International Statistical Review

International Statistical Review (2020), 88, S1, S114–S134 doi:10.1111/insr.12426

Deconfounding and Causal Regularisation for Stability and External Validity

Peter Bühlmann and Domagoj Ćevid

Seminar for Statistics, ETH Zürich, Zürich, Switzerland
E-mail: buhlmann@stat.math.ethz.ch

Summary

We review some recent works on removing hidden confounding and causal regularisation from a unified viewpoint. We describe how simple and user-friendly techniques improve stability, replicability and distributional robustness in heterogeneous data. In this sense, we provide additional thoughts on the issue of concept drift, raised recently by Efron, when the data generating distribution is changing.

Key words: anchor regression; causality; hidden confounding; high-dimensional models; instrumental variables regression; Lasso; structural equation model.

1 Introduction

Brad Efron, in his lecture at the occasion of receiving the International Prize in Statistics, brought up some fascinating thoughts on 'prediction, estimation and attribution', with particular attention to the new 'wide data era' that has entered statistics and data science more generally (Efron, 2019, 2020). Looking back almost 20 years ago, there has been a huge development in statistics since Leo Breiman's article 'Statistical Modeling: The Two Cultures' (Breiman, 2001). Even more broadly, data science has become an emerging new field and profession. It deals with information extraction from data, often in close proximity with other sciences. Its historical roots are in statistics, and statistical 'critical' thinking plays an ever important role in inference from data to models and prediction. There are many interesting facets of this broad topic, see, for example, David Donoho's '50 years of Data Science' (Donoho, 2017) or Bin Yu's 'Veridical Data Science' (Yu & Kumbier, 2020).

Efron (2019, 2020) has formulated intriguing ideas on prediction, estimation and attribution. We are presenting here a few additional considerations on the topic, as outlined in sections 1.1 and 1.2.

1.1 Stability of Predictions and Causal Thinking in Presence of Perturbations: Efron and Cox in Response to Breiman (2001)

Breiman (2001) argued strongly in favor of prediction and the corresponding feature importance measures. However, prediction in reality is often more subtle than the usual textbook definition where one assumes the same data generating mechanism for the training and the new test set data.

The illustration by Efron (2020) of concept drift where the data-generating distribution changes between training and test set, or also his question 'Were the test sets really a good test?' (Efron, 2019), nicely emphasises that prediction can be 'highly context-dependent and fragile': he illustrates with a certain data set that training on the first part of the observations and using the last ones as the test set give a widely different answer for the error rate than the average of taking many random divisions into training-data and test-data. Apparently, the last observations in the data set seem to have a rather different data generating distribution than the first ones from the training phase.

Similarly, Cox (2001) wrote in a response to Breiman's article wrote: '. . . Key issues are then the stability of the predictor as practical prediction proceeds, the need from time to time for recalibration and so on. However, much prediction is not like this. Often the prediction is under quite different conditions from the data; what is the likely progress of the incidence of the epidemic of v-CJD in the United Kingdom, what would be the effect on annual incidence of cancer in the United States of reducing by 10% the medical use of X-rays, etc.? That is, it may be desired to predict the consequences of something only indirectly addressed by the data available for analysis. As we move toward such more ambitious tasks, prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more and more tentative. Formulation of the goals of analysis solely in terms of direct prediction over the data set seems then increasingly unhelpful.'

Whereas Efron (2001) wrote in return to Breiman's article: 'Estimation and testing are a form of prediction: "In our sample of 20 patients drug A outperformed drug B; would this still be true if we went on to test all possible patients?" . . . (Peter Gregory) undertook his study for prediction purposes, but also to better understand the medical basis of hepatitis. Most statistical surveys have the identification of causal factors as their ultimate goal.'

In this paper, we build on the fact that stability of prediction and causality are naturally connected. As a result, new methods and algorithms emerge, which are easy to use and fairly 'automatic'. They will not replace careful statistical thinking, for example, in the way Cox (2001) describes it above. But they often act, in quite a few scenarios, more intelligently than plain vanilla 'black box' prediction algorithms: perhaps, such and many other new algorithms close to some extent the gap between 'the two cultures' from Breiman (2001). This is somewhat in line with Brad Efron's statements in his International Prize in Statistics lecture (Efron, 2019), namely, 'Two Trends: Making prediction algorithms better for scientific use' and 'Making traditional estimation/attribution methods better for large-scale $(n; p)$ problems'.

1.2 External Validity, Distributional Replicability, Robustness and Connections to Causality

One major problem with many modern algorithms and methods is their vulnerability to distributional changes in new data. Would we see a good amount of replication in a new study, or in a new environment? Can we do accurate prediction and estimation in changing scenarios? These questions tie in to some of the points raised by Efron (2020) and mentioned above, namely, about concept drift ('Were the test sets really a good test?', Efron, 2019), or to the comments by Cox (2001) that 'direct prediction over the data set seems then increasingly unhelpful'. They both refer to external validity and generalisation beyond the observed data. The latter is well understood if the future external data have the same generating distribution as the observed training data, but if not, external validity relates to distributional robustness (Sinha *et al.*, 2017; Gao *et al.*, 2017; Meinshausen, 2018), transfer learning (Pratt, 1993; Pan & Yang, 2010) and causality (Dawid & Didelez, 2010; Peters *et al.*, 2016; Bareinboim & Pearl, 2016; Rojas-Carulla *et al.*, 2018; Rothenhäusler *et al.*, 2018; Bühlmann, 2020; Dawid, 2020).

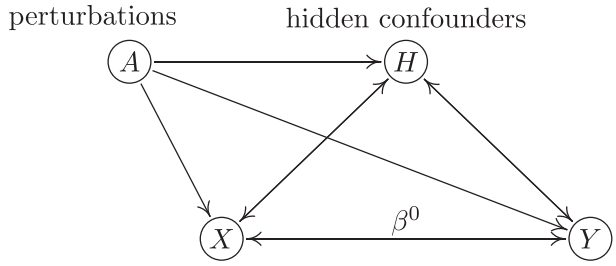

Figure 1. *The generic problem. The goal is inferring the regression parameter β^0, describing the relation between (the causal components of) X and Y. Additionally, there are hidden confounding variables H, and perturbations generated by observed external (exogenous) variables A are present. The cases without and with A are discussed in Sections 2 and 3, respectively. The directionality among different variables may be unknown. The variables X, H, and A can be multivariate but for simplicity, Y is univariate. The graph corresponds to the structure of a structural equation model: the arrows are bidirected, saying that the directions between some of the components can go either way and feedback loops are allowed as well.*

1.3 The Current Work

We review some of our more recent contributions on deconfounding, distributional robustness and replicability, and causality (Rothenhäusler *et al.*, 2018; Bühlmann, 2020; Ćevid *et al.*, 2018; Guo *et al.*, 2020). A unified treatment might enable us to clarify the connections more clearly. We aim for simplicity, demonstrating that at least some of the ideas and methods are simple and easy to use, yet they seem to be effective in achieving some form of distributional robustness. The latter term is rather different from the more standard formulation and procedures in robust statistics (Huber, 1964; Hampel *et al.*, 1986), where outliers occur in the training data and unlike test set distributional changes examined here.

The generic problem we are considering is loosely illustrated in Figure 1. We are interested in inferring the unconfounded regression parameter β^0 and in stable prediction of Y from X. We do not observe all the relevant variables and are thus confronted with hidden confounding. This scenario is discussed in Section 2. Additionally, we may observe data under various perturbations that are generated by external (exogenous) variables A, as discussed in Section 3. The graph in Figure 1 corresponds to a structural equation model (Bollen, 1989; Pearl, 2009), introduced in Equation (1) or (15) more formally. Of particular interest is the univariate response Y_i and its linear regression function of some of the components of a $(1 \times p)$-dimensional covariate X_i, where i denotes the i-th observation:

$$Y_i \leftarrow X_i\beta^0 + g(H_i, A_i) + \varepsilon_{Y,i},$$

where $\varepsilon_{Y,i}$ is a noise or innovation term being independent of all the variables arising 'earlier' or 'up-stream' of Y_i; the exogenous variables A_i are non-existent in our discussion in Section 2. The symbol '$\leftarrow$' is algebraically an equality sign. The variables corresponding to the support of β^0 are the *causal X-variables for Y*, because they are the only components of X_i, which directly enter the structural equation for Y_i. Thus, $\beta^0_j \neq 0$ if and only if the j-th component of X corresponds to a causal X-variable. More precise definitions of the model versions are given later.

1.3.1 A connection to causality

There is a fruitful link to causality. In a nutshell, one can represent the causal parameter β^0 (or also the function f^0) as the minimiser of a worst case risk such as

$$\beta^0 = \text{argmin}_\beta \max_{P \in \mathcal{P}} \mathbb{E}_P[(Y_i - X_i\beta)^2],$$

for particular classes of distributions $\mathcal{P}$ for (X, Y). Such a class can be thought as containing various perturbations of the original data generating distribution, and hence, there is an intrinsic connection between causality and distributional robustness (Dawid & Didelez, 2010; Peters *et al.*, 2016; Rojas-Carulla *et al.*, 2018; Rothenhäusler *et al.*, 2018; Bühlmann, 2020; Dawid, 2020). In this paper, we will not elaborate much on the causal interpretation; however, the operational procedures that have causal interpretability can be 'simply' used to increase robustness and the degree of external validity.

