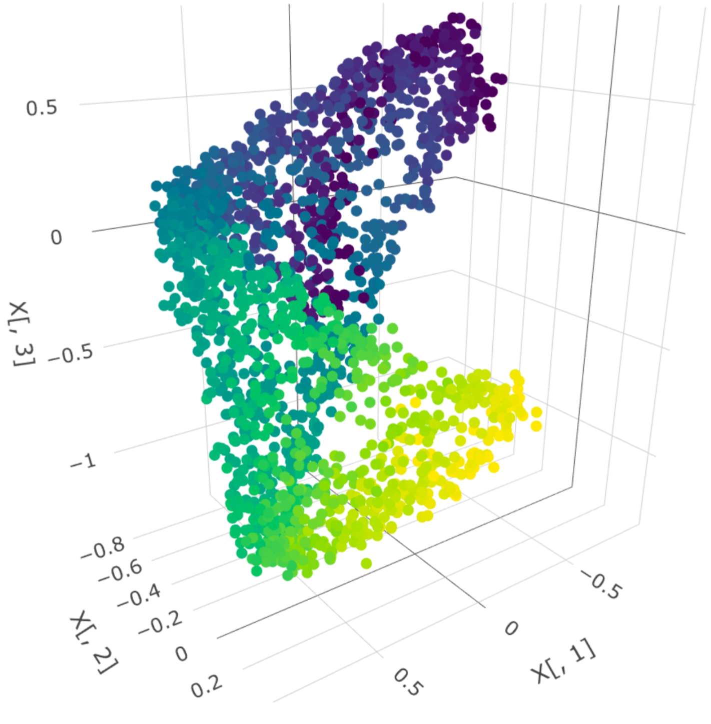
Confounding Adjustment for Causal Inference

Domagoj Ćevid



Diss. ETH No. 27870

Confounding Adjustment for Causal Inference

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZÜRICH
(Dr. sc. ETH Zürich)

presented by

Domagoj Ćevid

MMath, University of Cambridge
born on 03.12.1994
citizen of Croatia

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Nicolai Meinshausen, co-examiner

2021

DOI: ??????

ISBN: ??????

Abstract

The phenomenon of confounding, where both the treatment and the outcome variable of interest are affected by certain 'confounding' variables, is one of the biggest challenges for valid causal inference. It underpins many fallacies and misconceptions in statistics, such as Simpson's paradox or the examples where 'correlation does not imply causation'.

Therefore, confounding adjustment is at the heart of the field of causality. However, this is often not an easy task to do, even when the causal structure of our data is known. The dimensionality of the confounding variables can potentially be large, the confounders can be a mixture of discrete, continuous or categorical variables or they can affect the variable of interest in a non-parametric way.

There exist many different methods for confounding adjustment in the case when the confounding variables are known and observed in the data set at hand. However, very little research has considered the challenging case when the confounding is latent. Even though the assumption that there are no unobserved confounders is common in the causal literature, it often does not hold in practice. Such misspecification of the data model might lead to a decrease in performance of the conventional methods.

In this thesis we introduce novel methodology for confounding adjustment, addressing both the case when the confounding is unobserved, and the case when the confounding variables are observed, but their effect on the variables of interest is fairly complex and thus the conventional methods do not apply.

In **Paper A** we approach the problem of adjusting for latent confounding. Since this problem is extremely challenging, we consider the simple case where the data comes from a (high-dimensional) linear model and the confounding variables linearly affect the observed variables. We propose *Spectral Deconfounding* estimator which uses the standard Lasso after applying a carefully chosen linear transformation to the data. We derive interesting theoretical results and also empirically verify that it outperforms the conventional methods which ignore existence of latent confounding.

In **Paper B** we propose *Doubly Debiased Lasso* estimator, which can be viewed as the generalization of the Spectral Deconfounding estimator that has the advantage of having nicer asymptotic distribution, thus allowing for a construction of asymptotically valid

confidence intervals. The provided theoretical analysis is very elaborate and extends the theoretical results of Paper A.

Paper C considers an important problem in biostatistics of detecting the perturbations in the causal network between two conditions, such as, for example, cancer and normal cells. The proposed methodology is also extended to account for potential latent confounding. While it is not a direct application of the methods developed in Paper A and Paper B, it shares the main ideas developed there.

In **Paper D** we address the case when confounding is observed, but potentially very complex. We propose a versatile method called *Distributional Random Forests* that is able to non-parametrically estimate the multivariate joint conditional distribution. This is done in a model- and target-free way and can thus be used for many different learning problems beyond the original problem of confounding adjustment for causal effect estimation.

Résumé

Le phénomène de 'confounding' dans lequel la variable de traitement mais aussi la variable réponse sont toutes les deux afectées par certaines variables tierces (appelées variables de confounding) est un challenge en inférence. Beaucoup d'idées fausses sont liées à ce phénomène en statistiques, comme le paradoxe de Simpson ou encore les exemples illustratifs de l'adage 'une corrélation n'implique pas de causalité'.

L'ajustement pour ces variables de confounding est donc central en causalité. Cependant ce n'est pas toujours une tâche facile, même dans le cas où la structure causale des données est connue. La dimensionalité des variables de confounding est potentiellement large, ces variables peuvent être une mixture de variables discrètes, continues ou encore catégoriques. Elle peuvent aussi affecter la variable de réponse de manière non-paramètrique.

Il existe plusieurs méthodes d'ajustement quand les variables de confounding sont observées. Cependant peu de travaux existent sur le cas où les variables de confounding sont latentes. Même si l'hypothèse de latence des variables de confounding est souvent écartée dans la litératture, il se trouve qu'elle est souvent fausse dans les cas réels. En Faire abstraction peut amener une baisse de performance des méthodes conventionelles.

Dans cette thèse, nous introduisons une méthodologie nouvelle pour le problème de confounding applicable dans le cas où les variables de confounding ne sont pas observées et aussi dans le cas où les variables de confounding sont observées, mais leur effet sur les variables d'intérêt est suffisamment complexe pour que les approches classiques ne marchent pas.

Dans le **papier A**, nous étudions le problème d'ajustement pour des variables de confounding latentes. Comme ce problème est difficile, nous considérons le cas simple où les données proviennent d'un modèle linéaire (à haute dimension) et où les variables de confounding affectent linéairement les variables observées. Nous proposons un estimateur nommé *Spectral Deconfounding*, qui utilise la méthode classique du Lasso après avoir appliqué une transformation linéaire particulière aux données. Nous dérivons des résultats théoriques intéressants et nous vérifions de manière empirique que notre estimateur performe mieux que les méthodes classiques qui ignorent les variables de confounding latentes.

Dans le **papier B**, nous proposons un estimateur appelé *Doubly Debiased Lasso* qui peut être vu comme une généralisation de l'estimateur du papier A avec l'avantage en plus d'avoir de meilleurs propriétés asymptotiques qui permettent notamment de construire des intervalles de confiance. La théorie développée est élaborée et étend les résultats du papier A.

Dans le **papier C**, nous considérons un problème important en biostatistiques: la détection de perturbations dans un graphe causal entre les états, comme par exemple cellule cancéreuse ou normale. La méthodologie proposée est aussi étendue pour prendre en compte des potentielles variables latentes de confounding. Même si ce n'est pas une application directe des méthodes des papiers A et B, l'idée principale de la méthodologie est similaire.

Dans le **papier D**, nous nous intéressons au cas où les variables de confounding sont observées, mais dans lequel leur effet est potentiellement complexe. Nous proposons une méthode appelée *Distributional Random Forests* qui est une estimation non-paramétrique d'une distribution conditionnelle multivariée. La méthode ne requière pas d'hypothèse sur le modèle est n'est pas spécifique à une variable réponse particulière, ce qui ouvre un grand champ d'applications qui va plus loin que le problème de confounding en causalité.

