets. The exact size depends on the intervention variables with the following special cases: (i) if there are no interventions, it holds that $\mathrm{SB}_I(Y) = \mathrm{MB}(Y)$; (ii) if there are sufficiently many interventions, for example, on any node other than Y, it holds that $\mathrm{SB}_I(Y) = \mathrm{PA}(Y)$.

[1] The union of intervention stable sets is, itself, not necessarily intervention stable anymore.

3.2. *Stable blanket as a proxy for causality.* As alluded to in the previous section, the stable blanket $\mathrm{SB}_I(Y)$ can be seen as a proxy for the causal parents. In the most basic case of an SCM with an underlying directed acyclic structure, the Markov blanket can be decomposed into parents, children and parents of children, that is,

$$\mathrm{MB}(Y) = \mathrm{PA}(Y) \cup \mathrm{CH}(Y) \cup \big\{ j \in \{1, \dots, d\} \mid \exists k \in \mathrm{CH}(Y) : j \in \mathrm{PA}\big(X^k\big) \big\}.$$

As long as the intervention variables do not directly affect the response Y, this implies that the difference between the Markov blanket and the stable blanket consists only of variables that are children or parents of children of the response. We denote this difference as the nonstable blanket

$$\mathrm{NSB}_I(Y) := \mathrm{MB}(Y) \setminus \mathrm{SB}_I(Y).$$

Given the above decomposition, this implies that $\mathrm{PA}(Y) \subseteq \mathrm{SB}_I(Y)$ and $\mathrm{NSB}_I(Y) \subseteq \mathrm{CH}(Y) \cup \{j \in \{1, \dots, d\} \mid \exists k \in \mathrm{CH}(Y) : j \in \mathrm{PA}(X^k)\}$. Therefore, depending on whether we are either interested in the parents or in down-stream variables (or children) of Y, the sets $\mathrm{SB}_I(Y)$ and $\mathrm{NSB}_I(Y)$ can be used as proxies.

3.3. *Identifiability of generalizable sets.* In Section 2 we introduced the collection of generalizable and regression optimal predictor sets $\mathbb{O}_{\mathcal{E}^{\mathrm{tot}}}$ which lead to regressions that behave well on all potential environments $\mathcal{E}^{\mathrm{tot}}$. We saw that if one assumes an underlying causal model, as in Section 3, it is possible to compute the stable blanket $\mathrm{SB}_I(Y)$. This shows, since $\mathrm{SB}_I(Y) \in \mathbb{O}_{\mathcal{E}^{\mathrm{tot}}}$, that it is possible to construct a generalizable and regression optimal set whenever the underlying causal structure is known. In practice, we usually do not have access to the causal structure and only observe a (small) subset $\mathcal{E}^{\mathrm{obs}}$ of all potential environments $\mathcal{E}^{\mathrm{tot}}$. Intuitively, the best one can hope for in such cases is to find sets in $\mathbb{O}_{\mathcal{E}^{\mathrm{obs}}}$. Therefore, the question arises whether and when the sets in $\mathbb{O}_{\mathcal{E}^{\mathrm{obs}}}$ also generalize to any further environments not contained in $\mathcal{E}^{\mathrm{obs}}$. The answer depends on the assumptions one is willing to make on the data generating process and, in particular, on the types of environments that are observed and unobserved. In this section we discuss additional conditions to Setting 2, that allow generalization from $\mathcal{E}^{\mathrm{obs}}$ to $\mathcal{E}^{\mathrm{tot}}$.

Given Setting 2, we are interested in what additional conditions are sufficient to be able to infer the stable blanket and hence a generalizable and regression optimal set from data. In order to compute $\mathrm{SB}_I(Y)$, we need to be able to determine whether a given set is intervention stable based on data. We require two types of assumptions.

First, the faithfulness assumption (Pearl (2009)) ensures that any conditional independence in the data generating random variables corresponds to a d-separation in the graph. Given faithfulness and a sufficiently large sample size, it is possible, in most cases, to consistently recover the Markov blanket using, for example, an appropriate feature selection algorithm (Pellet and Elisseeff (2008)). This, in particular, does not require any type of heterogeneity and can be based purely on observational data.

Second, to check whether a subset $S \subseteq \{1, \dots, d\}$ is intervention stable requires to detect all conditional dependencies between the intervention variables and the response given the predictors in S. Since only the environments are observed and not the intervention variables, we require that

$$\begin{aligned} &\forall e, h \in \mathcal{E}^{\mathrm{obs}} : \mathbb{E}\big(Y_e \mid X_e^S = x^S\big) = \mathbb{E}\big(Y_h \mid X_h^S = x^S\big) \\ &\quad \Rightarrow \quad \forall \ell \in \{1, \dots, m\} : I^\ell \perp\!\!\!\perp Y \mid X^S. \end{aligned}$$

In other words, by contraposition we need that any conditional dependence between the intervention variables and the response leads to a shift in conditional mean across environments.

3.4. *Understanding stable blankets in linear models.* To get a better understanding of the relation between stable blankets and standard regression techniques, we consider linear models and analyze the behavior of the pooled ordinary least squares (OLS) estimator in our proposed multi-environment regression setting. We will show that the OLS only sets variables in the nonstable blanket to zero if the intervention strength goes to infinity. This means that, whenever the intervention strength is not sufficiently strong, OLS does not necessarily perform well on unobserved environments with stronger interventions.

For our results it is enough to consider population quantities since the ordinary least squares estimator is consistent. The following lemma gives an explicit expression of the population OLS applied to a linear SCM in terms of the (exogenous) noise variables and the coefficient matrix. It allows us to assess the behavior of the OLS under interventions.

LEMMA 3.6 (OLS in linear SCMs). *Assume the variables $(X, Y) \in \mathbb{R}^{d+1}$ satisfy a linear directed acyclic SCM, that is, there exists $B \in \mathbb{R}^{(d+1)\times(d+1)}$ and independent noise variable $\varepsilon = (\varepsilon^0, \dots \varepsilon^d) \in \mathbb{R}^{d+1}$ such that*

$$\begin{pmatrix} Y \\ X \end{pmatrix} := B \cdot \begin{pmatrix} Y \\ X \end{pmatrix} + \varepsilon \quad \text{with } B = \begin{pmatrix} 0 & \beta_{\mathrm{PA}}^\top \\ \beta_{\mathrm{CH}} & B_X \end{pmatrix},$$

where $B_X \in \mathbb{R}^{d\times d}$ and $\beta_{\mathrm{CH}}, \beta_{\mathrm{PA}} \in \mathbb{R}^{d\times 1}$. The parents and children of Y are given by the nonzero coefficients β_{PA} and β_{CH}, respectively. Then, the population ordinary least squares β^{OLS}, when regressing Y on X, is given by

$$\beta^{\mathrm{OLS}} = \beta_{\mathrm{PA}} + \big((\mathrm{Id} - B_X)^\top - \beta_{\mathrm{PA}}\beta_{\mathrm{CH}}^\top\big) D^{-1} \beta_{\mathrm{CH}} \bigg(1 - \frac{\sigma_0^2 \beta_{\mathrm{CH}}^T D^{-1} \beta_{\mathrm{CH}}}{1 + \sigma_0^2 \beta_{\mathrm{CH}}^T D^{-1} \beta_{\mathrm{CH}}}\bigg) \sigma_0^2,$$

where $D = \mathrm{Cov}(\varepsilon^1, \dots, \varepsilon^d)$ and $\sigma_0^2 = \mathrm{Var}(\varepsilon^0)$.