1.3.2 Notation

We use the standard notation in regression or classification and denote by X and Y the observed $n \times p$ design matrix of covariates and the $n \times 1$ response vector of the data, respectively; n is the sample size and p the dimensionality of the covariates. The i-th instance is denoted by X_i and Y_i, respectively, with X_i being a $1 \times p$ vector.

2 Deconfounding: In Presence of Dense Confounding

We consider the well-known problem of unobserved hidden confounding in a regression context. This is a special case of Figure 1, where the directions are known and without perturbations from external (exogenous) variables, see Figure 2.

There are several ways to explain it; we do so by using structural equation models (SEMs), see, for example, Bollen (1989) or Pearl (2009). We observe a univariate response variable Y_i, a $(p \times 1)$-dimensional covariate X_i and an unobserved $(q \times 1)$-dimensional hidden confounding variable H_i. In the linear case, the model is set up as follows:

$$\begin{aligned} Y_i &\leftarrow X_i \beta^0 + H_i \delta + \varepsilon_{Y,i}, \\ X_i &\leftarrow H_i \gamma + \varepsilon_{X,i}, \\ &\varepsilon_{X,i}, \varepsilon_{Y,i}, H_i \text{ jointly independent}, \end{aligned} \tag{1}$$

where β and δ are column vectors and γ a $q \times p$ matrix. We typically make an i.i.d. assumption across the indices $i = 1, \ldots, n$. The symbol '$\leftarrow$' is algebraically an equals sign, and it means in addition that the factorisation of the joint distribution of (all components of) Y_i, X_i, H_i, namely, $p(y, x, h) = p(y|x, h)p(x|h)p(h)$ with conditional distributions (densities), is precisely described by the equations, for example, $p(x|h) = p_{\varepsilon_X}(x - h\gamma)$. Figure 2 shows the corresponding graphical structure of the model. Of particular interest is the equation for the response variable Y; the goal is to infer the parameter β^0 from data. In the causality literature, β^0 is called the causal parameter of X_i on Y_i; but even without using the word causality, we can view it as the 'internal systems parameter'.

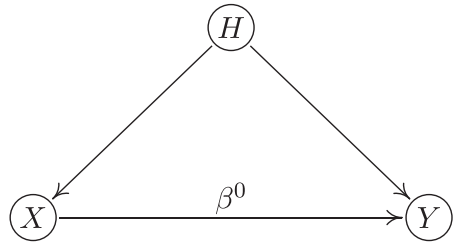

Figure 2. *Structure of the linear model with unobserved confounding variables H as in Equation (1).*

The parameter β^0 is not the regression parameter of Y versus X. In fact, due to confounding by the unobserved confounding variables H_i, we have

$$\begin{aligned} \operatorname{argmin}_\beta \mathbb{E} &= \beta^0 + b, \\ b &= \operatorname{Cov}(X_i)^{-1}\operatorname{Cov}(X_i, H_i)\delta. \end{aligned}$$

We can thus represent the model for the Y equation in Equation (1) as a standard linear model

$$Y_i = X_i(\beta^0 + b) + \varepsilon_i,\ \varepsilon_i = (H_i\delta - X_i b) + \varepsilon_{Y;i}, \tag{2}$$

where ε_i is uncorrelated with X_i, due to the property of the L_2 projection. A remarkable fact is that the bias b becomes small in case of high-dimensionality and 'dense' confounding as explained next.

A simple example of dense confounding.

Consider the case for Equation (1) where H is one-dimensional with $\operatorname{Var}(H_i) = 1$ and $\operatorname{Cov}(\varepsilon_{X;i}) = \xi^2 I$: $b = (\gamma^T\gamma + \xi^2 I)^{-1}\gamma^T\delta$, and for $\xi^2 \ll \|\gamma\|_2$ (in the context of dense confounding, $\|\gamma\|_2 \asymp \sqrt{p}$, see below), we obtain that

$$\|b\|_2 \approx |\delta|/\|\gamma\|_2.$$

Hence, if, say, all components of γ are of order one, that is, every component of X_i is affected by H_i with size of order one, which is some kind of dense confounding, we have that $\|b\|_2 = \mathcal{O}\left(|\delta|/\sqrt{p}\right)$. Therefore, this is a blessing of dimensionality when p is large.

One can see from the above example that the bias term of the population least squares principle becomes small in the case of high-dimensionality and dense confounding. However, with estimation based on a finite sample size, several issues become more delicate, and we propose to modify penalised least squares methods, as discussed next.

2.1 Deconfounding with Spectral Transformations

For estimating β^0 in Equation (1), we use a simple preprocessing technique that has some mathematical guarantees under an additional assumption of dense confounding.

2.1.1 Principal component adjustment

As a motivation, we consider first a commonly used approach to guard against hidden confounding as in Equation (1). We extract the first few principal components of X, denoted by $W^{(1)}, \ldots W^{(\hat{q})}$, ideally with $\hat{q}$ equal to or slightly larger than q. The $(n \times \hat{q})$ principal components $W = (W^{(1)}, \ldots, W^{(\hat{q})})$ serve as a proxy for the unobserved $(n \times q)$ H: the approximation is reasonable if the orthogonal projection $\Pi_W = W(W^TW)^{-1}W^T$ is similar to $\Pi_H = H(H^TH)^{-1}H^T$, that is, if X has approximately a low rank structure. One then adjusts for the principal components in W and builds partial residuals:

$$\tilde{X} = (I - \Pi_W)X,\ \tilde{Y} = (I - \Pi_W)Y, \tag{3}$$

and proceeds with (regularized) least squares regression of $\tilde{Y}$ versus $\tilde{X}$ to estimate the parameter β^0 in Equation (1). We can interpret this procedure in terms of singular values of X. Let

$$X = UDV^T$$

be the singular value decomposition of X. The singular values $D = \text{diag}(d_1, \ldots, d_m)$, with $m = \min(n, p)$ are ordered as $d_1 \geq d_2 \geq \ldots \geq d_m$. Consider a truncation of the singular values to

$$\tilde{d}_{\text{PCA},i} = 0\,(i = 1, \ldots, \hat{q}),\ \tilde{d}_{\text{PCA},i} = d_i\,(i = \hat{q} + 1, \ldots, m). \tag{4}$$

The PCA-adjusted matrix $\tilde{X}$ in Equation (3) can then be written as $\tilde{X} = U\tilde{D}V^T$, where $\tilde{D} = \text{diag}(\tilde{d}_1, \ldots, \tilde{d}_m)$ (note that $W^{(1)} \ldots, W^{(\hat{q})}$ are, when standardized to unit length, the first $\hat{q}$ column vectors of U). Alternatively, we can represent $\tilde{X}$ and also $\tilde{Y}$ in Equation (3) as a linear spectral transformation of the original quantities:

$$\begin{aligned} &\tilde{X} = FX,\ \tilde{Y} = FY, \\ &F = U\ \text{diag}(\tilde{d}_{\text{PCA},1}/d_1, \ldots, \tilde{d}_{\text{PCA},m}/d_m)U^T, \end{aligned} \tag{5}$$

and of course, we then have that $F = I - \Pi_W$.

2.1.2 The Trim transform and relations to Lava

One can think of other data transformations than the one in Equation (3) or (5). In fact, one may ask the question why the largest singular values in Equation (4) are shrunken to zero, making them the smallest singular values in the transformed $\tilde{X}$ (Figure 3).