Acknowledgments

I would like to first express my gratitude to my doctoral advisor Peter Bühlmann who has been great support throughout my time at ETH, both at the academic and at the personal level. With his great knowledge and experience, he has shown me the intricacies of the academic world and has guided me in my scientific pursuits.

Secondly, I would like to thank all my collaborators: Loris Michel, Nicolai Meinshausen, Jeffrey Näff, Zijian Guo, Kim Philipp Jablonski, Martin Pirkl, Niko Beerenwinkel, Rajen Shah and Matthias Löffler. Without their hard work, their motivation to create and the interesting and fruitful discussions we had, I would have had much harder and less enjoyable time during my PhD studies.

Furthermore, I am very happy to have had the pleasure to be surrounded by wonderful people at the Seminar für Statistik group. I thank all my colleagues and friends there and in particular Drago Plečko, Nicolas Bennett, Gian Thanei, Niklas Pfister, Yuansi Chen, Mona Azadkia, Geoffrey Chinot, Meta-Lina Spohn, Armeen Taeb, Emilija Perkovic, Jinzhou Li, Claude Renaux, Solt Kovacs and Dominik Rothenhäusler with whom I spent the most time and with whom I had the most enjoyable discussions. I specially thank the group secretaries Susanne Kaiser-Heinzmann and Letizia Maurer for their support and assistance with any issues that emerged.

Finally, my deepest gratitude goes to my entire family for their love and support and especially to my wife Paula who was with me every step of the way and with whom everything is much more meaningful.

Zürich, August 2021 Domagoj Ćevid

Contents

Abstract **iii**

Résumé **v**

1 Introduction **1**

2 Accompanying papers **9**

Paper A: Spectral Deconfounding via Perturbed Sparse Linear Models. 9

Paper B: Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding. 41

Paper C: Identifying cancer pathway dysregulations using differential causal effects 81

Paper D: Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression. 109

Bibliography **147**

Accompanying papers

A Spectral Deconfounding via Perturbed Sparse Linear Models
Ćevid, D., Bühlmann, P., Meinshausen, N.
Journal of Machine Learning Research 21 (2020): 232.

B Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding.
Guo, Z., Ćevid, D., Bühlmann, P.
Submitted to Annals of Statistics.

C Identifying cancer pathway dysregulations using differential causal effects.
Jablonski, K. P., Pirkl, M., Ćevid, D., Bühlmann, P., Beerenwinkel, N.
Submitted to Bioinformatics.

D Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression.
Ćevid, D., Michel L., Näff J., Bühlmann P., Meinshausen N.
Submitted to Journal of Machine Learning Research.

1 Introduction

Determining the causal relationships between different events, processes or states of objects is at the centre of scientific endeavour and human intelligence overall. Knowledge of cause and effect enables us to understand different mechanisms of nature more deeply, to be able to transfer our knowledge to new and unseen situations and to answer in general what would happen if certain action was performed, i.e. if we intervened on the observed system.

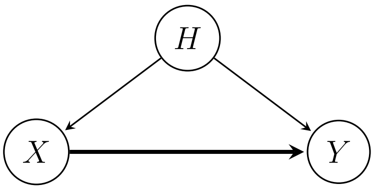
Despite being so fundamental, the discipline of causality has not been considered as an essential part of statistics for a long time. However, understanding only the associations between random variables through their observational distribution is often not enough, as we would like to derive useful conclusions from our data and to be able to transfer the obtained knowledge to other situations. For example, one might observe from the data that the population of a city and the air pollution tend to be highly (positively) correlated. However, based only on this data, we can not determine which of the following 3 scenarios is correct:

- the pollution is directly caused by the people,
- high air pollution at certain places causes people to move there,
- some other variable, such as e.g. the number of factories, causes both the air pollution to be high and people to settle nearby.

It has been only relatively recently, through work of many great statisticians (Rubin, 2005; Pearl, 2009), that the field of causality has been formalized within the statistical framework. Since then the field has been rapidly developing and the causal concepts have proved to be useful in a variety of different fields such as econometrics, finance, machine learning, biostatistics and many others. However, many challenges still remain to be solved and this thesis is hopefully a small step in that direction.

Confounding

Suppose we have a treatment variable $X \in \mathbb{R}^p$ affecting the response variable $Y \in \mathbb{R}$ and that we would like to determine the causal effect of X on Y. Any other variable $H \in \mathbb{R}^q$ affecting both X and Y is called a *confounder*. This is illustrated in the following plot:



If there are no confounding variables, the causal effect of X on Y can easily be determined directly from the observed conditional distribution $\mathbb{P}(Y \mid X)$. However, in presence of confounding, the observed distribution of X and Y is not a reliable indicator of the causal relationship between them. This is a common fallacy known under many names, such as "correlation does not imply causation".

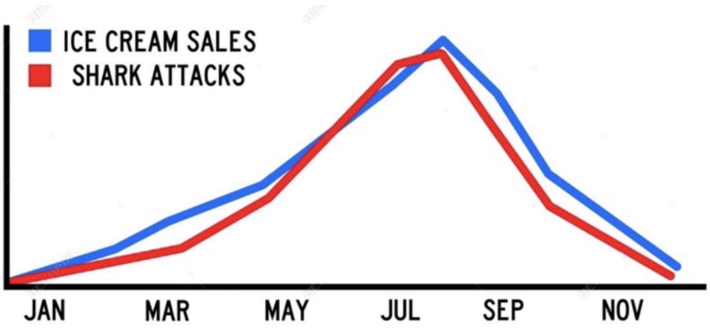


Figure 1.1: The average number of shark attacks and the ice cream consumption per calendar month are very correlated. However, this relationship is not causal but is due to confounding: hot weather is directly causes higher ice cream consumption and people to swim in the sea more, leading to increased number of shark attacks.

One funny example of this is given in Figure 1.1. However, sometimes confounding issues are not that obvious and could potentially be very subtle. For example, assume that some medication is effective against some lethal disease, but is much more likely to be prescribed to the most ill patients (maybe due to some serious side-effects). Then we might observe that mortality is higher in the group of patients who were treated with this medication. However, it is wrong to deduce from the data that this medication increases the mortality rate. In this example, the severity of the disease is a confounder as it simultaneously increases the mortality and the probability of being treated.

Confounding adjustment is necessary for determining the causal effects from the data and is thus at the heart of the causal inference. Confounding causes many difficulties in statistical analysis, such as false positives in model selection. Another problem is bad transferability of the fitted models to new environments, where the confounding mechanism changes. Therefore, our models need to take confounding into account and thus to be more 'causal' in order to have better robustness properties.

Confounding Adjustment

There exist many different ways to adjust for the confounding effects, depending on the causal structure and our model for the data generating mechanism. One of the most commonly used general approaches is based on the back-door adjustment formula (Pearl, 2009)

$$\mathbb{P}(Y \mid do(X=x)) = \int \mathbb{P}(Y \mid X=x, H=h) d\mathbb{P}(H=h), \tag{1.1}$$

which relates the observational conditional distribution to the interventional distribution, i.e. the distribution of the response Y if we had forcibly set the predictors X to attain value x. The back-door adjustment formula is related to data stratification with respect to the confounding variable, and aggregating the inferred causal effect over different strata.

However, one does not have a direct access to the observational distribution, but only to the data drawn from this distribution. Therefore, using the adjustment formulae such as (1.1) is not that straightforward and can be quite challenging. Many conventional methods such as regression adjustment, inverse propensity weighting (Rubin, 2005), propensity score matching or some modern ones such as double machine learning (Chernozhukov et al., 2018) or causal forests (Athey et al., 2019) assume some special structure, such as linearity or additivity of the true signal and the confounding effect; that the treatment or outcome variables are univariate and binary or maybe that the confounding variable is discrete or very low-dimensional.

Distributional Random Forests

In Paper D of this thesis we introduce a method called *Distributional Random Forests* (DRF), which is able to non-parametrically estimate any multivariate conditional distribution and thus can be used for many different applications in causality, such as confounding adjustment in an arbitrarily complex model. For example, if we are able to estimate the conditional distribution $\mathbb{P}(Y \mid X, H)$ well, we can use formula 1.1 and Monte-Carlo algorithm to estimate the causal effect $\mathbb{P}(Y \mid do(X=x))$, see Paper D for mode details.