A proof is given in Supplementary Material A (Pfister et al. (2021)). The result implies that the population OLS can be decomposed into the sum of the true causal parameter β_{PA} plus a correction term. It can be shown that this correction is zero for coordinates $j \notin \mathrm{MB}(Y)$ (see proof of Corollary 3.7) which is a well-known property of ordinary least squares. Moreover, we can explicitly analyze the behavior of the OLS in the multi-environment regression setting. In particular, it can be shown that $\beta^{\mathrm{OLS},j}$ converges to zero for variables $j \notin \mathrm{SB}_I(Y)$, as the variance of the interventions across environments increases. The exact result is given in the following corollary.

COROLLARY 3.7 (OLS under strong interventions). *Let $(I_{\bar{n}}, X_{\bar{n}}, Y_{\bar{n}})$ be a sequence of variables satisfying Setting 2 for the same directed acyclic linear SCM $\mathcal{S}^*$. Additionally, assume that each of the variables $I_{\bar{n}}$ has exactly one child, and the sum of the coefficients along directed paths starting at variables $I_{\bar{n}}$ are always nonvanishing. Moreover, for all $\bar{n} \in \mathbb{N}$, there are two observed environments $\mathcal{E}_{\bar{n}}^{\mathrm{obs}} = \{e_{\bar{n}}^+, e_{\bar{n}}^-\}$, where the interventions $e_{\bar{n}}^+$ and $e_{\bar{n}}^-$ satisfy for all $\ell \in \{1, \dots, m\}$ that*

$$I_{\bar{n}}^\ell = c_{\ell,\bar{n}}^+ \text{ in } \mathcal{S}_{e_{\bar{n}}^+} \quad \text{and} \quad I_{\bar{n}}^\ell = c_{\ell,\bar{n}}^- \text{ in } \mathcal{S}_{e_{\bar{n}}^-},$$

where $c_{\ell,\bar{n}}^+, c_{\ell,\bar{n}}^-$ are independent random variables with mean zero and variance $\sigma_{\bar{n}}^2$ such that $\lim_{\bar{n}\to\infty} \sigma_{\bar{n}} = \infty$. Then, the pooled OLS estimator $\beta_{\bar{n}}^{\mathrm{OLS}}$ when regressing $Y_{\bar{n}}$ on $X_{\bar{n}}$ (i.e., the minimizer of (2.2) over all linear functions) satisfies, for all $j \in \{1, \dots, d\} \setminus \mathrm{SB}_I(Y)$,

$$\lim_{\bar{n}\to\infty} \beta_{\bar{n}}^{\mathrm{OLS},j} = 0.$$

A proof is given in Supplementary Material A (Pfister et al. (2021)). We use $\bar{n}$ to make clear that this is a population result in which the limit is taken in terms of intervention strength and not in terms of sample size. Corollary 3.7 provides results in an asymptotic regime in which the interventions are sufficiently strong. In the numerical simulations in Section 5, we will see that, whenever the intervention strength is not sufficiently strong, the OLS can be outperformed.

4. Proposed method. Our goal is to fit a regression function which approximates a solution to (2.4). Instead of just finding a single set S for which the conditional mean based on X^S solves (2.4), we propose to approximate this function with a weighted average. The idea is that verifying the invariance constraint in (2.3) involves uncertainty which can be reduced by averaging over many invariant sets instead of deciding on a single set. For any subset $S \subseteq \{1, \dots, d\}$, let $\hat{f}^S : \boldsymbol{\mathcal{X}}^{|S|} \to \mathbb{R}$ be a regression estimate which minimizes (2.2) restricted to the predictors in S. Recall that the *stabilized regression* estimator is defined as the weighted average

$$\hat{f}_{\mathrm{SR}}(X) := \sum_{S \subseteq \{1,\dots,d\}} \widehat{w}_S \cdot \hat{f}^S(X^S), \tag{4.1}$$

where the weights are assumed to satisfy $\sum_S \widehat{w}_S = 1$. For this estimator to approximate a solution of (2.4), we select large weights for sets of predictors which are both generalizable and regression optimal.

4.1. *Estimating generalizable and regression optimal sets.* Let $\widehat{\mathbb{O}}$ be a subset of the power set of $\{1, \dots, d\}$ that estimates the collection of generalizable and regression optimal sets with respect to $\mathcal{E}^{\mathrm{obs}}$. Then, we propose to construct the weights as follows:

$$\widehat{w}_S := \begin{cases} 1/|\widehat{\mathbb{O}}| & \text{if } S \in \widehat{\mathbb{O}}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

The set $\widehat{\mathbb{O}}$ can be estimated by a score-based approach as follows. For each set $S \subseteq \{1, \dots, d\}$ compute two scores: (i) A stability score, denoted by $\mathbf{s}_{\mathrm{stab}}(S)$, which measures how well the regression based on predictors from S satisfies the invariance (2.1) and (ii) a prediction score, denoted by $\mathbf{s}_{\mathrm{pred}}(S)$, which measures how predictive the regression based on predictors from S is. Based on these scores, estimate the collection of generalizable sets as

$$\widehat{\mathbb{G}} := \{S \subseteq \{1, \dots, d\} \mid \mathbf{s}_{\mathrm{stab}}(S) \geq c_{\mathrm{stab}}\}$$

and the collection of generalizable and regression optimal sets as

$$\widehat{\mathbb{O}} := \{S \in \widehat{\mathbb{G}} \mid \mathbf{s}_{\mathrm{pred}}(S) \geq c_{\mathrm{pred}}\}.$$

The cutoff parameters c_{stab} and c_{pred} are tuning parameters. Depending on the data, the regression technique and potential domain knowledge, different types of scores and cutoffs can be selected.

Below, we discuss explicit options for constructing stability and prediction scores. We focus on settings where the response can be expressed as a function of the predictors with additive noise, that is, $Y = f(X) + \varepsilon$. For the stability score we propose an approximate hypothesis test for the null hypothesis $S \in \mathbb{G}_{\mathcal{E}^{\mathrm{obs}}}$ (see Section 4.1.1). For the prediction score, a bootstrap approach, based on mean squared errors, can be employed (see Section 4.1.2).

4.1.1. *Stability scores*. We propose to construct stability scores for each set $S \subseteq \{1, \dots, d\}$ by a test for the null hypothesis $S \in \mathbb{G}_{\mathcal{E}^{\text{obs}}}$, that is, whether S satisfies the invariance (2.1). Once such a test has been selected, we set, for any set $S \subseteq \{1, \dots, d\}$, the stability score $\mathbf{s}_{\text{stab}}(S)$ to be the p-value of this test. An intuitive parameterization is to set the cutoff c_{stab} to be the type-1 error control for the hypothesis test which controls the trade-off of how stringently we want to enforce stability. There are many ways in which a hypothesis test for this problem can be constructed. Here, we discuss some potential starting points for the general case and conclude with two well-known tests for Gaussian linear models. Assume we fit a regression function $\hat{f}_e^S$ on each observed environment $e \in \mathcal{E}^{\text{obs}}$ individually. Given the null hypothesis $S \in \mathbb{G}_{\mathcal{E}^{\text{obs}}}$, all of these regression functions should be approximately equal up to the estimation error, that is, $\hat{f}_e^S \approx \hat{f}_h^S$. As a consequence, the residuals $\hat{R}_e^S = Y_e - \hat{f}_e^S(X_e^S)$ on each environment should also have approximately the same distribution, that is, $\hat{R}_e^S \overset{d}{\approx} \hat{R}_h^S$. One can, therefore, construct a hypothesis test by explicitly quantifying the estimation error in either of these approximations. However, in order to be able to do this, one needs to make some assumptions on the data-generating process. In the case of linear regression, when the data generating process is a linear model with Gaussian noise ($Y = \beta X + \varepsilon$), we can explicitly test for equal regression parameters $\hat{\beta}_e$ and $\hat{\beta}_h$ using a Chow test (Chow (1960)). A slight disadvantage of this test is that it can only test equivalence between two environments at a time. This means one needs to correct for multiple testing whenever there are more than two environments. A second option in the Gaussian linear case is to use a resampling based test, as suggested by Shah and Bühlmann (2018). One can show that it is possible to exactly resample from the distribution of the scaled residuals $R_e/\|R_e\|_2$. This allows to construct a test for an arbitrary test statistic, based on $R_e/\|R_e\|_2$ (e.g., the sum of differences in mean across environments).