It might be advantageous to keep the ordering of singular values in the transformed design matrix while still shrinking the large ones. Two particular choices are as follows. The Trim transform (Ćevid *et al.*, 2018) uses

$$\tilde{d}_{\text{Trim},i} = \min(d_i, \tau),\ i = 1, \ldots, m,$$

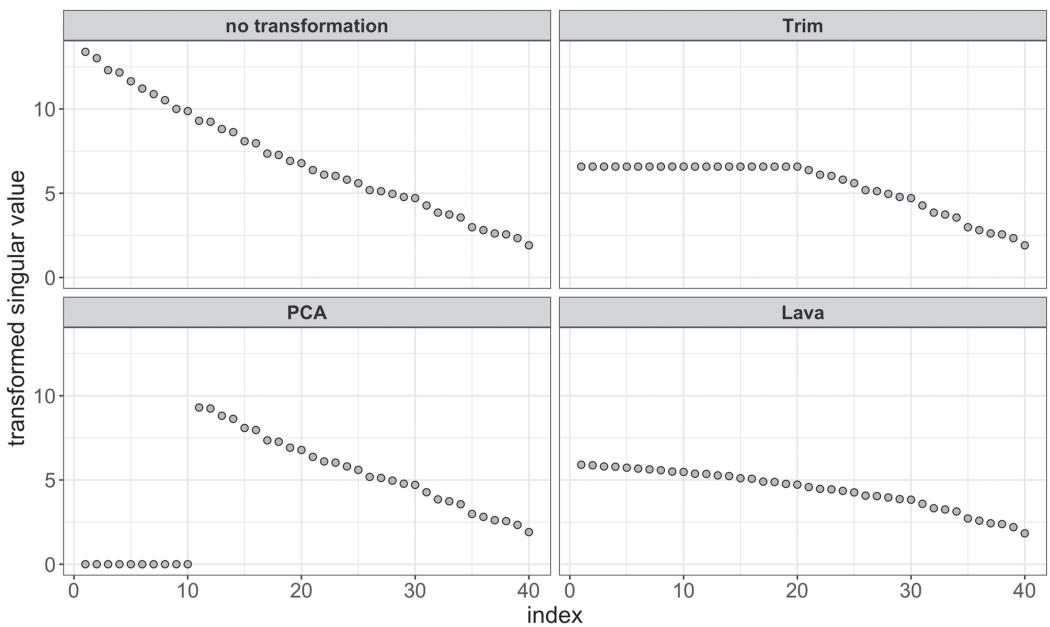

Figure 3. *Singular values of spectral transformed $\tilde{X}$. From top left to bottom right: no transformation with original X matrix, Trim-transform, PCA adjustment with 10 principal components, Lava. The figure is essentially taken from Ćevid et al. (2018).*

where τ is some threshold. A generic and often very good value is $\tau = d_{\lfloor m/2 \rfloor}$, the median of the singular values (Figure 3). The transformed variables are then of exactly the same form as in Equation (5), namely, premultiplying with the linear transformation F, involving now $\tilde{d}_{\text{Trim},i}$ being different than in Equation (4). Once we have the transformed data $\tilde{X}$ and $\tilde{Y}$, one can use 'any' reasonable technique for (high-dimensional) linear regression, say the Lasso (Tibshirani, 1996)

$$\hat{\beta}_{\text{TrimLasso}} = \text{argmin}_\beta \left(\|\tilde{Y} - \tilde{X}\beta\|_2^2/n + \lambda \|\beta\|_1 \right), \tag{6}$$

where $\lambda > 0$ is the regularisation parameter. The special case of ordinary least squares with $\lambda = 0$ is briefly mentioned in footnote 1. Other sparse estimators than the Lasso are possible as well, such as forward (Efroymson, 1960) or stagewise selection (Efron *et al.*, 2004), Elastic Net (Zou & Hastie, 2005), regularisation with the SCAD (Fan & Li, 2001), MCP (Zhang, 2010) or using guaranteed ℓ_0 (Bertsimas *et al.*, 2016) optimisation. We will describe in Fact 1 that the Lasso with the Trim transform $\hat{\beta}_{\text{TrimLasso}}$ estimates the parameter β^0 in Equation (1), assuming dense confounding. We note in particular, that the construction of the estimator is *very simple and easy to use,* requiring no further specialised software.

Another choice of a spectral transformation as in Equation (5) is implicit in the Lava (Chernozhukov *et al.*, 2017) estimator with

$$\tilde{d}_{\text{Lava},i} = \sqrt{\frac{n\lambda_2 d_i^2}{n\lambda_2 + d_i^2}},$$

where $\lambda_2 > 0$ is a tuning parameter. It is argued in Ćevid *et al.* (2018) that the choice

$$\lambda_2 = d_{\lfloor m/2 \rfloor}^2/n \tag{7}$$

is similar to the Trim transform with $\tau = d_{\lfloor m/2 \rfloor}$, see also Figure 3. We just point out that the Lava spectral transformation has an interesting representation in terms of estimating β^0 in Equation (1). It holds algebraically that for

$$(\hat{\beta}, \hat{b})_{\text{Lava}} = \text{argmin}_{\beta,b} \left(\|Y - X(\beta + b)\|_2^2/n + \lambda_1 \|\beta\|_1 + \lambda_2 \|b\|_2^2 \right),$$

we can represent, as in Equation (6),

$$\hat{\beta}_{\text{Lava}} = \text{argmin}_\beta \left(\|\tilde{Y} - \tilde{X}\beta\|_2^2/n + \lambda \|\beta\|_1 \right),$$

where $\tilde{X}$ and $\tilde{Y}$ are spectral transformed original quantities as in Equation (5) but with $\tilde{d}_{\text{Lava}}$ above. In view of the representation in Equation (2) with a sparse plus dense parameter vector, the Lava estimator $\hat{\beta}_{\text{Lava}}$ indeed estimates the sparse part β^0.

2.2 Guarantees for the Lasso After the Trim Transform

Once we have the Trim-transformed data $\tilde{X}$ and $\tilde{Y}$, we can use linear regression techniques for estimating β^0 in Equation (1). We note that for least squares estimation with $\text{rank}(X) = p < n$, nothing will happen.[1] But for higher dimensions and penalised methods, things change.

We consider the Lasso on the Trim-transformed data as in Equation (6) for some regularisation parameter λ. Standard software can be used, for example, `glmnet` in `R` (Friedman *et al.*, 2010). The choice of the regularisation parameter is perhaps a bit more delicate, but we propose the usual, for example, 10-fold cross-validation, see also below in Section 2.3. This

simple combination of deconfounding with the Trim transform in conjunction with the Lasso has interesting theoretical guarantees under the following main assumptions:

(A1) $\lambda_{\max}(\mathrm{Cov}(X_i, H_i)) \asymp \sqrt{p}$: the largest singular value of the $(p \times q)$ covariance matrix of (X_i, H_i) is of the order $\sqrt{p}$.
(A2) $d_{\lfloor n/2 \rfloor} = \mathcal{O}_P\left(\sqrt{p}\right)$: the median value of the singular value of X is of the order $\sqrt{p}$, with high probability.
(A3) The compatibility constant of $n^{-1}\tilde{X}^T\tilde{X}$ is of the same order as the minimal eigenvalue $\lambda_{\min}(\Sigma)$ of $\Sigma = \mathrm{Cov}(X_i)$.

Ćevid *et al.* (2018) give a detailed discussion when these assumptions hold, see also Guo *et al.* (2020). In particular, (A1) is an assumption on dense confounding; for example, if $p/q \to \infty$ and the number of non-zero columns of γ is of the order p (order p components of X are affected by H), and each of the non-zero columns of γ is sampled i.i.d. from a sub-Gaussian vector, then (A1) holds with high probability. This is an extension and along the lines of our simple example above on dense confounding. Assumption (A2) holds with high probability if the rows of X are realisations of i.i.d. random vectors (assuming sufficiently many finite moments).

Fact 1. (Ćevid *et al.*, 2018) Consider the confounding model in Equation (1) with $p \geq n$ and $\max_j \Sigma_{jj} = \mathcal{O}(1)$, where $\Sigma = \mathrm{Cov}(X)$. Assume (A1)–(A3). Then, for some $\lambda \asymp \sqrt{\log(p)/n}$ in Equation (6), the usual rate of convergence as in the unconfounded high-dimensional linear model holds, namely,

$$\|\hat{\beta}_{\mathrm{TrimLasso}} - \beta^0\|_1 = \mathcal{O}_P\left(\frac{\sigma s_0}{\lambda_{\min}(\Sigma)}\sqrt{\log(p)/n}\right),$$

where $s_0 = |\mathrm{supp}(\beta^0)|$ is the number of non-zero components of β^0 and $\sigma^2 = \mathrm{Var}(H_i\delta + \varepsilon_{Y;i})$.

The asymptotics is to be understood as the usual one in high-dimensional statistics where both $p \geq n \to \infty$.

Other methods based on the Trim-transformed data, such as forward selection or regularisation with sparsity inducing penalties other than the ℓ_1-norm, have not yet been theoretically established to exhibit certain convergence rates. Fact 1 above serves as an indication that algorithms and methods are expected to behave well when using them on Trim-transformed data.

2.3 Choosing the Regularisation Parameter

Choosing the regularisation parameter for Lasso or other algorithms with cross-validation is conceptually somewhat different than in the standard setting with no confounding.

For the sake of illustration, consider the Lasso $\hat{\beta}_{\mathrm{TrimLasso}}(\lambda)$ on the Trim-transformed data as in Equation (6). When using cross-validation, aiming for best prediction, the chosen λ would be typically too small because the best prediction would also try to capture the unwanted signal component Xb in Equation (2). To partially correct for this issue, cross-validation should be run on the deconfounded data $\tilde{X}, \tilde{Y}$ and ignoring the issue that the spectral transformation has used the full data; that is, we simply spectral-transform the full data set first and then proceed as usual. This strategy should make the additional signal $\tilde{X}b$ smaller and hence cross-validation aiming for best prediction is expected to perform reasonably well.