However, DRF is very versatile and can be used for a variety of applications, as illustrated in the paper. It is based on the standard Random Forest algorithm (Breiman, 2001), but where the splits are performed based on some distributional metric, i.e. we split such that the difference in the distribution of the response in the left and right child node is the largest. Having constructed our forest, we can obtain for any point of interest $\mathbf{x}$ a weighting function $w_{\mathbf{x}}(\cdot)$ which assigns a weight to each training point indicating how relevant it is for the given test point $\mathbf{x}$. The weighting function can be used in the second step for computing any target of interest, as illustrated in the following diagram:

$$\begin{tikzcd}
\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}) \arrow[r, "\text{1) get } w_{\mathbf{x}}(\cdot) \text{ with DRF}"] \arrow[d, dashed, "\text{objective}"'] & \hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}) \arrow[d, "\text{2) compute from } w_{\mathbf{x}}(\cdot)"] \\
\tau(\mathbb{P}) & \tau(\hat{\mathbb{P}}) \arrow[l, dashed, "\text{induced estimator}"']
\end{tikzcd}$$

Unobserved Confounding Adjustment

Almost all confounding adjustment approaches require the confounding variables to be observed. The case when the confounding is unobserved has not received much attention in the literature, mostly because this case is so difficult that at first one might think that it is not possible to address. After all, how could one adjust for something that is not observed? This is why almost all problems in causal literature are solved under the assumption that there is no latent confounding.

However, such assumption might not hold in practice. It is very plausible that some unobserved external variables, such as e.g. demographic factors, laboratory conditions or batch effects, could affect our data and thus to introduce spurious associations. Therefore, it is very important to address for potential latent confounding. However, since the data need not be confounded, our method needs to have a comparable performance to the conventional methods which ignore existence of confounding. A significant portion of this thesis considers the problem of attaining robustness against potential hidden confounding. Since this problem is very difficult, we start with some simpler, linear, models but explain later how those ideas could be generalized to more complicated models, which we leave for future research.

Linear Factor Model

We consider first the simplest confounding model, where the confounding variables linearly affect the observed covariates. We assume that the predictors $X \in \mathbb{R}^p$ are generated as follows

$$X \leftarrow \Psi^T H + E, \tag{1.2}$$

where $\Psi \in \mathbb{R}^{q \times p}$ is the loading matrix of coefficients and $E \in \mathbb{R}^p$ is the random error term.

It is evident that, up to the noise term E, the covariates lie on a q-dimensional hyperplane in $\mathbb{R}^p$, spanned by the rows of Ψ. This is illustrated in Figure 1.2.

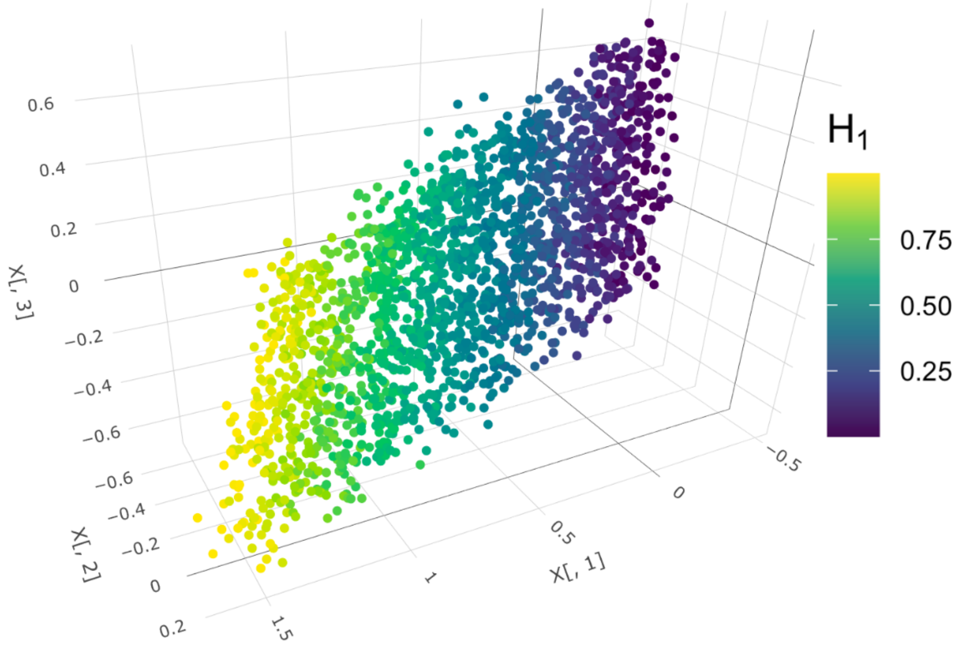


Figure 1.2: When the confounders affect the covariates linearly, as in (1.2), our data lies approximately on a low-dimensional hyperplane.

Principal Component Analysis (PCA) (Price et al., 2006) tries to find linear combinations of the predictors that capture the most variability of the data. If our data approximately lies on the low-dimensional hyperplane, top several (around q) principal components will account for a lot of variability. This is very related to the spiked covariance structure (Paul, 2007), where the first few singular values of the covariance matrix are much larger than the rest. By inspecting singular values of the data matrix we can also test whether the data is confounded, i.e. whether it comes from the Linear Factor Model (1.2). It has been known for a long time in the field of biostatistics that top principal components contain some information about the confounding effects (Leek and Storey, 2007; Gagnon-Bartsch et al., 2013), see also Figure 1.3.

In Paper A, we develop *Spectral Deconfounding* estimator for the special case where the predictors follow the Linear Factor Model (1.2) and the response variable comes from the linear model

$$Y = \beta^T X + \delta^T H + \epsilon,$$

where $\beta \in \mathbb{R}^p, \delta \in \mathbb{R}^q$ are coefficient vectors and ϵ is a random error. It can be viewed as a standard Lasso estimator applied on the transformed data:

$$\widehat{\beta} = \arg\max_{\beta} \frac{1}{n} \|FY - FX\beta\|_2^2 + \lambda \|\beta\|_1,$$

where the spectral transformation matrix $F \in \mathbb{R}^{n \times n}$ is chosen such that it transforms

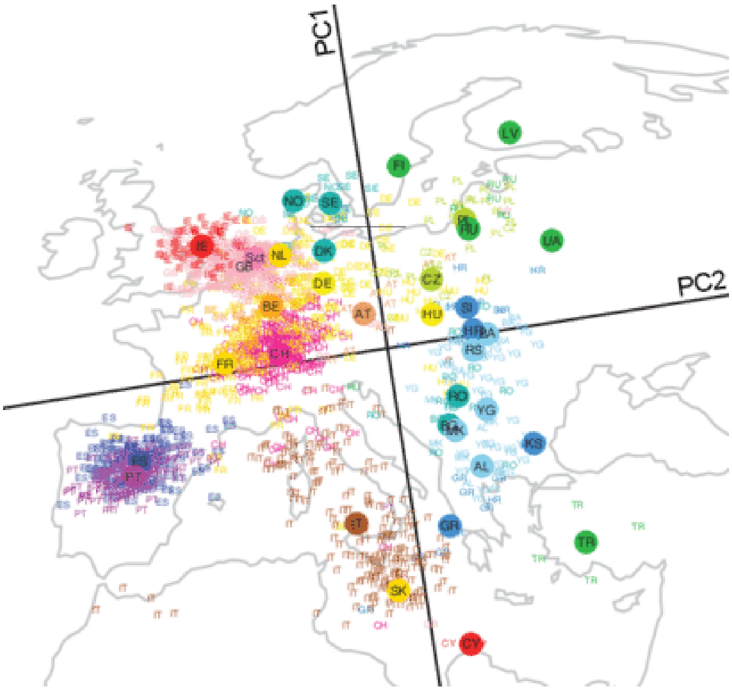


Figure 1.3: Top principal components capture information about confounding. In this example, the first 2 principal components of the gene expression matrices match closely the geographic distribution of the samples. The plot is borrowed from Novembre et al. (2008).

the singular values of the design matrix X. By default, we propose the *trim transform* which caps all singular values at a given threshold (e.g. the median singular values). The intuition is that the confounding is captured by the first several spiked singular values and shrinking them helps to reduce the confounding effects. In Paper A, we rigorously show that under some assumptions, one can get the same error rate as the Lasso for the data model without confounding. Additionally, we provide simulation results which empirically verify that in presence of latent confounding, we outperform standard Lasso for coefficient estimation.