4.1.2. *Prediction scores*. For the prediction score we propose to either use the negative mean squared prediction error or the negative minimal environment-wise mean squared prediction error. We use negative values to ensure that large values imply predictive and small values nonpredictive. To make the cutoff interpretable and easier to select, one can use the following bootstrap procedure. For every set $S \subseteq \{1, \dots, d\}$, let $\mathbf{s}_{\text{pred}}(S)$ be the chosen prediction score. Construct B bootstrap samples, $(\mathbf{X}_1^*, \mathbf{Y}_1^*), \dots, (\mathbf{X}_B^*, \mathbf{Y}_B^*)$, and define for every $S \subseteq \{1, \dots, d\}$ the bootstrap distribution function of the prediction score for all $t \in \mathbb{R}$ as

$$F^*_{\mathbf{s}_{\text{pred}}(S)}(t) := \sum_{i=1}^{B} \mathbb{1}_{\{\mathbf{s}_{\text{pred}}(S)(\mathbf{X}_i^*, \mathbf{Y}_i^*) \le t\}}.$$

Moreover, let $Q \in \widehat{\mathbb{G}}$ be the set of predictors with maximal prediction score, that is, $Q := \arg\max_{S \in \widehat{\mathbb{G}}} \mathbf{s}_{\text{pred}}(S)(\mathbf{X}, \mathbf{Y})$. Then, we choose the cutoff parameter to be $c_{\text{pred}} = (F^*_{\mathbf{s}_{\text{pred}}(Q)})^{-1}(\alpha_{\text{pred}})$, where $\alpha_{\text{pred}} \in (0, 1)$ specifies how strongly to focus on the most predictive set.

4.2. *Variable importance*. Based on the stabilized regression estimator, it is possible to define several types of variable importance measures that can then be used to recover either the Markov blanket, the stable blanket or the nonstable blanket.

Assume we have computed the stabilized regression estimator given in (1.1). Then, for each variable $j \in \{1, \dots, d\}$, define the weight variable importance as follows:

$$v_j^{\text{weight}} := \sum_{S \subseteq \{1, \dots, d\}} \widehat{w}_S \cdot \mathbb{1}_{\{j \in S\}}.$$

This means the importance of a variable depends on how often it appears with a positive weight. In the case of linear regression, a similar importance measure can be defined. To that end, let the individual regression functions be given by $\widehat{f}_S : x \mapsto \widehat{\beta}_S^\top x$, where $\widehat{\beta}_S$ is the (scaled) ordinary least squares estimator based on the predictor set S with zeros at all other coordinates. Then, define the coefficient variable importance as

$$v_j^{\mathrm{coef}} := \sum_{S \subseteq \{1,\dots,d\}} \widehat{w}_S \cdot |\widehat{\beta}_S^j|.$$

A third option that can be used for a general regression procedure is a permutation based approached. Let $\mathbf{X}_1^{*,j}, \dots, \mathbf{X}_B^{*,j}$ be permuted versions of the data in which the jth coordinate is permuted while the remaining coordinates are fixed. Then, the permutation importance is defined as

$$v_j^{\mathrm{perm}} := \frac{1}{B} \sum_{i=1}^{B} \Big(\frac{\mathrm{RSS}_i^{*,j} - \mathrm{RSS}}{\mathrm{RSS}} \Big),$$

where RSS and $\mathrm{RSS}_i^{*,j}$ are the residual sum of squares of the estimator $\hat{f}_{\mathrm{SR}}$, based on the training data $\mathbf{X}$ and the permuted data $\mathbf{X}_i^{*,j}$, respectively.

Since stabilized regression averages over the sets that are estimated to be generalizable and regression optimal, using any of these variable importance measures should rank variables higher if they indeed are part of a generalizable and regression optimal set. In terms of Section 3, this means that variables in the stable blanket are ranked higher. Similarly, if the stability test cutoff is removed or, equivalently, set to $-\infty$ the variable importance should rank variables higher that are in the Markov blanket. A sensible ranking for whether a variable belongs to the nonstable blanket is thus given by

$$v_j^{\mathrm{SRdiff}} := v_j^{\mathrm{SRpred}} - v_j^{\mathrm{SR}},$$

where v_j^{SR} and v_j^{SRpred} are one of the variable rankings above, based on stabilized regression with and without stability cutoff, respectively.

4.3. *Implementation*. Given a regression procedure, stabilized regression is straightforward to implement and pseudo-code is given in Algorithm 1. The framework is modular and most components, such as stability score, prediction score, variable screening and subsampling of subsets can all be adjusted according to the application at hand.

In Algorithm 1 we added a variable screening step in line 1, since exhaustive subset search becomes infeasible as soon as more than about 15 variables are involved. Instead, we propose to combine a variable screening with subsequent subsampling of predictor sets. Any type of variable screening can be employed, as long as it focuses on selecting predictive variables and removing irrelevant variables. In the linear case, two reasonable approaches would be either plain correlation screening (Fan and Lv (2008)) or an ℓ^1-penalty type screening as, for example, used in the Lasso (Tibshirani (1996)). How many variables to keep after screening depends on the application. In general, our empirical analysis suggested to screen as much as possible without removing any potentially relevant predictors. To make computations feasible after screening, one can additionally subsample subsets randomly. There are several ideas that appear to work well in practice. First, only sample random sets up to a certain size. If one has an idea about how many variables are required to get a stable set (this can often be checked empirically), it empirically seemed to help to sample more sets with this size and less sets with different sizes. Second, the number of subsampled sets should depend both on the expected number of stable and predictive sets and on the number of variables after screening.

Algorithm 1: StabilizedRegression

input : predictor matrix $\mathbf{X}$
response matrix $\mathbf{Y}$
environments $\mathcal{E}^{\text{obs}}$
parameters $\alpha_{\text{pred}}, \alpha_{\text{stab}} \in (0, 1)$

perform variable screening (optional)
select collection of sets $\{S_1, \dots, S_M\}$ (all or subsampled)
for $k \in \{1, \dots, M\}$ **do**
  fit regression function $\hat{f}^{S_k}$
  compute stability score $\mathbf{s}_{\text{stab}}(S)$
  compute prediction score $\mathbf{s}_{\text{pred}}(S)$
end
$\widehat{\mathbb{G}} \leftarrow \{S \in \{S_1, \dots, S_M\} \mid \mathbf{s}_{\text{stab}}(S) \geq \alpha_{\text{stab}}\}$
$c_{\text{pred}} \leftarrow (F^*_{\text{MSE}_Q})^{-1}(1 - \alpha_{\text{pred}})$
$\widehat{\mathbb{O}} \leftarrow \{S \in \widehat{\mathbb{G}} \mid \mathbf{s}_{\text{pred}}(S) \geq c_{\text{pred}}\}$
compute weights $\widehat{w}_S$ according to (4.2)

output: weights $\widehat{w}_S$
regressors $\hat{f}^S$

In our simulations it was often sufficient to subsample about 1000 sets, but, generally, the number should be selected in a data driven fashion, similar to how the number of trees in a random forest (Breiman (2001)) is selected.

In Supplementary Material A (Pfister et al. (2021)), we give a proposal on how to choose default parameters.