As an alternative to cross-validation, one can use Stability Selection (Meinshausen & Bühlmann, 2010) on the original data. This amounts to directly choosing an amount of regularisation for selecting the relevant components of β^0, that is, for variable selection. It does *not* lead to an estimate for the tuning parameter λ in Equation (6). Instead, Stability Selection is linking a different stability-based regularisation with the expected number of false positives, assuming

an exchangeability condition for i.i.d. generated data. However, the methodology aiming for stability is also useful for heterogeneous data where the underlying distribution has changed as discussed in next.

2.4 Robustification Against Hidden Confounding and External Validity

Perhaps the main value of the deconfounded Lasso procedure, that is, Trim-transforming the data and using Lasso, is the degree of robustification against hidden confounding. The assumptions (A1)–(A3) in Section 2.2 might be partially unrealistic, but Ćevid *et al.* (2018) report empirically that 'there is not much to lose, but potentially a lot to be gained'. This can be summarised as follows: (i) the procedure is extremely simple requiring in addition only one singular value decomposition and (typically) three lines of code; (2) the method is very effective in estimating the underlying unconfounded regression parameter β^0 in scenarios of dense confounding and a sparse β^0; (3) in the case of no confounding, the deconfounded Lasso is essentially as good as the plain Lasso; (4) in between the settings in (2) and (3), there is improvement with the deconfounded Lasso over its plain version, yet it still does not entirely remove the bias due to confounding. We refer also to Section 2.6.

The unconfounded parameter β^0 is the parameter where other sources of unmeasured variation have been removed. This is very relevant for improving replicability. Suppose that we estimate the regression parameter on one (training) data set and would like to have it replicated on another (test) data set. If the two data sets differ in their distribution, the regression parameter is not replicable. However, the unconfounded parameter β^0 is replicable under the following assumption:

$$\begin{aligned} \text{the training data set}: &\quad \text{is generated from the model in (1)},\\ \text{the test data set satisfies}: &\quad Y_i' \leftarrow X_i'\beta^0 + H_i'\delta' + \varepsilon'_{Y,i},\\ &\quad X_i' \leftarrow H_i'\gamma' + \varepsilon'_{X,i},\\ &\quad \varepsilon'_{X,i}, \varepsilon'_{Y,i}, H_i' \text{ jointly independent}, \end{aligned}$$

where the unconfounded parameter β^0 is the same, but the other parameters are allowed to change; the notation with the superscript $'$ denotes the quantities corresponding to the test data set (but β^0 in the test data is the same as in Equation 1). We will illustrate such a replicability phenomenon on real data below.

2.4.1 An Illustration on Data from the GTEx Consortium

The Genotype-Tissue Expression (GTEx) project is studying tissue-specific gene expression and regulation in human samples (https://gtexportal.org, Lonsdale *et al.* (2013)). Here, we consider a small aspect of the publicly available data.

For the specific skeletal muscle tissue, we have 14 713 gene expression measurements for $n = 491$ samples. In addition, there are 65 additional covariates that are believed to be good proxies of confounding variables, including genotyping principal components and so-called PEER factors. Thus, we have two data sets: the raw data with covariates X and response Y, and another with X' and Y', where we linearly regress out the 65 proxies for hidden confounding and X', Y' are the corresponding residuals. The response variable is the expression of one (randomly chosen) gene while the covariates comprise all other expressions. If there is hidden (linear) confounding and the proxy variables indeed capture the true underlying hidden confounding variables, that is, the linear span of the 65 proxy variables equals the linear span of the unobserved hidden variables, the X', Y' data are unconfounded. This in turn would imply

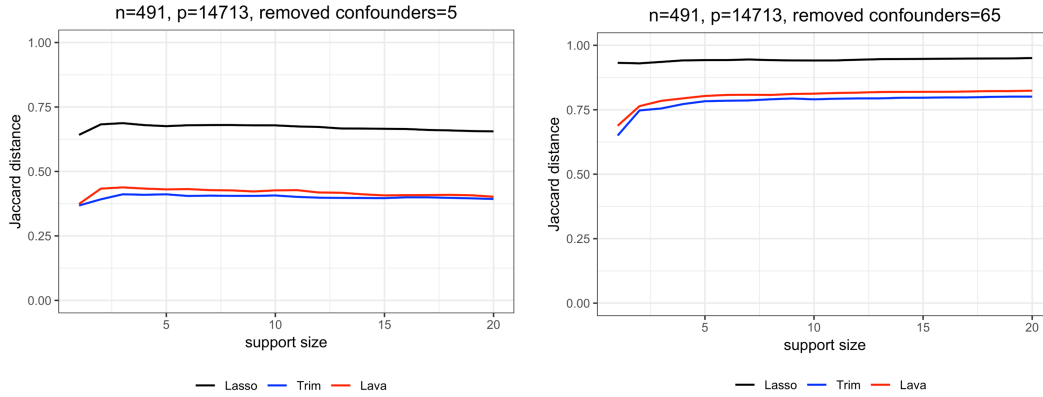

Figure 4. *Stability of selected features in GTEx data. Jaccard distances for sets of selected variables on original and proxy adjusted (for approximate deconfounding) data. Adjustment with 5 (left) and 65 (right) proxies for hidden confounding. x-axis: number of top selected features (support size); y-axis: Jaccard distance between sets of top selected features (of fixed support sizes) based on original and proxy adjusted data. Black: plain Lasso; blue: trim transform followed by Lasso; red: Lava. The figure is essentially taken from Ćevid et al. (2018).*

that the deconfounded Lasso as in Equation (6) would give similar results on (X, Y) and (X', Y'), while this would not be the case for the plain Lasso as it would be subject to some bias when running it on the confounded data (X, Y). We aim here to illustrate that the deconfounded Lasso is indeed more robust than the plain Lasso when estimation is done once on the original (X, Y) and once on the approximately deconfounded data (X', Y').

The response variable is one randomly selected gene and the remaining $p = 14712$ gene expressions are the covariates. We compute the regularisation paths of the Lasso with the Trim transformed data and of the plain Lasso: this leads to sets of active variables with non-zero coefficient estimates $\hat{S}^{(r)}_{\text{TrimLasso}}(\lambda) \subset \{1, \ldots, p\}$ and $\hat{S}^{(r)}_{\text{Lasso}}(\lambda) \subset \{1, \ldots, p\}$, where $r = 1, 2$ denotes the original (X, Y) and the proxy adjusted data (X', Y'), respectively. As a measure of robustness or consensus, we compute the Jaccard distance between $r = 1$ and $r = 2$ for sets of the same cardinality (by varying λ). For comparison, we consider also the Lava estimator with the tuning parameter as in Equation (7). Figure 4 reports the results when adjusting once with 5 and once with all 65 proxy confounding variables and when averaging over 500 randomly chosen response variables. The problem is very high-dimensional and with a high noise level; nevertheless, we clearly see that deconfounding with the Trim transform, and Lava likewise, provide more robustness than the plain Lasso, across a large range of cardinalities of the active sets. We note that the robustness or consensus decreases with more proxy adjustment; this is mainly due to the fact that the data sets (X, Y) and (X', Y') become more different with more adjustment. But the advantage remains when considering differences between the methods.

We mention here that Shah *et al.* (2020) provide vaguely related results on robustness for the GTEx data for another Ridge-type procedure for undirected graphical models.

2.5 The Doubly Debiased Lasso

Assigning uncertainty is a core task in statistical inference. Substantial progress has been made for low-dimensional parameters in high-dimensional models. The prime example is about inference for single components of a high-dimensional regression parameter. The debiased or desparsified Lasso has become a basic machinery for constructing hypothesis tests and confidence intervals (Zhang & Zhang, 2014; van de Geer *et al.*, 2014), see also Dezeure *et al.* (2015)

for some review of the earlier work. An interesting property of the debiased or desparsified Lasso is its efficiency, assuming sparsity conditions (van de Geer *et al.*, 2014).

We briefly review here the approach of Guo *et al.* (2020) on the doubly debiased Lasso to obtain hypothesis tests and confidence intervals for single regression coefficients β_j^0 in model (1) in presence of hidden confounding.