In Paper B, we propose the *Doubly Debiased Lasso* estimator for the same data model considered in Paper A. It is analogous generalization of the Spectral Deconfounding estimator as the Debiased Lasso (Zhang and Zhang, 2014) generalizes the standard Lasso (Tibshirani, 1996). It has the advantage that its asymptotic distribution is nicely behaved which enables us to construct asymptotically valid confidence intervals. However, the performance of the plain Debiased Lasso deteriorates in the presence of latent confounding. On the other hand, Doubly Debiased Lasso provides additional robustness against hidden confounding. This is achieved by applying carefully chosen spectral transformations, analogously as in Paper A, for both the initial estimator and the construction of the proposed estimator. The main emphasis of the paper is on the rigorous theoretical analysis of the estimation error and the asymptotic distribution of the estimator and the obtained

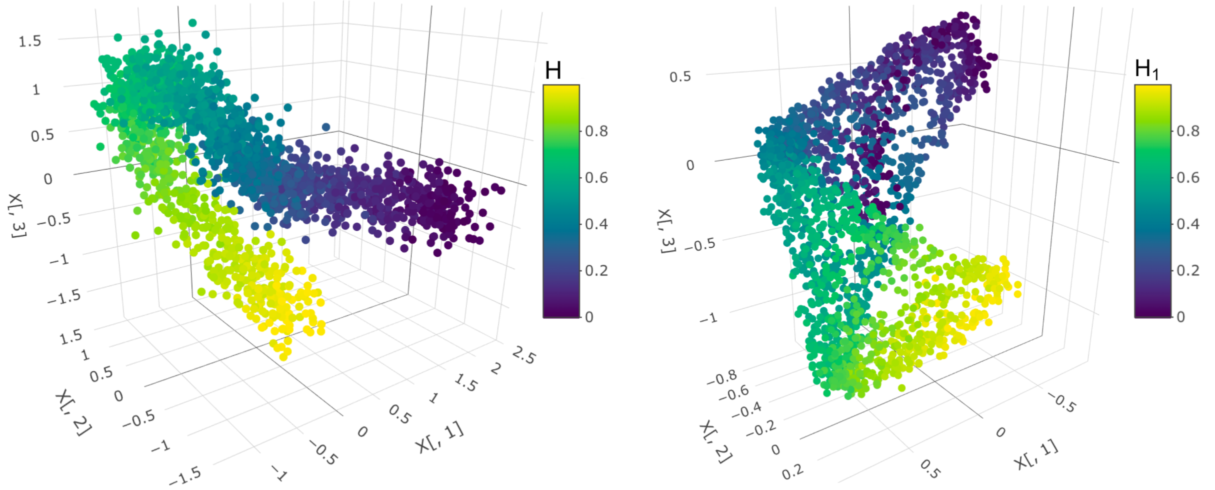


Figure 1.4: When the predictors are affected by a small number of confounding variables, the data lies approximately on a low-dimensional manifold. In the left plot we have only one confounding variable, whereas in the right plot there are two. In both cases the confounding variables affect the predictors in a highly non-linear way, but one can still recover some information about the latent confounders from the data.

results generalize also the results obtained in Paper A.

Extensions of the Methodology

As we have seen, top several principal components capture a lot of information about the confounding in the Linear Factor Model (1.2). Therefore, one can compute the corresponding scores of the principal components and use those variables in the further analysis as the proxies for the confounding variables.

For example, in Paper C we extend the proposed *Differential Causal Effects* method to be more robust against hidden confounding. Given a biological pathway (i.e. the causal graph) and the gene-expression data from two conditions, the goal is to determine which part of the network has been dysregulated between the conditions, for example between cancerous and normal cells. This is done by performing nodewise regression with interaction terms in order to detect the changes in edge weights. Computing the scores of principal components and adding them to the regression as the source nodes in the pathway helps to reduce the confounding bias which can lead to false findings.

Similar idea can be used to generalize our approach to more complicated, nonlinear models. When a small number of confounding variables affects a large number of the observed covariates, our data will approximately lie on a low-dimensional manifold, just as it lies on the low-dimensional hyperplane in the linear case (1.2). This is illustrated in Figure 1.4.

This manifold structure can be used to get some information about the latent variables, which can be used for confounding adjustment. More specifically, one can first apply some manifold learning (or dimensionality reduction) algorithm from the machine learning literature in order to obtain confounding proxies, just as one can use the principal components scores in the linear case. Those confounding proxies can in turn be used in the downstream analysis in order to adjust for the latent confounding. Exploring this approach in more detail is left for future research.

Paper

A

Spectral Deconfounding via Perturbed Sparse Linear Models.

Ćevid, D., Bühlmann, P., Meinshausen, N.

Journal of Machine Learning Research 21 (2020)

Spectral Deconfounding via Perturbed Sparse Linear Models

Domagoj Ćevid*, Peter Bühlmann*, Nicolai Meinshausen*

*Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland.

Abstract

Standard high-dimensional regression methods assume that the underlying coefficient vector is sparse. This might not be true in some cases, in particular in presence of hidden, confounding variables. Such hidden confounding can be represented as a high-dimensional linear model where the sparse coefficient vector is perturbed. For this model, we develop and investigate a class of methods that are based on running the Lasso on preprocessed data. The preprocessing step consists of applying certain spectral transformations that change the singular values of the design matrix. We show that, under some assumptions, one can achieve the usual Lasso ℓ_1-error rate for estimating the underlying sparse coefficient vector, despite the presence of confounding. Our theory also covers the Lava estimator (Chernozhukov et al., 2017) for a special model class. The performance of the methodology is illustrated on simulated data and a genomic dataset.

Keywords. confounding, data transformation, Lasso, latent variables, principal components

1 Introduction

Many datasets nowadays include measurements from many variables. The corresponding models are typically high-dimensional with many more parameters than the sample size. For statistical estimation and inference, there is a vast literature which assumes sparsity. For example, see the monographs by Bühlmann and van de Geer (2011), Giraud (2014) or Hastie et al. (2015).

However, the performance of many high-dimensional regression methods might suffer in presence of unobserved confounding variables which affect both the predictors and the response. Confounding is a severe issue when interpreting regression parameters, often, but not necessarily, in connection with causal inference. A prime example are genetic studies where unobserved confounding can easily lead to spurious correlations and partial dependencies (Novembre et al., 2008). Even when one is concerned with only prediction, the causal parameter leads to predictive robustness against perturbations of the confounding variables.

Adjusting for unobserved confounding variables is very important in practice and several deconfounding methods have been suggested for various settings (Gerard and Stephens, 2017; Leek and Storey, 2007; Gagnon-Bartsch and Speed, 2012; Wang and Blei, 2018; Paul et al., 2008). Often, the methods try to estimate the confounding variables directly from the data, usually by using some factor analysis technique. There are not many theoretical results justifying the methods, especially since some of them are quite complicated and therefore difficult to analyze.

Our focus is on linear models. In absence of confounding variables, when the response is affected only by a small number of predictors, i.e. the coefficient vector is sparse, one can efficiently estimate the active set and the corresponding coefficients with the Lasso and related methods and thus achieve the minimax optimal ℓ_1-norm estimation error rate, see, for example, Bickel et al. (2009) or the monographs by Bühlmann and van de Geer (2011) or Wainwright (2019). However, these methods are not adequate in presence of confounding in linear model, since in addition to just a few predictors that indeed affect the response, we have additional association of the response with many other predictors, as they contain information about the confounding variables.