5. Numerical simulations. In this section we assess the empirical performance of stabilized regression. We restrict ourselves to the linear model setting, as this is the setting of our biological application. First, in Section 5.1 we consider low-dimensional linear regression and in Section 5.2 high-dimensional sparse linear regression. In both cases we assess how well stabilized regression recovers the sets $\text{SB}_I(Y)$ and $\text{NSB}_I(Y)$ as well as the predictive performance on unseen new environments.

Stabilized regression. Throughout this section we use the implementation of stabilized regression given in Algorithm 1. We consider two versions, both using ordinary least squares as regression but based on different choices of weights $\widehat{w}_S$. First, we use a vanilla version denoted by SR. It uses the mean squared error as prediction score and the p-value of a resampling test using the differences of environmentwise means as test statistic as stability score (see Section 4.1.1). The tuning parameters α_{pred} and α_{stab} are both selected to be 0.01. Second, we use a predictive version, denoted by SRpred. It uses the lowest environmentwise mean squared error as prediction score (again, with $\alpha_{\text{pred}} = 0.01$) and does not include any type of stability score. For both methods we rank the variables according to the score v_j^{coef}, defined in Section 4.2. By construction, we expect SR to rank variables in the stable blanket highest, while SRpred should rank variables in the Markov blanket highest (as long as they are predictive in at least one environment). We combine both procedures to get a further variable ranking, denoted by SRdiff which ranks variables according to $v_j^{\text{SRdiff}} = v_j^{\text{SRpred}} - v_j^{\text{SR}}$, defined in Section 4.2. We expect that this will recover variables in the nonstable blanket. For the high-dimensional example we combine both stabilized regression procedures with ℓ^1 prescreening and screen to 10 variables.

Competing methods. As our simulations are all focused on the linear case, we consider the following linear methods: (i) *Ordinary linear least squares*. This method can only be applied in the low-dimensional setting and will be denoted by OLS. (ii) *ℓ^1-penalized linear regression*, also known as Lasso (Tibshirani (1996)), is a regularized version of linear regression that is often employed to high-dimensional problems. We select the penalty parameter, based on cross-validation, and denote the method by Lasso. (iii) *Anchor regression* (Rothenhäusler et al. (2021)) which explicitly incorporates heterogeneity. We consider two versions, one for the low-dimensional case, based on OLS, and one for the high-dimensional case, based on Lasso, denoted by AR and AR (Lasso), respectively. The tuning parameter for both is based on an environment-wise cross-validation. (iv) *Instrumental variables regression*, which allows to guard against arbitrary shift strengths. We compute it via the anchor regression estimate, based on a penalty parameter of $\gamma = 1000$. As with anchor regression, there will be two versions, based either on OLS or Lasso, denoted by IV and IV (Lasso), respectively. For each method we get a variable importance measure by taking the scaled regression parameter. All methods, except IV, should recover the Markov blanket. On the other hand, in our simulation settings IV should recover the stable blanket (see Section 3.3), given a sufficient sample size and strong enough interventions.

5.1. *Low-dimensional linear regression*. In our first numerical experiment we consider a standard low-dimensional linear SCM. We want to assess both the predictive generalization performance as well as the variable selection. To this end, we simulate 1000 data sets, according to Simulation 1, and apply stabilized regression and all competing methods to each.

SIMULATION 1 (Low-dimensional linear regression). Randomly sample a DAG with $d = 11$ variables as follows: (i) Sample a causal ordering by randomly permuting the variables. (ii) Iterate over the variable and sample for each variable at most four parents from all variables with higher causal ordering. Next, select a random node to be the response Y, extend the DAG by randomly sampling four variables from the remaining $d - 1$ variables and add a parent intervention node I to each of them. Denote the adjacency matrix of the resulting DAG by B, that is, $B_{i,j} \neq 0$ if and only if there is an edge from node i to node j. For each nonzero entry in B, sample an edge weight uniformly from $(-1.5, -0.5) \cup (0.5, 1.5)$. Based on this DAG, generate data from different environments consisting of random mean shifts in the noise of the intervention variables. The random mean shifts are sampled differently, depending on whether the environment is used for training or for testing. Specifically, for training the mean shift is sampled uniformly from $(-1, 1)$ and for testing it is sampled uniformly from $(-10, 10)$. Based on these settings, sample five training and 10 testing environments, each consisting of $n = 250$ observations using Gaussian noise. More specifically, for each environment e generate data according to $\mathbf{X}_e = (\mathrm{Id} - B)^{-1}\boldsymbol{\varepsilon}_e$, where $\boldsymbol{\varepsilon}_e \in \mathbb{R}^{n\times(d+1)}$ and each row is sampled multivariate normal with covariance matrix $0.25 \cdot \mathrm{Id}$ and mean vector μ which specifies the random mean shift for the intervention variables and is zero everywhere else.

The prediction performance (in terms of mean residual sum of squares) on the testing environments is given in Figure 5. The 1000 repetitions are split, depending on whether $\mathrm{MB}(Y) = \mathrm{SB}_I(Y)$ or $\mathrm{MB}(Y) \neq \mathrm{SB}_I(Y)$ (542 repetitions in the first and 458 repetitions in the second case). In the case that $\mathrm{MB}(Y) = \mathrm{SB}_I(Y)$, we expect all procedures to perform similarly, as all prediction methods should be generalizable in this case. Only the IV method performs slightly worse, which is expected, since it generally is an estimator with higher variance. On the other hand, in the case $\mathrm{MB}(Y) \neq \mathrm{SB}_I(Y)$ not all methods generalize to the training method. Only SR and IV are expected to be generalizable in this case. However, IV

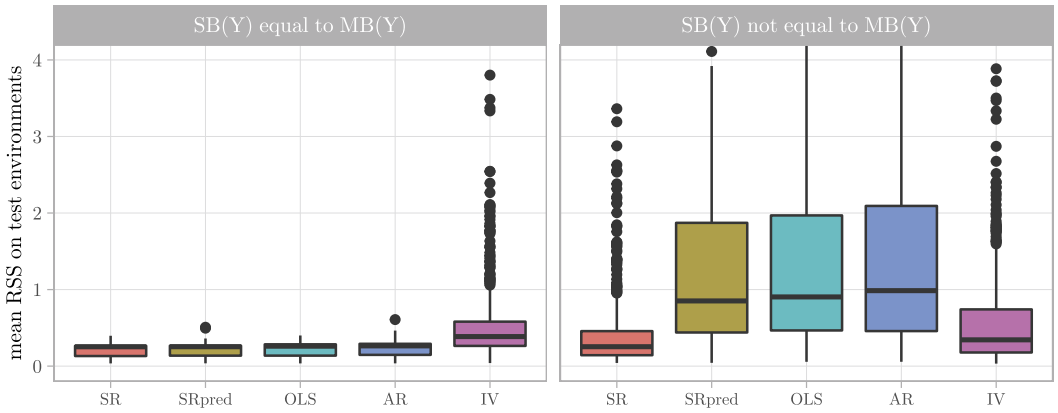


FIG. 5. *Prediction results for* 1000 *repetitions from Simulation* 1. *SR performs well both when* $\mathrm{MB}(Y) = \mathrm{SB}_I(Y)$ (542 *repetitions*) *and when* $\mathrm{MB}(Y) \neq \mathrm{SB}_I(Y)$ (458 *repetitions*). *Apart from SR and IV, no other method is expected to generalize. The different performance between SR and IV is a finite sample property and shows that averaging can outperform direct optimization of* (2.4).

again performs worse than SR. The reason that AR does not generalize here is that the test shifts are chosen to be stronger than the training environments. It, therefore, is not able to guard against these types of shifts.