2.5.1 The standard debiased Lasso

The idea of debiasing the Lasso is based on partial regression. For ordinary least squares estimation in the $p < n$ regime, we obtain the estimator $\hat{\beta}_{\mathrm{OLS};j}$ as follows:

$$\begin{aligned}&\text{do least squares regression of } X^{(j)} \text{ versus } X^{(-j)} \text{ and denote the residuals by } Z^{(j)},\\ &\hat{\beta}_{\mathrm{OLS};j} = (Z^{(j)})^T Y / \|Z^{(j)}\|_2^2 = (Z^{(j)})^T Y / ((Z^{(j)})^T X^{(j)}),\end{aligned} \tag{8}$$

where $X^{(-j)}$ is the $(n \times (p-1))$ matrix arising from deleting the j-th column of X. The first regression in Equation (8) is ill-posed if $p > n$. The debiased Lasso then uses instead

$$\text{a Lasso regression of } X^{(j)} \text{ versus } X^{(-j)} \text{ and denote the residuals again by } Z^{(j)}. \tag{9}$$

When using them in the second regression, we obtain

$$\frac{(Z^{(j)})^T Y}{(Z^{(j)})^T X^{(j)}} = \beta_j + \sum_{k \neq j} \frac{(Z^{(j)})^T X^{(k)} \beta_k}{(Z^{(j)})^T X^{(j)}} + \frac{(Z^{(j)})^T \varepsilon_Y}{(Z^{(j)})^T X^{(j)}},$$

where we assume an unconfounded model $Y = X\beta + \varepsilon_Y$ with $\mathbb{E} = 0$. Unlike as for least squares, $(Z^{(j)})^T X^{(k)} \neq 0$ for $k \neq j$, and hence, there is a bias term

$$B = \sum_{k \neq j} \frac{(Z^{(j)})^T X^{(k)} \beta_k}{(Z^{(j)})^T X^{(j)}}.$$

An obvious estimator for the bias arises by plugging in a Lasso estimate of Y versus X, resulting in

$$\hat{B} = \sum_{k \neq j} \frac{(Z^{(j)})^T X^{(k)} \hat{\beta}_{\mathrm{Lasso};k}}{(Z^{(j)})^T X^{(j)}}, \tag{10}$$

and the debiased or desparsified Lasso is then defined as

$$\hat{\beta}_{\mathrm{debiasedLasso};j} = \frac{(Z^{(j)})^T Y}{(Z^{(j)})^T X^{(j)}} - \hat{B}.$$

In case of hidden confounding in model (1), both regressions in Equations (8) and (9) are exposed to bias from hidden confounding and standard methodology does not work. Following the ideas developed in Sections 2.1–2.2, we propose to Trim transform the data twice, once before applying the Lasso in the $X^{(j)}$ versus $X^{(-j)}$ regression (9) and once before using the Lasso in Y versus X for being plugged-in to the bias estimator in Equation (10). By doing so, we remove bias thanks to spectral transformations, and hence, the words 'doubly debiased'. Of course, there are tuning parameters to be chosen, namely, for each of the Lasso regressions appearing in Equations (9) and (10). This issue is analogous as for the standard debiased or desparsified Lasso but perhaps one tick more difficult as indicated in Section 2.3. The details

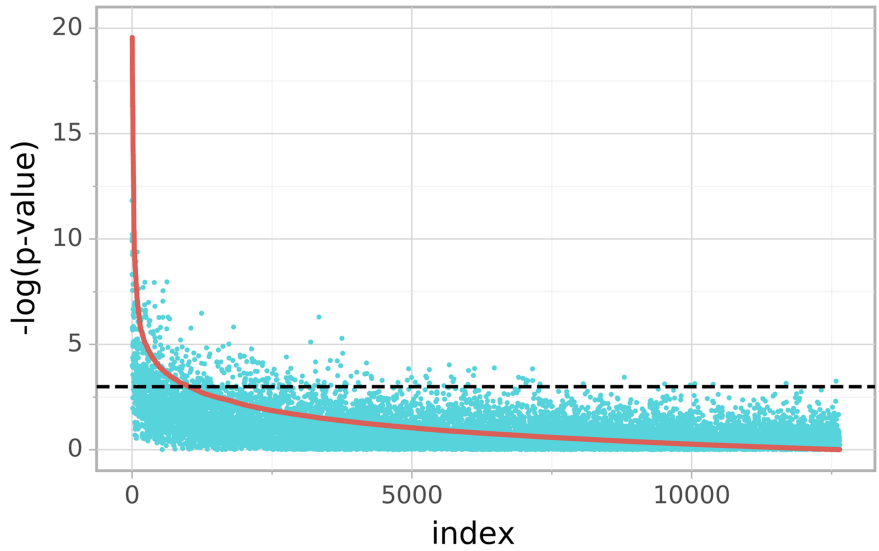

Figure 5. *P values for two-sided test of the hypothesis* $H_{0,j} : \beta_j^0 = 0$ *with two-sided alternative for GTEx data. Doubly debiased Lasso (blue) and debiased Lasso (red) for the expression of a predetermined landmark gene (being the response Y) versus all other gene expressions (being the covariates X). x-axis: index of covariates, ordered by decreasing significance based on the Debiased Lasso.; y-axis: negative log P value. Black dotted line indicates the 5% significance level with the value* $-\log(0.05)$. *The figure is essentially taken from Guo et al. (2020).*

are given in Guo *et al.* (2020), and the resulting estimator is called the doubly debiased Lasso $\hat{\beta}_{\text{DDLasso}}$.

The following result holds.

Fact 2 (Guo *et al.*, 2020) Consider the confounding model in Equation (1) with $\max_j \Sigma_{jj} = \mathcal{O}(1)$, where $\Sigma = \text{Cov}(X)$. Under similar conditions as in (A1)–(A3) and assuming sparsity for both the regressions of Y versus X and the one in Equation (9),

$$V_j^{-1/2}(\hat{\beta}_{\text{DDLasso};j} - \beta_j^0) \Rightarrow \mathcal{N}(0,1) \; (P \geq n \to \infty),$$
$$V_j = n^{-1}\text{Var}(\varepsilon_Y) g_j(X),$$

with a known specific function $g_j(X)$ of the design matrix X that is of order 1 as $p \geq n \to \infty$. In addition, if the trimming threshold is such that the fraction of the shrunken singular values converges to zero (only the very large singular values are Trimmed to the corresponding quantile value), the doubly debiased Lasso is as efficient as the ordinary least squares estimator in low dimensional settings, that is,

$$V_j \asymp n^{-1}\text{Var}(\varepsilon_Y)(\text{Cov}(X))_{jj}^{-1}.$$

2.5.2 An illustration on the GTEx data

We consider again the GTEx data mentioned in Section 2.4.1, but now with a somewhat smaller dimensionality $P = 12\,646$ but an increased sample size of $n = 706$ (removing some of the covariates with missing values due to larger sample size). There are approximately one thousand landmark genes of particular importance and interest.

Figure 5 illustrates the difference between the P values from the doubly Debiased Lasso in comparison with the plain debiased Lasso that does not protect against potential hidden confounding. The plot considers a particular landmark gene whose expression is the response variable, and all other 12 645 gene expressions are covariates. The doubly debiased Lasso claims less significance, which seems a plausible finding (and it is not primarily due to larger variance that is not shown here).

We also illustrate increased robustness of the doubly debiased Lasso. As explained already in Section 2.4.1, there are 65 additional proxy variables that are aimed to approximate unobserved hidden confounding. Figure 6 shows P values for 10 response landmark genes (and the plots comprise all P values from the 10 regressions). We can see from the left plot that the doubly debiased Lasso is much more conservative for the potentially confounded original (X, Y) data. The cloud of points is skewed upwards showing that the standard debiased Lasso declares many more predictors as significant. On the other hand, in the right plot, the P values obtained by the two methods are much more similar for the proxy-adjusted unconfounded (X', Y') data, and the point cloud is now much less skewed upwards. The remaining deviation from the line $y = x$ might be due to the remaining confounding, not accounted for by regressing out the given confounder proxies. Figure 6 describes the results.

2.6 When Dense Confounding Fails

When hidden confounding is substantial but fails to be dense in the sense that it affects many of the components of X, the deconfounding Trim-transformation technique does not effectively remove the entire bias, and the parameter β^0 is not identifiable from the data generating distribution. In such situations, other assumptions are required, see also Section 3.

However, deconfounding leads to robustification, as mentioned in Section 2.4; the points (1)–(4) there are relevant in general, also for inference with the doubly debiased Lasso. We repeat here again that deconfounding can lead to substantial improvements and seems to never