Some approaches for relaxing the sparsity assumption are (i) the notion of weak sparsity (Van de Geer, 2016), where the regression parameter β fulfills the condition that $\|\beta\|_q$ is small for some $0 < q < 1$ or (ii) assuming the structure that the regression parameter can be represented as a sum of a sparse and a dense vector. The case (i) does not call for a new method or algorithm: in fact, the Lasso still exhibits optimal convergence rate if $\|\beta\|_q$ is sufficiently small (Van de Geer, 2016). On the other hand, case (ii) requires a different method such as, for example, Lava (Chernozhukov et al., 2017).

Here we investigate how to deal with the confounding by analyzing the second case where the parameter is a sum of a sparse and a dense part. If many predictors are affected by the confounding variables, the true underlying regression vector will be changed by some small, dense perturbation. We propose left multiplying the response Y and the design matrix X consisting of the values of the predictors by a carefully chosen spectral transformation matrix F which transforms the singular values of X. The transformed response and design matrix can then be used as the input for a high-dimensional sparse regression technique: we consider the Lasso as a prime example. We investigate the theoretical properties and empirical performances for the class of spectral transformations. As a result, we conclude that certain spectral transformations that shrink the large singular values, such as the Trim transform which we introduce in this paper, perform well over a range of scenarios, pointing out also some advantages over other techniques and approaches.

1.1 Relation to Other Work and Our Contribution

For adjusting for the effect of unobserved confounding, the most prominent method in practice is to adjust for the top several principal components of the predictors, see for example (Novembre et al., 2008). Such PCA adjustment is also a special case of the FarmSelect estimator (Fan et al., 2020) for the linear model, which considers the problem of high-dimensional variable selection where the latent variables cause the correlations of the predictors, but do not directly affect the response. PCA adjustment is a special case of a spectral transformation. Our presented theory explains when and why this method works well and proposes an alternative transformation, called Trim transform, which has an advantage that one does not need to estimate the number of principal components to adjust for.

The Puffer transform, which maps all singular values to 1, has also been suggested for improving the variable selection properties of the Lasso for a sparse high-dimensional linear model (Jia et al., 2015). Our theory gives a more precise result about the Puffer transform for the estimation problem: the Trim transform is at least as good as Puffer transform and substantially better when the sample size is close to the number of predictors. In Shah et al. (2020), the Puffer transform in combination with bootstrap aggregation is used in order to estimate the covariance matrix in presence of confounding variables, a very different quantity than the precision matrix or regression coefficients.

Chandrasekaran et al. (2012) address the problem of estimating the precision matrix in presence of a few hidden confounding variables. Then the observed precision matrix can be represented as a sum of the initial sparse precision matrix and a low-rank perturbation due to the confounding variables. Their model is similar to the one we consider, but the assumptions and the goals differ. We aim to estimate just the regression coefficients instead of the whole precision matrix and the method we propose is much simpler. Furthermore, the theoretical conclusions are substantially different: we establish the convergence rates in terms of the ℓ_1-norm estimation error, while they consider support recovery and ℓ_∞-norm bounds for the low-dimensional setting, assuming strong conditions. Also Fan et al. (2013) have considered low rank plus sparse problems from the viewpoint of factor models: their contribution provides a rich source of references from an area which is vaguely related to our current work.

The Lava estimator (Chernozhukov et al., 2017) is the most similar to our Trim transform. The theory we develop, covering also the Lava, gives a result for the ℓ_1-norm estimation error rate for the sparse coefficient vector. This goes well beyond the theory given by Chernozhukov et al. (2017) for justifying the original and interesting Lava method. There, the authors mostly consider the Gaussian sequence model but also provide general bounds for high-dimensional regression whose (e.g. asymptotic) behavior is not further analyzed in terms of restricted eigenvalues and the sparse and dense component of the underlying unknown parameter vector. Our presented theory exploits the specific structure of a hidden confounding model which provides a different motivation than the one in Chernozhukov et al. (2017), where no confounding was considered. In addition, our developments suggest a simple rule for the choice of the ℓ_2-norm regularization parameter for the Lava estimator, leaving only the ℓ_1-norm regularization parameter as the single parameter to be tuned by cross-validation.

Our contribution can be seen as threefold. We describe a class of spectral transformations and propose a simple spectral transformation called Trim transform, which is perhaps slightly easier to use than the Lava or the PCA adjustment estimator. Furthermore, for the linear model where the underlying sparse parameter has been perturbed, we provide novel theory establishing for a certain class of spectral transformations a fast convergence rate for the ℓ_1-norm estimation error of the true underlying sparse parameter. Finally, and as our primary goal, we use these results to show how the issue of hidden confounding can be addressed by using a wisely chosen spectral transformation, such as e.g. Trim transform, with the Lasso afterwards: we establish under certain assumptions the same convergence rate as the one of the Lasso for a linear model without confounding and illustrate the empirical performance of our method on simulated and real genomic data. Our method is entirely modular and can be used not only in conjunction with the Lasso, but also any other reasonable high-dimensional linear regression method.

2 The Models

In this section we consider a linear model with additional confounding. We also introduce a perturbed linear model and show how it relates to the confounding model. Our theoretical results apply to the perturbed linear model as well and it is useful for better understanding of the confounding model.

2.1 Confounding Model

Consider a standard (high-dimensional) linear model with n observations and p predictors $X_1, \ldots, X_p$ linearly affecting the response Y. Suppose further that q additional unobserved confounding variables linearly affect the response as well. The confounding variables are correlated with the predictors, introducing additional spurious correlations between the response and the predictors.

The model for n i.i.d. observations is given by:

$$Y = X\beta + H\delta + \nu \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$ is the matrix of predictors and $H \in \mathbb{R}^{n \times q}$ represents the hidden confounding variables, which exhibit correlation with X, i.e., $\mathrm{Cov}(H, X) \neq 0$ (with a slight abuse of notation, we write $\mathrm{Cov}(H,\, X)$ as the covariance of any row of H and X). We assume that X and H have i.i.d. rows that are jointly Gaussian and that $\nu \in \mathbb{R}^n$ is a vector of sub-Gaussian errors with mean zero and standard deviation σ_ν, independent of X and H. The vectors $\beta \in \mathbb{R}^p$ and $\delta \in \mathbb{R}^q$ are fixed coefficients; we additionally assume that β is sparse with exactly s non-zero components. Since the model does not change under the transformation $H \leftarrow H\mathrm{Cov}(H)^{-1/2}$, $\delta \leftarrow \mathrm{Cov}(H)^{1/2}\delta$, we can assume without loss of generality that $\mathrm{Cov}(H) = I_q$, i.e. the confounding variables are uncorrelated.

Note that by L_2 projection, X can also be written as

$$X = H\Gamma + E, \tag{2}$$

where we choose $\Gamma \in \mathbb{R}^{q \times p}$ such that $\mathrm{Cov}(H, E) = 0$:

$$\Gamma = \mathrm{Cov}(H)^{-1}\mathrm{Cov}(H, X) = \mathrm{Cov}(H, X).$$

The matrix $\Gamma \in \mathbb{R}^{q \times p}$ describes the linear effect of confounding variables on X. The random term $E \in \mathbb{R}^{n \times p}$ can be seen as the unconfounded design matrix; without confounding, i.e. when $H = 0$, it equals X. The columns of E are allowed to be correlated and we denote its covariance matrix by Σ_E; if the components of E are (weakly) uncorrelated, X is generated from an (approximate) factor model (Anderson, 1958; Chamberlain and Rothschild, 1982). Here the hidden variables do not encode a factor structure for X alone, but also in addition generate confounding effects.