At first sight it might seem surprising that the performance of SR (and also IV) is substantially worse when $\mathrm{MB}(Y) \neq \mathrm{SB}_I(Y)$, compared to when $\mathrm{MB}(Y) \neq \mathrm{SB}_I(Y)$. The reason is that in these cases we need to capture two types of signals: (i) predictiveness of a set of predictors (is a set good at explaining the response) and (ii) stability of a set of predictors (does a set lead to the same model across all environments). The second signal type requires sufficient heterogeneity in the data and is substantially harder to detect. In contrast, if $\mathrm{MB}(Y) = \mathrm{SB}_I(Y)$, it is sufficient to only detect signal type (i). The outliers in the boxplot for SR in Figure 5 correspond to settings in which SR was not able to correctly distinguish between stable predictive sets and nonstable predictive sets. Whenever there are sufficiently strong interventions in the training environments, SR will perform similarly well in the test as in the training environment independent of the intervention strength in the test environment. In Supplementary Material A (Pfister et al. (2021)) (Section F), we show that similar results are obtained when choosing a different stability threshold.

Based on Simulation 1, we can compute the ground truth sets $\mathrm{MB}(Y)$, $\mathrm{SB}_I(Y)$ and $\mathrm{NSB}_I(Y)$ and check how well each method recovers each of these sets. To this end, we compute true and false positive rates for each method, based on its variable importance ranking. Results are given in Figure 6, where we only consider the 386 cases of the 1000 repetitions for which $\mathrm{SB}_I(Y) \neq \varnothing$ and $\mathrm{NSB}_I(Y) \neq \varnothing$. The prediction performance on this subset of the data is very similar to Figure 5 (right) and is given in Supplementary Material A (Pfister et al. (2021)) (Section F). As one would expect from the prediction results, SR outperforms the other methods in terms of recovering the stable blanket. Since SR down-weights variables in the nonstable blanket, it is not expected to recover $\mathrm{MB}(Y)$ and $\mathrm{NSB}_I(Y)$ well. However, SRpred is better in recovering the Markov blanket (comparable with OLS), and hence SRdiff allows good recovery of the $\mathrm{NSB}_I(Y)$. As expected, AR and OLS both are good at recovering $\mathrm{MB}(Y)$. However, they perform bad in terms of recovery of both $\mathrm{SB}_I(Y)$ and $\mathrm{NSB}_I(Y)$ and hence themselves do not allow to distinguish between them. IV, on the other hand, solves the same optimization as SR and hence aims at recovering $\mathrm{SB}_I(Y)$. Similarly, it also down-ranks variables from $\mathrm{NSB}_I(Y)$ but is not quite as good as SR in this respect.

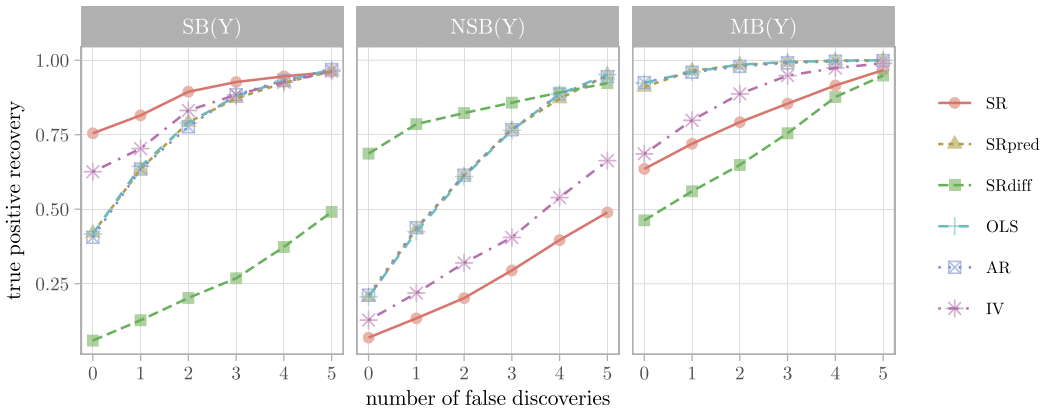


FIG. 6. *Recovery performance based on* 386 *repetitions* (*only using repetitions with* $\mathrm{SB}_I(Y) \neq \varnothing$ *and* $\mathrm{NSB}_I(Y) \neq \varnothing$) *from Simulation* 1. *Each of the different versions of stabilized regression recovers one set well*: *SR has the best recovery of* $\mathrm{SB}_I(Y)$, *SRdiff has the best recovery of* $\mathrm{NSB}_I(Y)$ *and SRpred performs competitive in recovering* $\mathrm{MB}(Y)$.

5.2. *High-dimensional linear regression*. To illustrate that stabilized regression adapts to high-dimensional settings, we consider the high-dimensional linear simulation described in Simulation 2. We look at both prediction and variable selection properties of all methods. Results are given in Figure 7 and Figure 8 and substantiate the conclusions drawn in Section 5.1.

SIMULATION 2 (High-dimensional linear regression). Randomly sample a DAG with $d = 1001$ variables as follows: (i) Sample a causal ordering by randomly permuting the variables. (ii) From the full graph, based on this causal order, select edges with a probability of $p = 2/(d-1)$, so the expected number of edges is d. Fix the first variable to be the response Y, and denote the adjacency matrix of the resulting DAG by B, that is, $B_{i,j} \neq 0$ if and only if there is an edge from node i to node j. For each nonzero entry in B, sample an edge weight uniformly from $(-1.5, -0.5) \cup (0.5, 1.5)$. Based on this DAG, generate data from different environments, consisting of random mean shifts on a subset of the children of Y which is selected by randomly choosing each child with probability $q = 0.9$. The mean shifts are

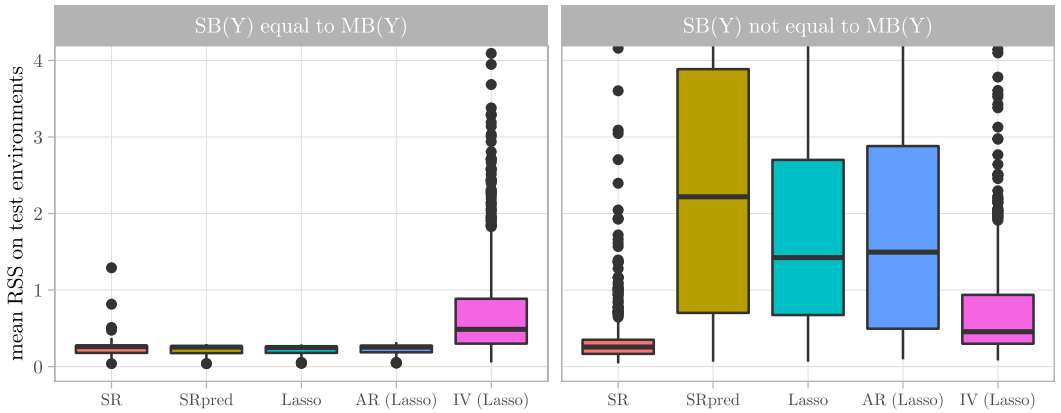


FIG. 7. *Prediction results from* 1000 *repetitions from Simulation* 2. *SR performs well both when* $\mathrm{MB}(Y) = \mathrm{SB}_I(Y)$ (643 *repetitions*) *and when* $\mathrm{MB}(Y) \neq \mathrm{SB}_I(Y)$ (357 *repetitions*). *Apart from IV, no other method is expected to generalize to these settings. The different performance between IV and SR is even more pronounced in the high-dimensional settings*.