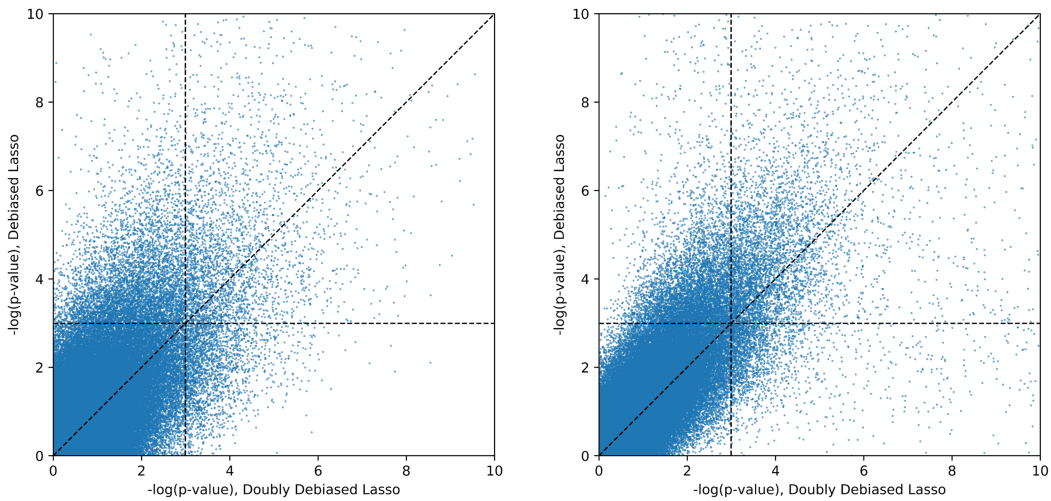

Figure 6. *Stability of P values in GTEx data. Comparing P values from original data and its proxy adjusted version for approximate deconfounding. Two-sided tests of the hypothesis* $H_{0,j} : \beta_j^0 = 0$, *for 10 landmark gene expression as responses and all other expressions as covariates, obtained by doubly debiased Lasso (x-axis, - log P value) and standard debiased Lasso (y-axis, - log P value). Original (X, Y) data (left) and adjusted* (X', Y') *with 65 proxies for hidden confounding (right). Horizontal and vertical black dashed lines indicate the 5% significance level. The figure is taken from Guo et al. (2020). [Colour figure can be viewed at* wileyonlinelibrary.com*]*

make things substantially worse than not taking any action against hidden confounding. The price to be paid for deconfounding is typically a slightly larger variance of the estimator resulting in a somewhat reduced efficiency for data without any confounding.

3 Anchor Regression: Towards Causality, Distributional Robustness and Distributional Replicability

We consider now the general situation from Figure 1; it extends Figure 2 because the directions between the variables X, Y, and H are unknown. Furthermore, we abandon here the major assumption of dense confounding from Section 2. Both are important relaxations in practice. However, this comes with the price of requiring access to exogenous variables A as indicated in Figure 1; as an example, we mention the case where the variables A represent mean shift perturbations (Figure 1), where exogeneity (source node in the graph in Figure 1) is often a reasonable assumption. This graphical exogeneity assumption is the reason for the name 'anchor regression', and we will explain more below.

Instrumental variables regression is a popular proposal for a special case with perturbations (Bowden & Turkington, 1990; Angrist *et al.*, 1996; Stock & Trebbi, 2003; Imbens, 2014; Imbens & Rubin, 2015). The SEM in Equation (1) is extended to

$$\begin{aligned} &Y_i \leftarrow X_i\beta^0 + H_i\delta + \varepsilon_{Y,i}, \\ &X_i \leftarrow A_i\kappa + H_i\gamma + \varepsilon_{X,i}, \\ &A_i,\ H_i,\ \varepsilon_{Y,i},\ \varepsilon_{X,i} \text{ jointly independent,} \end{aligned} \tag{11}$$

where the random variables are i.i.d. across $i = 1, \ldots, n$. The main assumption here is that the so-called instrumental variables A_i do not directly affect the hidden variable H_i nor the response variable Y_i. The well-known two-stage least squares (TSLS) estimator is then defined as the least squares estimator on linearly transformed data:

$$\begin{aligned} &\tilde{Y} = \Pi_A Y,\ \tilde{X} = \Pi_A X,\ \Pi_A = A(A^TA)^{-1}A^T, \\ &\hat{\beta}_{\text{TSLS}} = \text{argmin}_\beta \|\tilde{Y} - \tilde{X}\beta\|_2^2/n = \text{argmin}_\beta \|\Pi_A(Y - X\beta)\|_2^2/n = \text{argmin}_\beta \|Y - \Pi_A X\beta)\|_2^2/n, \end{aligned} \tag{12}$$

where the right-hand side is perhaps the most common representation.

3.1 Causal Regularisation and Anchor Regression

More generally than two-stage least squares estimation in Equation (12), we can look at its regularised version, called anchor regression (Rothenhäusler *et al.*, 2018):

$$\hat{\beta}^{(\gamma)}_{\text{anchor}} = \text{argmin}_\beta \left(\|(I - \Pi_A)(Y - X\beta)\|_2^2/n + \gamma \|\Pi_A(Y - X\beta)\|_2^2/n \right), \tag{13}$$

for some regularisation parameter $0 \le \gamma \le \infty$. The name 'anchor regression' originates from the fact that A does not necessarily need to fulfil the assumption of an instrument in IV regression: A only needs to be an exogenous variable, that is, an 'anchor' for leading to meaningful results. More details are given in Sections 3.2 and 3.3. With $\gamma = 0$, we obtain adjustment with respect to A, that is, partialling out the linear effect of A, $\gamma = 1$ corresponds to ordinary least squares and $\gamma = \infty$ is two-stage least squares. This regularisation has been proposed much earlier in a different but equivalent form under the name of K-class estimators, mainly for reducing the large (or infinite) variance of two-stage least squares for estimation of β (Theil, 1958; Jakobsen

& Peters, 2020). The computation of anchor regression is extremely easy and modular; one can simply transform the data

$$\tilde{X} = W_\gamma X,\ \tilde{Y} = W_\gamma Y,$$
$$W_\gamma = I - \left(1 - \sqrt{\gamma}\right) \Pi_A,$$

and then use least squares estimation of $\tilde{Y}$ versus $\tilde{X}$. One can also consider sparsity-regularised anchor regression with, for example, the ℓ_1-norm penalty:

$$\hat{\beta}^{(\gamma)}_{\text{anchor}} = \text{argmin}_\beta \left(\|(I - \Pi_A)(Y - X\beta)\|_2^2/n + \gamma \|\Pi_A(Y - X\beta)\|_2^2/n + \lambda \|\beta\|_1 \right),$$

which can be solved by running standard Lasso of $\tilde{Y}$ versus $\tilde{X}$.

The anchor regression method is also called causal regularisation because it regularises least squares towards the causal parameter (whereas the motivation for Equation (13) above has been to regularise the TSLS estimator towards least squares to reduce variance). For $\gamma \to \infty$, we approximate a causal solution under the assumptions of instrumental variables regression. More generally, one can improve robustness and replicability when choosing γ clearly larger than 1 as discussed in Sections 3.2 and 3.3.

We can also connect anchor regression to deconfounding from Section 2. When taking the anchor variables A as the first $\hat{q}$ principal components of X, then $\gamma = 0$ corresponds to PCA adjustment as described in Equation (4); for $\gamma > 0$ but small, such an anchor regression would shrink the first $\hat{q}$ singular values of X, but it is not a spectral transform any longer of the form as in Equation (5) with any transformed singular values $\tilde{d}_i$. Also, when using A as the first principal components of X, the exogeneity assumption as in Figure 1 is violated, a crucial condition for what we discuss next.

3.2 Distributional Robustness of Anchor Regression

Rothenhäusler *et al.* (2018) take a very different view of Equation (13) than improving the mean squared error of the two-stage least squares estimator (12) in IV regression, namely, that anchor regression is sensible even when the main assumptions of IV regression fail. That is, if we allow that A directly affects the hidden variables H or the response Y in Equation (11), which implies that β^0 is not identifiable from the data, anchor regression is still estimating an interesting parameter, as we discuss next. The main assumption is that A is an exogenous variable, that is, a source node in the graphical representation of a SEM in Figure 1, as used in our main working model (15). Consider the population version of anchor regression in Equation (13):

$$\beta^{(\gamma)} = \text{argmin}_\beta \left(\mathbb{E}[|(I - P_A)(Y_i - X_i\beta)|^2] + \gamma \mathbb{E}[|P_A(Y_i - X_i\beta)|^2] \right), \tag{14}$$

where $P_A(\cdot) = \mathbb{E}$ is the population version of Π_A under a linearity assumption as in Equation (15) below, and the index i is arbitrary (because we assume that the data are i.i.d. across samples). This population parameter $\beta^{(\gamma)}$ is a *regularised* population parameter, where the regularisation is not used to obtain better statistical finite sample properties. Instead, the regularisation has a direct relation to distributional robustness.

To explain such a robustness, assume that the training data are i.i.d. realisations of the following structural equation model:

$$\begin{pmatrix} X_i \\ Y_i \\ H_i \end{pmatrix}^T = B \begin{pmatrix} X_i \\ Y_i \\ H_i \end{pmatrix}^T + \varepsilon_i + \text{M}A_i^T = (I - B)^{-1}(\varepsilon_i + \text{M}A_i^T), \tag{15}$$

where (all the components of) A_i, ε_i are jointly independent and M is a coefficient matrix of dimension $(\dim(X_i, Y_i, H_i) \times \dim(A_i))$. Note that $I - B$ is always invertible if the model structure corresponds to an acyclic directed graph.

We define the system under shift perturbations v by the same equations as in Equation (15) but replacing the term MA from the contributions of the anchor variables by a deterministic or stochastic perturbation vector v. That is, the system under shift perturbations satisfies:

$$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix}^T = B \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix}^T + \varepsilon + v = (I - B)^{-1}(\varepsilon + v), \tag{16}$$

with ε having the same distribution as ε_i in Equation (15). The shift vector v is assumed to be in the span of M, that is, $v = \mathrm{M}\delta$ for some vector δ. Thus, the vector v shifts the variables X^v, Y^v, H^v in the same direction as A_i, according to the range (or span) of M but with possibly different strengths. The variables X^v, Y^v can be interpreted as the test data coming from a different distribution than the training data from model (15).