A main example of the above model is a structural equation model (SEM)

$$\begin{aligned} X &\leftarrow H\Gamma + E, \\ Y &\leftarrow X\beta + H\delta + \eta \end{aligned}$$

and thus β is the direct causal effect of X on Y. In a standard SEM with no further hidden variables, the components of E would be assumed independent.

We will show in Section 4 that one can recover the coefficient β if the confounding is dense in a certain sense, e.g. when the rows or columns of $\Gamma = \mathrm{Cov}(H, X)$ are realizations of independent and identically distributed random variables with mean zero.

2.2 Perturbed Linear Model

The confounding model (1) is related to the perturbed linear model

$$Y = X(\beta + b) + \epsilon, \tag{3}$$

where the sparse coefficient vector β has been perturbed by the perturbation vector $b \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^n$ is the vector of sub-Gaussian errors independent of X with standard deviation σ. Here we assume that the rows of X are i.i.d. sub-Gaussian vectors with mean zero and covariance matrix $\Sigma = \mathrm{Cov}(X)$.

The relationship between models arises by rewriting (1) as

$$Y = X(\beta + b) + (H\delta - Xb) + \nu,$$

where b satisfies that $\mathrm{Cov}(X, H\delta - Xb) = 0$, i.e., Xb is the L_2-projection of $H\delta$ onto X. This gives us the formula

$$\begin{aligned} b &= \mathrm{Cov}(X)^{-1}\mathrm{Cov}(X,H)\delta \\ &= \big(\mathrm{Cov}(X,H)\mathrm{Cov}(H)^{-1}\mathrm{Cov}(H,X) + \mathrm{Cov}(E)\big)^{-1}\,\mathrm{Cov}(X,H)\delta \end{aligned} \tag{4}$$

The error is given by $\epsilon = (H\delta - Xb) + \nu$, which by construction of b is uncorrelated with X and thus independent of X, because the rows of X and H are assumed to be jointly Gaussian in the confounding model. We require such independence (induced by joint Gaussianity) in the proof of Theorem 2, although ϵ being uncorrelated with X might be sufficient. The variance of the error is given by

$$\sigma^2 = \mathrm{Var}(H\delta - Xb + \nu) \leqslant \|\delta\|_2^2 + \sigma_\nu^2.$$

One can think of $H\delta - Xb$ as the part of the confounding that can not be explained by X and which just increases the variance of the additive error. Xb is the part of the confounding effect $H\delta$ that is correlated with X and, as is well known, the bias b due to the confounding makes the estimation of β more difficult.

In conclusion, the confounding model (1) can be thought of as a special case of the perturbed linear model (3), but with additional relationship between the design matrix X, the perturbation vector b, given by (4), and the additive error ϵ.

The perturbed linear model is in general unidentifiable since we can only infer $\beta + b$ from the data generating distribution. This makes the estimation of β impossible, unless b has a certain structure; we will be able to asymptotically retrieve the sparse coefficient vector β, by assuming, for example, that b converges to 0 in some norm. In Section 4, we investigate under which conditions we are able to infer the sparse part β and how efficiently in terms of statistical accuracy.

It could be interesting to estimate the coefficient vector $\beta + b$ rather than just β, but it is impossible to do in general in the high-dimensional case; even if we knew β exactly, estimating b would mean estimating p coefficients from $n < p$ data points, which is impossible without additional assumptions about the structure of b.

2.3 Relationship with the Factor Model Literature

Even though the confounding variables are hidden, we are able to infer some of their properties if they affect many of the observed predictors X. This is the essence of factor analysis, where a lot of interesting work has been done. If the latent factors H linearly affect the covariates, as it is the case in the confounding model (1), they can be estimated well (up to a rotation) from the principal components of the design matrix $X = H\Gamma + E$ (Chamberlain and Rothschild, 1982; Bai, 2003), especially if one additionally imposes certain assumptions on the factor loadings Γ (Bing et al., 2017).

There are several related models considered in the literature. In certain cases (Paul et al., 2008; Bing et al., 2019) we assume that only the latent factors affect the response and the observed covariates are only used to obtain information about the latent factors:

$$Y = H\delta + \nu, \qquad X = H\Gamma + E.$$

In Bai and Ng (2006) one has an additional contribution of some other known low-dimensional covariates W:

$$Y = W\beta + H\delta + \nu, \qquad X = H\Gamma + E.$$

Another line of work assumes that the latent factors do not directly affect the response:

$$Y = X\beta + \nu, \qquad X = H\Gamma + E,$$

but that they only cause the predictors to be correlated (Huang and Jojic, 2011; Fan et al., 2020). Such correlation makes the analysis much more difficult, especially for the problem of variable selection, and one can use the factor analysis to address this issue.

In this paper we allow the latent confounders to affect both the predictors and the response and focus on the estimation of the sparse coefficient vector β, which has a causal interpretation as it describes the direct effect of the predictors on the response. The key difficulty is to handle the bias b in the observational data caused by the latent confounders. The assumption of dense confounding, expressed in detail in Section 4, is related to the spiked covariance assumptions common in the factor analysis literature (Paul et al., 2008; Bai, 2003). It is used to make conclusions about the structure of the coefficient perturbation b rather than about the factor identifiability. We avoid estimating the factor variables directly, but instead we adjust for them implicitly, by transforming the singular values of X.

3 Methodology

In the following, we propose and motivate some methods based on a class of spectral transformations.

3.1 Spectral Transformations

Let $X = UDV^T$ be the singular value decomposition of X, where $U \in \mathbb{R}^{n\times r}, D \in \mathbb{R}^{r\times r}, V \in \mathbb{R}^{p\times r}$, where $r = \min(n,p)$ is the rank of X. We write $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_r$ for the diagonal elements of D. We use the truncated form of SVD, which uses only non-zero singular values.

The idea is to first transform our data by applying some specific linear transformation $F : \mathbb{R}^n \to \mathbb{R}^n$ and then perform the Lasso algorithm:

$$\begin{gathered} X \to \widetilde{X} := FX \\ Y \to \widetilde{Y} := FY \\ \widehat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n} \|\widetilde{Y} - \widetilde{X}\beta\|_2^2 + \lambda\|\beta\|_1 \right\}. \end{gathered} \tag{5}$$

We restrict our attention to the class of spectral transformations, which transform the singular values of X, while keeping its singular vectors intact. Let $\widetilde{D}$ be an arbitrary $r \times r$ diagonal matrix with diagonal elements $\widetilde{d}_1, \ldots, \widetilde{d}_r$. Our spectral transformation matrix is given by

$$F = U \begin{bmatrix} \widetilde{d}_1/d_1 & 0 & \ldots & 0 \\ 0 & \widetilde{d}_2/d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \widetilde{d}_r/d_r \end{bmatrix} U^T \tag{6}$$

and then we have

$$\widetilde{X} = FX = U\widetilde{D}V^T$$

In this paper we explore the question of what is a good choice of F for the estimation of β. In general, the Lasso performs best when the predictors are uncorrelated and when the errors are independent. Therefore, a good choice of F needs to find a good balance between a well behaved error term $\widetilde{\epsilon} = F\epsilon$, well behaved design matrix $\widetilde{X}$ and well behaved perturbation term $\widetilde{X}b$.

One such transformation is the **Trim transform** which limits all singular values to be at most some constant τ:

$$\widetilde{d}_i = \min(d_i, \tau). \tag{7}$$

We show in Section 4 that it can, under some assumptions, achieve the same ℓ_1-norm error rate for the estimation of the unknown sparse coefficient vector β as the Lasso in the case of no confounding. We also show that the median singular value is a good choice of τ:

$$\tau = d_{\lfloor r/2 \rfloor}$$

3.2 Existing Methods and Motivation

We discuss some existing methods which are related to the spectral transformation method described above and provide further explanations and relationships between them. We also present intuitive explanation why our suggested method should work well against dense confounding.

3.2.1 Examples of Spectral Transformations

Several existing methods consist of first transforming the data with a certain matrix F (some of which fall into class of spectral transformations (6)), and then using some regression method, such as the Lasso.