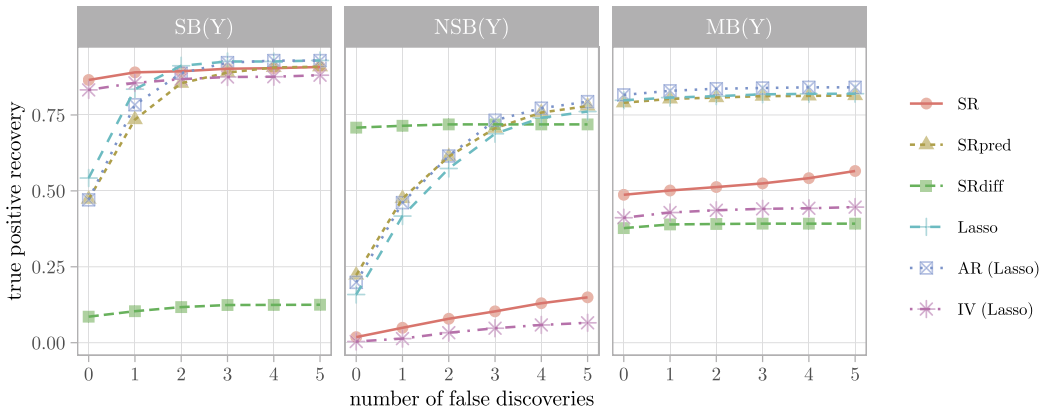


FIG. 8. *Recovery performance based on* 248 *repetitions* (*only using repetitions with* $\mathrm{SB}_I(Y) \neq \varnothing$ *and* $\mathrm{NSB}_I(Y) \neq \varnothing$) *from Simulation* 2. *Each of the different versions of stabilized regression recovers one set well*: *SR has the best recovery of* $\mathrm{SB}_I(Y)$, *SRdiff has the best recovery of* $\mathrm{NSB}_I(Y)$ *and SRpred performs competitive in recovering* $\mathrm{MB}(Y)$.

sampled differently, depending on whether the environment is used for training or for testing. Specifically, for training it is sampled uniformly from $(-1, 1)$, and for testing it sampled uniform from $(-10, 10)$. Based on these settings, sample five training and 10 testing environments, each consisting of $n = 100$ observations using Gaussian noise. More specifically, for each environment e, generate data according to $\mathbf{X}_e = (\mathrm{Id} - B)^{-1}\boldsymbol{\varepsilon}_e$, where $\boldsymbol{\varepsilon}_e \in \mathbb{R}^{n\times(d+1)}$ and each row is sampled multivariate normal with covariance matrix $0.25 \cdot \mathrm{Id}$ and mean vector μ which specifies the random mean shift for the children that are intervened on and is zero everywhere else.

6. Application to biological pathway analysis. In our biological application we aim to generate novel biological hypotheses about gene function. More specifically, we are interested in two types of questions: (1) If we examine canonical metabolic pathways, can we identify novel gene relationships interacting with the known pathway and (2) can we classify gene targets by whether they have a fixed or switching functional dependence on a pathway's activity depending on the environment. To answer these questions, we propose applying two versions of stabilized regression and visualizing the results as in Figure 2. The following steps describe the procedure:

1. *Input*: A response variable Y, representing a quantity of interest (e.g., average activation levels of a pathway), a collection of gene expression levels $X^1, \ldots, X^d$ and an environment variable E indicating different conditions in which the data have been recorded.
2. *Stabilized regression*: Compute the following two versions of stabilized regression:

   (a) SR: Use the p-value of a stability test as stability score and pooled mean squared prediction error as prediction score.

   (b) SRpred: Use the minimum environmentwise mean squared prediction error as prediction score and no stability cutoff.

   In both cases, we propose a correlation prescreening to screen to approximately $\min_e \frac{n_e}{2}$ variables and a subsampling of subsets of a fixed maximum size (see Section 4.3).
3. *Variable importance*: Based on these two versions of stabilized regression, compute variable importance scores v_j^{SR}, v_j^{SRpred} and v_j^{SRdiff}, using one of the variable importance measures from Section 4.2.

4. *Stability selection*: Use stability selection (Meinshausen and Bühlmann (2010)) to compute selection probabilities for the two selection criteria $v_j^{\mathrm{SR}} > 0$ and $v_j^{\mathrm{SRdiff}} > 0$. This introduces sample stability into the estimates, hence increasing reliability of the results.
5. *Visualization*: Plot the two types selection probabilities on different axes (x-axis: SRdiff, y-axis: SR).

The resulting plot visualizes the relation of all predicting genes with the response. It allows explicitly distinguishing between genes that have a stable functional dependence with the response across all environments and genes that are predictive but have a functional shift with respect to the response across environments. The stability selection procedure adds a theoretical guarantee on the false discovery rate which can be selected by practitioners (green regions in the plot correspond to the threshold at which the expected number of wrongly selected variables is at most 1).

In the following sections, we apply stabilized regression to the systems biology application discussed in Section 1.

DATA SET (Biological pathway analysis). The data set is due to Roy et al. (2019), Williams et al. (2020), Čuklina et al. (2021) and consists of multiomic data from the transcriptome and proteome of a mouse population of 57 different inbred strains that was split into two groups, fed either with a low fat or a high fat diet. Liver tissue from these cohorts was then collected at multiple timepoints across their natural lifespans, providing diet as an independent biological (environment) variable. In the following application we work with two parts of this data: (1) Proteomic data consisting of $d = 3939$ measured genes from $n = 315$ mice of which 150 had a high and 165 a low fat diet. (2) Transcriptomic data consisting of $d = 25,391$ measured from $n = 291$ mice of which 129 had a high and 162 a low fat diet. The preprocessed data is part of the Supplementary Material B (Pfister et al. (2021)).

To assess the variable selection performance of stabilized regression on this data set, we first benchmark our method with other common approaches used to find functionally related genes (Section 6.1). Second, we discuss whether our proposed method and visualization procedure is able to distinguish between stable and unstable dependencies (Section 6.2). As there are only two environments (high fat/low fat diet), it is not feasible to evaluate the predictive performance of stabilized regression on this data set.

In all of the following experiments, we use a stability score based on the Chow test and set the cutoff parameters to $\alpha_{\mathrm{pred}} = 0.01$ and $\alpha_{\mathrm{stab}} = 0.1$. Furthermore, we use correlation pre-screening to screen to 50 variables, subsample 5000 subsets, consisting of at most six variables, and use v_j^{coef} as variable importance measure.

6.1. *Gene recovery.* Validation on real data is often difficult and can only be as good as the ground truth known about the underlying system. Here, as a rough approximation, we assume that genes belonging to the same canonical metabolic pathways are functionally closer than genes not belonging to the same pathway (Francesconi et al. (2008)). Furthermore, data-driven network approaches to functional gene annotation have proven successful in independent de novo reconstitution of functional gene ontology sets which have been curated over decades through molecular experimentation (Dutkowski et al. (2013)). This assumption is key to any correlation-based discovery approach in biology and is known to be particularly well satisfied in larger protein complexes (Roumeliotis et al. (2017)). Our validation is based on taking a set of genes from known metabolic pathways, iteratively taking each of these genes as a response Y, and then observing how many canonical genes from the known pathway are recovered. We selected seven pathways for this analysis taken from the KEGG database (Kanehisa and Goto (2000)) and the Reactome Pathway Knowledgebase (Fabregat