An example with discrete anchors, encoding different environments. We often have the following situation in mind. The data are heterogeneous from various subpopulations or environments labeled by $\{1, \ldots, \ell\}$. These are then encoded with ℓ-dimensional anchor variables A_i in the form of dummy variables. The heterogeneity of the data enters as distributional additive shifts (or perturbations) in terms of $\mathrm{M}A_i^T$. As an environment, it is often reasonable to assume that A_i is exogenous, that is, a source node in the graph in Figure 1. The data generated by Equation (16) is typically the test data where realisations of the anchor variable A are not available.

We do not have to rely on discrete anchors as in the example above and all subsequent results hold in general.

Fact 3. (Rothenhäusler *et al.*, 2018) Consider random variables X_i, Y_i as in Equation (15) and X^v, Y^v as in Equation (16). Then, for any $b \in \mathbb{R}^p$, it holds that

$$\sup_{v \in C_\gamma} \mathbb{E}[(Y^v - X^v b)^2] = \mathbb{E}[((I - P_A)(Y_i - X_i^T b))^2] + \gamma \mathbb{E}[(P_A(Y_i - X_i^T b))^2],$$

where

$$C_\gamma = \{v; \quad v = \mathrm{M}\delta \text{ for random or deterministic } \delta, \text{ uncorrelated with } \varepsilon \\ \text{and } \mathbb{E}[\delta\delta^T] \preccurlyeq \gamma \mathbb{E}[AA^T]\}. \tag{17}$$

Fact 3 establishes distributional robustness of the population version of anchor regression; the parameter γ has an exact correspondence to the class C_γ of shift perturbations. From a practical view-point, Fact 3 tells us that we can construct an estimator on the training data only, by employing causal regularisation, which protects on new test data that arise from shift perturbations as in Equation (17).

Rothenhäusler *et al.* (2018) give finite sample versions of the result in Fact 3 and show empirical examples how prediction can be improved thanks to distributional robustness; if the test data are a perturbed version of the training data, formalised with X_i^v, Y_i^v for some perturbation vector v, then the expected worst case squared error loss on the test data can be optimised by anchor regression.

3.2.1 Choosing γ and specifying anchor variables

The choice of γ in anchor regression or causal regularisation can be addressed from different angles. If we want to insure ourselves against bad perturbations in span(M) of a certain size, as defined in Equation (17) and aiming for worst case optimal prediction, then γ corresponds to the multiplication factor of the observed heterogeneity in the data. That is, for example, $\gamma = 5$ corresponds to perturbations $\sqrt{5}$ times as large as the ones we have observed in the data. Alternatively, we can consider leave-one-environment-out cross-validation and choose γ that optimises the worst case performance among the left-out environments (being the test data).

Regarding the specification of anchor variables, as mentioned above, A_i should be exogenous; we describe below an example with heterogeneity arising from different environments. The general idea is to *stabilise* the estimator $\hat{\beta}_{\text{anchor}}$ over the values of A_i by taking a large value of γ that enforces that the residuals are nearly orthogonal to A_i; see also (Pfister *et al.*, 2019c) and a replicability result in Fact 4. This is the opposite action than using A_i as an additional covariate that would correspond to $\gamma = 0$. Exogeneity and $\gamma = \infty$ (two stage least squares estimation) plays also a prominent role in IV regression for deconfounding the effects of hidden confounders H_i. Thus, the stabilising anchor regression estimator is ideally pursued with exogenous anchor variables and large values of γ.

3.3 Distributional Replicability and External Validity with Anchor Regression

Anchor regression also leads to an improved replicability on new data, say from a related study. We argue here that the parameter

$$\beta^{(\to\infty)} = \limsup_{\gamma\to\infty} \beta^{(\gamma)}$$

can be replicated on new data from a different distribution than the training data. In view of Fact 3, $\beta^{(\to\infty)}$ leads to distributional robustness for arbitrarily large perturbations $v = \mathrm{M}\delta$ in the span of M. If the assumptions from instrumental variables regression hold, then $\beta^{(\to\infty)} = \beta^0$ which is the causal parameter, but that is not the case in general. The causal parameter has an invariance property with respect to certain arbitrarily strong perturbations. Also $\beta^{(\to\infty)}$ leads to an invariance of the residuals, namely,

$$Y^v - X^v\beta^{(\to\infty)}$$
has the same distribution for all arbitrarily strong perturbations $v$as in Equation (16). (18)

According to a general relation between causality and invariance, and due to the residual invariance from Equation (18), we call $\beta^{(\to\infty)}$ the 'diluted causal' parameter.

We consider now the following setting. The first data set is generated from the model (15) whose distribution induces the diluted causal parameter $\beta^{(\to\infty)}$ (being a function of the data generating probability distribution). The second (test or validation) data set is generated from a perturbed version as in model (16) whose distribution generates the diluted causal parameter $b'^{(\to\infty)}$.

For our replicability result in Fact 4, we require the so-called projectability assumption:

$$I = \{\beta; \mathbb{E}[Y - X\beta|A] \equiv C\} \neq \varnothing \text{ for any constant C.} \tag{19}$$

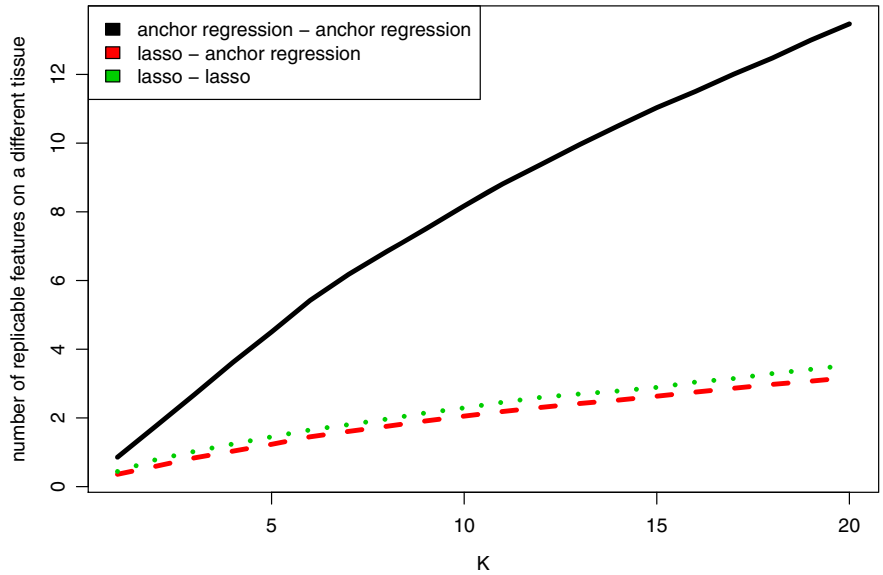

Figure 7. *Replicability of diluted causal parameter $\beta^{(\to\infty)}$ on GTEx data. x-axis: number K of top ranked features; y-axis: average overlap (number) among the top K features among $\binom{13}{2}$ tissue pairs, and averaged over 200 randomly selected response variables each being one of the available gene expressions. Anchor regression with $\gamma = 16$ on both tissues in each pair (solid black), Lasso on both tissues in each pair (dotted green), anchor regression with $\gamma = 16$ on one tissue and Lasso on the other tissue for each tissue pair (dashed red). The figure is taken from Rothenhäusler et al. (2018 figure 4).*

This condition holds if and only if

$$\text{rank}(\text{Cov}(A, X)) = \text{rank}[\text{Cov}(A, X), \text{Cov}(A, Y)],$$

where $[\text{Cov}(A,X), \text{Cov}(A, Y)]$ denotes the extended matrix by concatenating the columns of the two matrices. For example, if $\text{rank}(\text{Cov}(A,X))$ is full rank and $\dim(A) \leq \dim(X)$, the projectability condition (19) holds.

Fact 4. (Rothenhäusler *et al.*, 2018) Consider the diluted causal parameters $\beta^{(\to\infty)}$ from model (15) and $b'^{(\to\infty)}$ from model (16). Assume the projectability condition (19) for the model (15). Then, $\beta^{(\to\infty)} = b'^{(\to\infty)}$, that is, the diluted causal parameter is replicable on the new perturbed data set.

3.3.1 An illustration on the GTEx data

We illustrate the distributional replicability for the diluted causal parameter $\beta^{(\to\infty)}$ on the GTEx data mentioned already in Section 2.4.1 and 2.5.2.

Here, we consider 13 different tissues for which $p = 12948$ gene expression measurements and 65 proxies of confounding are measured. The 13 different tissues correspond to 13 different data sets consisting of response variables Y being one of the gene expressions, covariates X comprising all other gene expressions and anchor variables A being the 65 proxy variables. The sample size varies between 300 and 700 across the 13 tissues.