Lava One such example is the Lava estimator (Chernozhukov et al., 2017), designed for the linear model where the coefficient vector can be written as a sum of a dense and a sparse vector. It is originally given by (with a slight change of notation)

$$(\widehat{\beta}, \widehat{b}) = \arg\min_{\beta, b} \left\{ \frac{1}{n} \|Y - X(\beta + b)\|_2^2 + \lambda_2 \|b\|_2^2 + \lambda_1 \|\beta\|_1 \right\},$$

which can be seen as a combination of Lasso and Ridge regression. It is shown in Chernozhukov et al. (2017) that the solution of this optimization problem is given by

$$\begin{aligned} F &= (I_p - X(X^T X + n\lambda_2 I_p)^{-1} X^T)^{1/2}, \\ \widehat{\beta} &= \arg\min_{\beta} \left\{ \frac{1}{n} \|\widetilde{Y} - \widetilde{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}, \\ \widehat{b} &= (X^T X + n\lambda_2 I_p)^{-1} X^T (Y - X\widehat{\beta}). \end{aligned}$$

From here, one can see that the estimator of the sparse part is just a Lasso estimator applied to the transformed data, where

$$\widetilde{d}_i = \sqrt{\frac{n\lambda_2 d_i^2}{n\lambda_2 + d_i^2}}.$$

This transformation is visualized in Figure 3.

Puffer transform Another example is the Puffer transform introduced in Jia et al. (2015), which uses the Lasso after mapping all non-zero singular values d_i to a constant $\widetilde{d_i} = 1$. The algorithm is analyzed as a preconditioning method for the variable selection problem without any coefficient perturbation. This transformation decreases the correlations between the columns of the design matrix, but it can inflate the errors, especially when p is close to n. It can also be thought of as a special case of the Lava transformation in the case when $\lambda_2 \to 0$, since then $\frac{\widetilde{d_i}}{\sqrt{n\lambda_2}} \to 1$ (the denominator here is just a scaling factor). The transformation is displayed in Figure 3.

PCA adjustment Another example of a spectral transformation is given by PCA-based methods for adjusting for hidden confounders (Novembre and Stephens, 2008; Fan et al., 2020; Bai, 2003). In the confounding model (1), the effect of confounding variables will approximately lie in the span of the first few principal components of X (see Figure 1). One adjusts for a first few principal components from the columns of the design matrix X before further analysis in hope of removing the effect of the confounding variables (Paul et al., 2008; Huang and Jojic, 2011). This procedure is in fact analogous to applying a spectral transformation, where the matrix $\widetilde{D}$ is obtained from D by mapping the first several singular values to 0. See also Figure 3 for an illustration. The slight difficulty with this approach is knowing exactly the number of principal components to remove. Asymptotically, this can be done with high probability (Bai, 2003) under certain assumptions on the separation of the singular values. However, for finite samples or if there is a slight model misspecification, it might not be that easy to estimate q, see e.g. our real data genomic dataset in Figure 10.

3.2.2 Some Intuition

Since our method (5) is invariant under transformation $F \to cF$, for arbitrary constant $c \in \mathbb{R}$, we can assume without loss of generality that the singular values of F are at most 1, i.e. the transformation F shrinks all vectors, with different shrinkage in directions of its singular vectors. Ideally, we would like to shrink in a way such that the perturbation term $\widetilde{X}b$ becomes much smaller compared to the signal $\widetilde{X}\beta$.

Trim transform has the highest shrinkage along directions of the singular vectors corresponding to large singular values. The more b is aligned with the first few singular vectors of X (those corresponding to large singular values), the larger $\|Xb\|_2$ will be. Therefore, shrinking those large singular values ensures that $\|\widetilde{X}b\|$ stays small regardless of the direction b is pointing to. It is especially the case in the confounding model that b approximately lies in the span of the first few singular vectors (see Figure 1).

As can be seen from definition of b, Xb is the part of the confounding effect $H\delta$ which

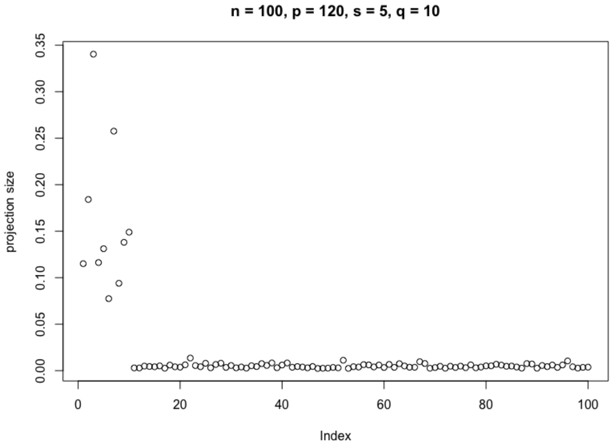


Figure 1: Size of the projection of b onto V_i for different i, for a random dataset drawn from the confounding model with $q = 10$ confounding variables, as described in Section 5.1.1. We see that the projections of b onto the first 10 singular values are substantially larger than the rest.

is correlated with X. Therefore, $\|Xb\|_2$ can be just as large as $\|H\delta\|_2 = \mathcal{O}(\sqrt{n}\|\delta\|_2)$. However, after applying the Trim transform we have that

$$\|\widetilde{X}b\|_2 \leqslant \lambda_{\max}(\widetilde{X})\|b\|_2 = \mathcal{O}\left(\sqrt{p} \times \sqrt{\frac{\|\delta\|_2^2}{p}}\right) = \mathcal{O}(\|\delta\|_2),$$

which is substantially smaller than before. $\lambda_{\max}(\widetilde{X})$ is the largest singular value of $\widetilde{X}$, which will be shown in Lemma 4.2 to be of order $\sqrt{p}$ for the Trim transform and we have $\|b\|_2 = \mathcal{O}\big(\sqrt{\|\delta\|_2^2/p}\big)$ under certain model assumptions by Lemma 4.1.

On the other hand, the signal $X\beta$ lies in the span of a sparse set of predictors. Therefore, the signal $\widetilde{X}\beta$ will be approximately of the same size as the signal $X\beta$ before transformation, unless β is aligned with the large singular vectors, which are shrunk the most. This is very unlikely if they are sufficiently random. This is illustrated in Figure 2. Therefore, by shrinking large singular values, $\|Xb\|_2$ will decrease much more compared to $\|X\beta\|_2$.

4 Theoretical Results

In this section we analyse the behaviour of the ℓ_1-estimation error for the sparse coefficient β for an arbitrary spectral transformation F. We derive results for the perturbed linear model (3) and relate them to the confounding model (1) by using the relationship between them. The proofs of the results can be found in the appendix of Ćevid et al. (2020a).

We show that if our spectral transformation fulfils certain criteria, and the confounding

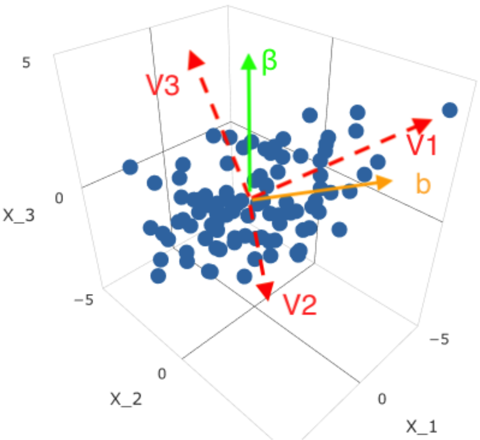


Figure 2: Visualisation of the relationship between the perturbation b, signal β and singular vectors of X. In the confounding model b will be much more aligned with the singular vectors corresponding to large singular values than β.

is dense in the sense that every confounding variable affects many predictors, we achieve in the high-dimensional case the same ℓ_1-error rate as the Lasso in the case when we have no confounding, despite the presence of the coefficient perturbation caused by the confounding variables. Furthermore, in Section 4.4, we discuss specific choices of spectral transformations and verify that the Trim transform (7), as well as Lava and PCA adjustment, can be used in order to achieve this error rate.