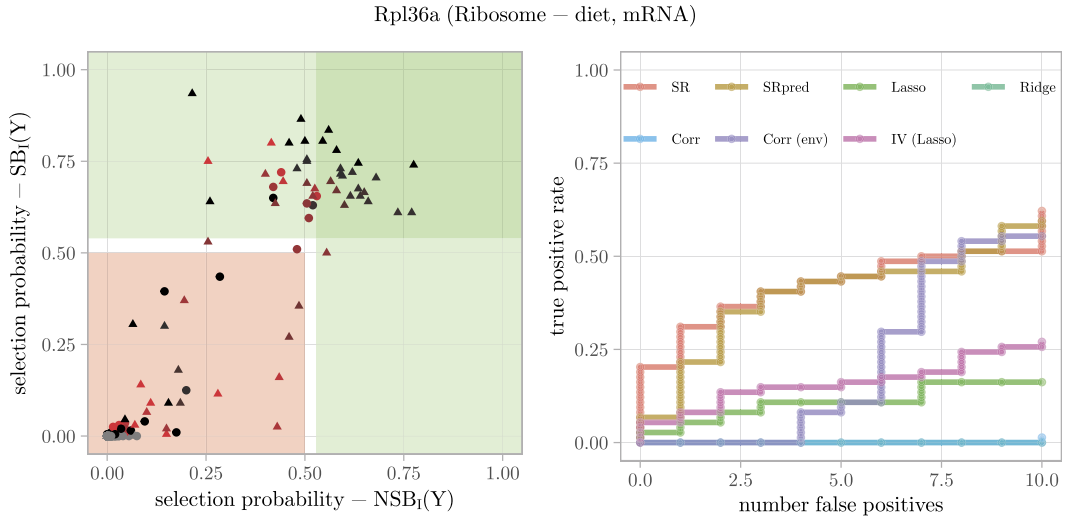


FIG. 9. *Recovery analysis for gene Rpl36a from the* Ribosome *pathway. (Left) Visualization for Rpl36a as response and all remaining genes as potential predictors. Canonical* Ribosome *genes are marked with a triangle all other genes with a circle. Many correct genes are ranked high. (Right) pROC for different methods, where canonical* Ribosome *genes are considered true positives and all other genes false positives.*

et al. (2017)). More details are given in Supplementary Material A (Pfister et al. (2021)). In our analysis we use diet (low-fat vs. high-fat) as an environment variable. The result of applying the procedure described above to a single gene from the *Ribosome* pathway results in Figure 9 (left); Figure 2 shows the same analysis for a different pathway. To visualize which other genes belong to this pathway, we have drawn these genes as triangles. We compare the recovery performance of our method with the following competing variable selection methods: (i) *Corr*: pairwise correlation on the pooled data (including both diets), (ii) *Corr (env)*: maximum of the pairwise correlation on each diet individually, (iii) *Lasso*: ℓ^1-penalized regression, (iv) *Ridge*: ℓ^2-penalized regression and (v) *IV (Lasso)*: a Lasso based version of anchor regression with $\gamma = 1000$. The performance is then assessed by computing partial receiver operator curves (pROC) with up to 10 false positives, as shown in Figure 9 (right). We did this for all genes from the pathway and summarized the resulting pROCs using the normalized area under these curves, called pAUC (partial area under the receiver operator curve). The results for the *Ribosome* pathway are shown in Figure 10.

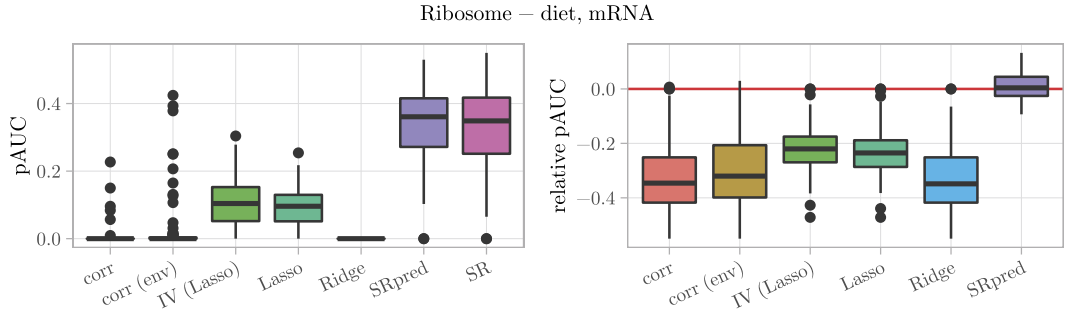


FIG. 10. *Recovery analysis with mRNA data based on the* Ribosome *pathway using diet as an environment variable. (Left) Box plot of pAUC values for recovery of different genes belonging to the Ribsome pathway. (Right) Relative difference of each pAUC value compared to SR. Values below* 0 *imply worse pAUC, compared to SR. Stabilized regression outperforms the competing methods.*

The results for all seven pathways (both for mRNA and protein data) are given in Supplementary Material A (Pfister et al. (2021)). While in many cases the results are not as pronounced as for the Ribsome pathway, one can see that, in most cases, stabilized regression performs at least as good as other competitors and often better. The differences between methods is less obvious for protein data for which basic correlation screening often performs very well. We believe this might be due to the fact that proteins are one step closer to the biological processes, and hence these measurements capture the functional relations more directly.

6.2. *The advantage of intervention stability.* A key advantage of our method is that it allows to group genes based on whether their dependence on the response is stable or unstable with respect to some exogenous environment variable. We illustrate this with Figure 2 (and Figure 9 (left)). The green region of significant findings can be divided into three parts that should be interpreted differently. The first region is the top left area of the plot. Genes that appear there are detected only by SR and not by SRpred which implies that they might not be the most predictive genes but depend on the response in a stable fashion across all environments. The second region is the bottom right part of the plot. These genes are only found by SRpred and not by SR. This means that they are strongly predictive for the response in at least one of the environments, but the dependence with the response changes across environments. Finally, the third area is the top right corner of the plot in which the green areas overlap. Genes in this area are significantly reduced in importance in SR, compared to SRpred but still remain significant in terms of SR. This can happen if the stability cutoff is not consistently removing the same genes in all cases which means that the variations across environments are not sufficiently strong to distinguish whether these genes are stable or unstable. While no conclusion can be drawn on whether these genes are stable or unstable, they can be considered to be predictive for the response.

7. Discussion. We propose a regression framework for multi-environment settings. Our novel algorithm, stabilized regression, averages over regression estimates, based on subsets of predictors, in order to regularize the final predictions to be both predictive and stable across environments. We relate this setting to causal models and prove that, under mild conditions, there exists an optimal subset of predictors called the stable blanket which generalizes across environments while minimizing the mean squared prediction loss. Furthermore, we show that one can separate the Markov blanket into the stable blanket and the nonstable blanket which allows to characterize predictive variables by whether they have a stable or unstable functional dependence on the response. Using this framework, we propose a procedure that assists hypothesis generation in systems biology and demonstrate its usefulness on a current multiomic data set. The procedure is shown to perform well in terms of recovery on known biological pathways and, additionally, allows to separate findings into stable and unstable predictors.

While our framework can be combined with any regression procedure, we focus on the case of linear models. Future research should, therefore, assess how these ideas perform on nonlinear regression problems. In those settings, one needs to be more careful about how to deal with shift interventions, since extrapolation might not be well defined anymore. A further interesting direction would be to consider different notions of stability, other than the one considered here, based on the conditional invariance defined in (2.1).

Acknowledgements. We thank Yuansi Chen and Nicolai Meinshausen for helpful discussions. Most of this work was done while NP and EGW were at ETH Zürich.

Funding. The first and fifth authors were supported by European Research Council (CausalStats—AdG grant 786461). The second author was supported by an NIH F32 Ruth Kirchstein Fellowship (F32GM119190). The third author was supported by VILLUM FONDEN research grant 18968 and the Carlsberg Foundation. The fourth author was supported by European Research Council (Proteomics4D—AdG grant 670821 and Proteomics v3.0—AdG grant 233226) and by Swiss National Science Foundation (31003A-140780, 31003A-143914, CSRII3-136201, 310030E_173572 and 310030B_185390).