We consider anchor regression with $\gamma = 16$ (being chosen as a large value, yet still improving the variance in comparison with choosing $\gamma = \infty$, that is, two-stage least squares) and cross-validated choice of the tuning parameter λ for an ℓ_1-norm penalty as an estimator for

$\beta^{(\to\infty)}$, for each of the 13 tissues (data sets). The goal is to evaluate the degree of replicability and external validity of the anchor regression estimator. Figure 7 illustrates the results. The anchor regression estimate $\hat{b}^{t,(\gamma=16)}$ for one tissue t is compared with another one $\hat{b}^{t\prime,(\gamma=16)}$ for another tissue t'. The overlap (number) among the top K variables (features), according to the absolute value of the estimates, is counted and averaged over all $\binom{13}{2}$ pairs of tissues (data sets) and 200 random choices of a response as one of the available gene expressions. Figure 7 displays the results. There is some evidence for the GTEx data that indeed, anchor regression for the diluted causal parameter $\beta^{(\to\infty)}$ has higher degree of replicability on new perturbed data sets.

The interpretation of the diluted causal parameter $\beta^{(\to\infty)}$ is, however, different from the usual least squares parameter and leads to invariance of residuals as described in Equation (18).

4 Discussion

4.1 Extensions

We have explained here the concepts for linear models only. Modifications for generalised linear models or non-linear models are certainly of interest. In the context of non-linear anchor regression, some methodological and algorithmic proposals have been illustrated empirically in Bühlmann (2020). In general, for models with non-linear regression functions, we can view the proposed methods as to perform deconfounding or distributionally robustifying the linear component of a general regression function. In fact, from a transfer learning perspective, for replicability on new data, it seems hard to go beyond linear extrapolation for strong perturbations arising in new data (Christiansen *et al.*, 2020).

Robustification and stabilising over different environmental conditions or different data sets from a 'causal structural equation model' point of view has been worked out also for independent component analysis (Pfister *et al.*, 2019b) or dynamical systems modeling (Pfister *et al.*, 2019a).

4.2 Summary

We have argued that deconfounding and causal regularisation (i.e. anchor regression) are powerful tools for improving replicability or distributional robustness; see Sections 2.4, 2.5.1, 3.2 and 3.3. For linear systems, the operational procedures are extremely simple and modular: it is just linearly pretransforming the data and then using any modern regression technique on such transformed data. Such pretransformations are also crucial for the important issue of 'attribution' (Efron, 2019, 2020): for high-dimensional densely confounded linear models, the doubly debiased Lasso (Guo *et al.*, 2020) leads to hypothesis tests and confidence intervals for the unconfounded (causal) parameter and thus, as an important consequence, to improved replicability even though the data are corrupted by latent perturbations or '(context) drifts' (Efron, 2020).

ACKNOWLEDGEMENTS

We thank an anonymous reviewer for insightful comments. The research of P. Bühlmann and D. Ćevid was supported by the European Research Council under the Grant Agreement 786461 (CausalStats - ERC-2017-ADG).

References

Angrist, J., Imbens, G. & Rubin, D. (1996). Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.*, **91**, 444–455.

Bareinboim, E. & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proc. Nat. Acad. Sci.*, **113**, 7345–7352.

Bertsimas, D., King, A. & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Ann. Stat.*, **44**, 813–852.

Bollen, K.A. (1989). *Structural Equations With Latent Variables*, Vol. **210**. New York: John Wiley & Sons.

Bowden, R. & Turkington, D. (1990). *Instrumental Variables*. Cambridge: Cambridge University Press.

Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Stat. Sci.*, **16**(3), 199–231.

Bühlmann, P. (2020). Invariance, causality and robustness (with discussion). *Stat. Sci.*, **35**, 404–426.

Ćevid, D., Bühlmann, P. & Meinshausen, N. (2018). Spectral deconfounding and perturbed sparse linear models. Preprint arXiv:1811.05352.

Chernozhukov, V., Hansen, C. & Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *Ann. Stat.*, **45**, 39–76.

Christiansen, R., Pfister, N., Jakobsen, M., Gnecco, N. & Peters, J. (2020). The difficult task of distribution generalization in nonlinear models. Preprint arXiv:2006.07433.

Cox, D.R. (2001). Comment on "Statistical modeling: The two cultures". *Stat. Sci.*, **16**(3), 216–218.

Dawid, A.P. (2020). Decision-theoretic foundations for statistical causality. Preprint arXiv:2004.12493.

Dawid, A.P. & Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Stat. Surv.*, **4**, 184–231.

Dezeure, R., Bühlmann, P., Meier, L. & Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R-software hdi. *Stat. Sci.*, **30**, 533–558.

Donoho, D. (2017). 50 years of data science. *J. Comput. Graph. Stat.*, **26**(4), 745–766.

Efron, B. (2001). Comment on "statistical modeling: the two cultures". *Stat. Sci.*, **16**, 218–219.

Efron, B. (2019). Prediction, estimation, and attribution. Talk on the occasion of receiving the International Prize in Statistics, see https://statprize.org/pdfs/materials/2019Efron‐presentation.pdf

Efron, B. (2020). Prediction, estimation, and attribution. *J. Am. Stat. Assoc.*, **115**, 636–655.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Stat.*, **32**, 407–451.

Efroymson, M. (1960). Multiple regression analysis. In *Mathematical Methods for Digital Computers*, Ed. M. Efroymson, pp. 191–203. New York: John Wiley & Sons.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Gao, R., Chen, X. & Kleywegt, A.J. (2017). Wasserstein distributional robustness and regularization in statistical learning. arXiv preprint arXiv:1712.06050.

Guo, Z., Ćevid, D. & Bühlmann, P. (2020). Doubly debiased Lasso: high-dimensional inference under hidden confounding and measurement errors. Preprint arXiv:2004.03758.

Hampel, F., Ronchetti, E., Rousseeuw, P. & Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**, 73–101.

Imbens, G.W. (2014). Instrumental variables: an econometrician's perspective. *Stat. Sci.*, **29**, 323–358.

Imbens, G. & Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Jakobsen, M. & Peters, J. (2020). Distributional robustness of K-class estimators and the PULSE. Preprint arXiv:2005.03353.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F. & Young, N. (2013). The genotype-tissue expression (GTEx) project. *Nat. Gen.*, **45**, 580–585.

Meinshausen, N. (2018). Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pp. 6–10. IEEE: Lausanne, Switzerland.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection (with discussion). *J. R. Stat. Soc. Ser. B*, **72**, 417–473.

Pan, S.J. & Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, **22**, 1345–1359.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, (2nd ed.). Cambridge: Cambridge University Press.

Peters, J., Bühlmann, P. & Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence interval (with discussion). *J. R. Stat. Soc. Ser. B*, **78**, 947–1012.

Pfister, N., Bauer, S. & Peters, J. (2019a). Learning stable and predictive structures in kinetic systems. *Proc. Nat. Acad. Sci.*, **116**, 25405–25411.

Pfister, N., Weichwald, S., Bühlmann, P. & Schölkopf, B. (2019b). Robustifying independent component analysis by adjusting for group-wise stationary noise. *J. Machine Learn. Res.*, **20**, 1–50.

Pfister, N., William, E., Peters, J., Aebersold, R. & Bühlmann, P. (2019c). Stabilizing variable selection and regression. Preprint arXiv:1911.01850.

Pratt, L.Y. (1993). Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 204–211.

Rojas-Carulla, M., Schölkopf, B., Turner, R. & Peters, J. (2018). Invariant models for causal transfer learning. *J. Machine Learn. Res.*, **19**, 1309–1342.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P. & Peters, J. (2018). Anchor regression: heterogeneous data meets causality. *To appear in J. R. Stat. Soc. (Ser. B)*. preprint arXiv:1801.06229.

Shah, R.D., Frot, B., Thanei, G.-A. & Meinshausen, N. (2020). Right singular vector projection graphs: fast high-dimensional covariance matrix estimation under latent confounding. *J. R. Stat. Soc. (Ser. B)*, **82**, 361–389.

Sinha, A., Namkoong, H. & Duchi, J. (2017). Certifiable distributional robustness with principled adversarial training. arXiv preprint arXiv:1710.10571. Presented at Sixth International Conference on Learning Representations (ICLR 2018).

Stock, J.H. & Trebbi, F. (2003). Retrospectives: who invented instrumental variable regression? *J. Econ. Perspect.*, **17**, 177–194.

Theil, H. (1958). *Economic Forecasts and Policy North-Holland*. Amsterdam: Netherlands.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.*, **42**, 1166–1202.

Yu, B. & Kumbier, K. (2020). Veridical data science. *Proc. Nat. Acad. Sci.*, **117**(8), 3920–3929.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, **38**, 894–942.

Zhang, C.-H. & Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B*, **76**, 217–242.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.

[*Received August 2020, Revised September 2020, Accepted October 2020*]