SUPPLEMENTARY MATERIAL

Supplement A: Supporting material (DOI: 10.1214/21-AOAS1487SUPPA; .pdf). Provides supporting material consisting of proofs, proposed default settings, gene annotations for cholesterol biosynthesis example, details on pathways and complete results for biological pathway analysis.

Supplement B: Pre-processed biological data set and R code (DOI: 10.1214/21-AOAS1487SUPPB; .zip). Contains biological data and all R code.

REFERENCES

ALDRICH, J. (1989). Autonomy. *Oxf. Econ. Pap.* **41** 15–34.

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BÜHLMANN, P. (2020). Invariance, causality and robustness: 2018 Neyman Lecture. *Statist. Sci.* **35** 404–426. MR4148216 https://doi.org/10.1214/19-STS721

BURNHAM, K. and ANDERSON, D. (1998). Practical use of the information-theoretic approach. In *Model Selection and Inference* 75–117. Springer.

CANNINGS, T. I. and SAMWORTH, R. J. (2017). Random-projection ensemble classification. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 959–1035. With discussions and a reply by the authors. MR3689307 https://doi.org/10.1111/rssb.12228

CHOW, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28** 591–605. MR0141193 https://doi.org/10.2307/1910133

CONSTANTINOU, P. and DAWID, A. P. (2017). Extended conditional independence and applications in causal inference. *Ann. Statist.* **45** 2618–2653. MR3737904 https://doi.org/10.1214/16-AOS1537

ČUKLINA, J., LEE, C. H., WILLIAMS, E. G., SAJIC, T., COLLINS, B. C., RODRIGUEZ MARTINEZ, M., SHARMA, V. S., WENDT, F., GOETZE, S., KEELE, G. R. et al. (2021). Molecular systems biology. Batch effects in large-scale proteomics studies: diagnostics and correction.

DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *Int. Stat. Rev.* **70** 161–189.

DUTKOWSKI, J., KRAMER, M., SURMA, M. A., BALAKRISHNAN, R., CHERRY, J. M., KROGAN, N. J. and IDEKER, T. (2013). A gene ontology inferred from molecular networks. *Nat. Biotechnol.* **31** 38–45.

FABREGAT, A., JUPE, S., MATTHEWS, L., SIDIROPOULOS, K., GILLESPIE, M., GARAPATI, P., HAW, R., JASSAL, B., KORNINGER, F. et al. (2017). The reactome pathway knowledgebase. *Nucleic Acids Res.* **46(D1)** D649–D655, 11. https://doi.org/10.1093/nar/gkx1132

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

FRANCESCONI, M., REMONDINI, D., NERETTI, N., SEDIVY, J. M., COOPER, L. N., VERONDINI, E., MILANESI, L. and CASTELLANI, G. (2008). Reconstructing networks of pathways via significance analysis of their intersections. *BMC Bioinform.* **9** 9.

GANIN, Y., USTINOVA, E., AJAKAN, H., GERMAIN, P., LAROCHELLE, H., LAVIOLETTE, F., MARCHAND, M. and LEMPITSKY, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17** Paper No. 59, 35. MR3504619

HAAVELMO, T. (1944). The probability approach in econometrics. *Econometrica* **12 S** 118. MR0010953 https://doi.org/10.2307/1906935

HEINZE-DEML, C. and MEINSHAUSEN, N. (2021). Conditional variance penalties and domain shift robustness. *Mach. Learn.* **110** 303–348. MR4207502 https://doi.org/10.1007/s10994-020-05924-1

HEINZE-DEML, C., PETERS, J. and MEINSHAUSEN, N. (2018). Invariant causal prediction for nonlinear models. *J. Causal Inference* **6**.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. MR1765176 https://doi.org/10.1214/ss/1009212519

HOOVER, K. D. (1990). The logic of causal inference. *Econ. Philos.* **6** 207–234.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

KANEHISA, M. and GOTO, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** 27–30. https://doi.org/10.1093/nar/28.1.27

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523 https://doi.org/10.1111/j.1467-9868.2010.00740.x

PAN, S., TSANG, I., KWOK, J. and YANG, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **22** 199–210.

PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

PELLET, J.-P. and ELISSEEFF, A. (2008). Using Markov blankets for causal structure learning. *J. Mach. Learn. Res.* **9** 1295–1342. MR2426044

PERRONE, M. and COOPER, L. (1992). When networks disagree: Ensemble methods for hybrid neural networks. Technical report, Brown Univ., Providence RI, Institute for Brain and Neural Systems.

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. With comments and a rejoinder. MR3557186 https://doi.org/10.1111/rssb.12167

PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3822088

PFISTER, N., BAUER, S. and PETERS, J. (2019). Learning stable and predictive structures in kinetic systems. *Proc. Natl. Acad. Sci. USA* **116** 25405–25411. MR4047351 https://doi.org/10.1073/pnas.1905688116

PFISTER, N., BÜHLMANN, P. and PETERS, J. (2019). Invariant causal prediction for sequential data. *J. Amer. Statist. Assoc.* **114** 1264–1276. MR4011778 https://doi.org/10.1080/01621459.2018.1491403

PFISTER, N., WILLIAMS, E. G., AEBERSOLD, R. and BÜHLMANN, P. (2021). Supplement to "Stabilizing variable selection and regression." https://doi.org/10.1214/21-AOAS1487SUPPA, https://doi.org/10.1214/21-AOAS1487SUPPB

RICHARDSON, T. and ROBINS, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Center for the Statistics and the Social Sciences, Univ. Washington Series. Working Paper 128, 30 April 2013.

ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. and PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** Paper No. 36, 34. MR3862443

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2021). Anchor regression: Heterogeneous data meet causality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 215–246. MR4250274 https://doi.org/10.1111/rssb.12398

ROUMELIOTIS, T. I., WILLIAMS, S. P., GONÇALVES, E., ALSINET, C., DEL CASTILLO VELASCO-HERRERA, M., ABEN, N., GHAVIDEL, F. Z., MICHAUT, M., SCHUBERT, M. et al. (2017). Genomic determinants of protein abundance variation in colorectal cancer cells. *Cell Rep.* **20** 2201–2214.

ROY, S., SLEIMAN, M. B., JHA, P., WILLIAMS, E. G., INGELS, J. F., CHAPMAN, C. J., MCCARTY, M. S., HOOK, M., SUN, A. et al. (2019). Modulation of longevity by diet, and youthful body weight, but not by weight gain after maturity. Preprint bioRxiv:776559.

SCHÖLKOPF, B., JANZING, D., PETERS, J., SGOURITSA, E., ZHANG, K. and MOOIJ, J. M. (2012). On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)* 1255–1262. Omnipress.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHAH, R. D. and BÜHLMANN, P. (2018). Goodness-of-fit tests for high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 113–135. MR3744714 https://doi.org/10.1111/rssb.12234

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WANG, S., NAN, B., ROSSET, S. and ZHU, J. (2011). Random Lasso. *Ann. Appl. Stat.* **5** 468–485. MR2810406 https://doi.org/10.1214/10-AOAS377

WILLIAMS, E. G., PFISTER, N., ROY, S., STATZER, S., INGELS, J., BOHL, C., HASSAN, M., ČUKLINA, J., BÜHLMANN, P. et al. (2020). Multi-omic profiling of the liver across diets and age in a diverse mouse population. Preprint bioRxiv. Available at https://www.biorxiv.org/content/10.1101/2020.08.20.222968v2.

WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20** 557–585.

YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. MR3102560 https://doi.org/10.3150/13-BEJSP14

YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

ZHANG, K., SCHÖLKOPF, B., MUANDET, K. and WANG, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning* (*ICML*) 819–827.