We assume first for simplicity that we are in the high-dimensional case, where $p \geqslant n$. However, the theory developed in this section also holds for the case $n > p$ with small adjustments. We discuss the case $n > p$ in more details in Section 4.6.

4.1 Notation

For a matrix M we write

$$\phi_M := \inf_{\|\alpha\|_1 \leqslant 5\|\alpha_S\|_1} \frac{\sqrt{\alpha^T M \alpha}}{\frac{1}{\sqrt{s}}\|\alpha_S\|_1},$$

where S is the support set of β, s is the size of S and α_S is a vector consisting only of the components of α which are in S.

Let us also write $\widetilde{\Sigma} := \frac{1}{n}\widetilde{X}^T\widetilde{X}$, and $\widehat{\Sigma} := \frac{1}{n}X^T X$. We denote the k-th largest diagonal element of the transformed singular values $\widetilde{D}$ by $\widetilde{d}_{(k)}$. We denote the the largest, the smallest and i-th (non-zero) singular value of any rectangular matrix A by $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ and $\lambda_i(A)$ respectively. The condition number is defined as $\mathrm{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$.

Finally, we use the notation $A = \Omega(B)$ if $\frac{B}{A} = \mathcal{O}(1)$, i.e. if A has asymptotically at least the same rate as B and $A \asymp B$ if A and B have asymptotically the same rate. $A = \mathcal{O}_p(B)$

means that there exists a constant $c > 0$ such that $\mathbb{P}(A > cB) \to 0$ and Ω_p is defined analogously.

4.2 Main Result for the Confounding Model

We present here the main result for the confounding model (1), which we derive below by considering the relationship with the corresponding perturbed linear model, as described in Section 2.

Theorem 1. *Consider the model in (1) with* $\max_i \Sigma_{ii} = \mathcal{O}(1)$ *and cond*$(\Sigma_E) = \mathcal{O}(1)$ *and suppose that* $\lambda_{\min}(\Sigma)$ *is bounded away from zero. Assume that the model satisfies*

(A1) $\lambda_{\min}(\Gamma) = \lambda_{\min}(Cov(X, H)) = \Omega(\sqrt{p})$.

Assume additionally that a spectral transformation F *in (5) with* $\lambda_{\max}(F) = 1$ *satisfies*

(A2) $\lambda_{\max}(\widetilde{X}) = \mathcal{O}_p(\sqrt{p})$

(A3) $\phi^2_{\widetilde{\Sigma}} = \Omega_p(\lambda_{\min}(\Sigma))$.

Then for the penalty level $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$*, despite the confounding variables, the* ℓ_1*-estimation error has the following rate:*

$$\|\widehat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right).$$

The assumption **(A1)** means that the confounding is dense in the sense that each confounding variable is correlated with many predictors: The condition $\lambda_{\min}(\Gamma) = \Omega_p(\sqrt{p})$ is satisfied, for example, if $\frac{q}{p} \to 0$ and Γ is drawn at random with either rows or columns of Γ being independent, identically distributed sub-Gaussian random vectors, as shown in Lemma 4.1.

We also show in Section 4.4 that certain choices of the spectral transformation, such as the Trim transform (7) with $\tau = d_{\lfloor tn \rfloor}$, where $t \in (0, 1)$ is an arbitrary constant, or the PCA adjustment, which maps first several singular values to zero, satisfy with high probability the conditions **(A2)** and **(A3)** in the high-dimensional setting under certain conditions.

4.3 ℓ_1-estimation Error of β in the Perturbed Linear Model

In this section we derive an upper bound for the ℓ_1-estimation error of β in the perturbed linear model and show that we can achieve the usual Lasso error rate in the high-dimensional case, provided the perturbation b is sufficiently small. Then the main theorem for the confounding model, Theorem 1, follows from Corollary 4.1 by using the relationship between the models described in Section 2.

The following result describes the effect of an arbitrary linear transformation F on the ℓ_1-estimation error of the Lasso:

Theorem 2. *Assume the model in (3) with* $\max_i \Sigma_{ii} = \mathcal{O}(1)$*. Let* $F \in \mathbb{R}^{n\times n}$ *be an arbitrary linear transformation and* $A > 0$ *an arbitrary fixed constant. Then for the method described in* (5) *with transformation* F *and penalty level* $\lambda = A\sigma\sqrt{\frac{\log p}{n}}\lambda_{\max}(F)^2$*, with probability at least* $1 - 2p^{1-A^2/(32\max_i \Sigma_{ii})} - pe^{-n/136}$*, we have*

$$\|\widehat{\beta} - \beta\|_1 \leqslant C_1 \frac{s\lambda}{\phi^2_{\widetilde{\Sigma}}} + C_2 \frac{\|\widetilde{X}b\|_2^2}{n\lambda},$$

where C_1, C_2 *are constants depending only on* A*.*

Remark. *One can get a better bound*

$$\|\widehat{\beta} - \beta\|_1 \leqslant C_1 \frac{s\lambda}{\phi^2_{\widetilde{\Sigma}}} + C_2 \frac{\sqrt{s}}{\phi_{\widetilde{\Sigma}}} \frac{\|\widetilde{X}b\|_2}{\sqrt{n}}$$

by taking larger penalty λ *than the one above, but then* λ *depends on the unknown quantity* $\|\widetilde{X}b\|_2$*. For that reason we will use the bound above with standard penalty level* λ*, since it does not matter when* $\|\widetilde{X}b\|_2$ *is small, which holds in our case, as shown later.*

The first term is the standard bound for the ℓ_1-error of the Lasso, with only difference that the compatibility constant is for the matrix $\widetilde{\Sigma} = \frac{\widetilde{X}^T\widetilde{X}}{n}$ rather than the matrix $\widehat{\Sigma} = \frac{X^TX}{n}$. The second term shows the dependence of the error on the term $\widetilde{X}b$. It is also worth noting that the penalty λ has standard form up to the scaling correction factor $\lambda_{\max}(F)^2$, which equals 1 for the Trim transform and the PCA adjustment.

In order to control the error caused by the coefficient perturbation b, we need to make $\|\widetilde{X}b\|_2$ small by shrinking the singular values enough, e.g. by ensuring that $\widetilde{d}_{(1)}$, the largest singular value after transformation, is small. On the other hand, we must not shrink the singular values too much, since we need $\phi_{\widetilde{\Sigma}}$ to stay large. If we have that $\phi^2_{\widetilde{\Sigma}}$ is bounded away from 0 with high probability, as it is the case with $\phi^2_{\widehat{\Sigma}}$ (see Bühlmann and van de Geer (2011)), and that $\|\widetilde{X}b\|_2$ is sufficiently small, we get from Theorem 2 that our estimator achieves the usual Lasso error rate:

Corollary 4.1. Consider the model in (3) with $\max_i \Sigma_{ii} = \mathcal{O}(1)$ and suppose that $\lambda_{\min}(\Sigma)$ is bounded away from zero. For the coefficient perturbation b as in (4), assume that

(A1') $\|b\|_2^2 = \mathcal{O}\left(\frac{s\sigma^2 \log p}{p}\right)$.

Assume additionally that the spectral transformation F in (5) with $\lambda_{\max}(F) = 1$ satisfies

(A2) $\lambda_{\max}(\widetilde{X}) = \mathcal{O}_p(\sqrt{p})$

(A3) $\phi^2_{\widetilde{\Sigma}} = \Omega_p(\lambda_{\min}(\Sigma))$.

Then for the penalty level $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$, despite the coefficient perturbation, the ℓ_1-estimation error has the following rate:

$$\|\widehat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right).$$

We show in the following section that in the perturbed linear model that arises from the confounding model (1), the induced coefficient perturbation b, given in (4), satisfies the condition **(A1')**, provided that the dense confounding assumption **(A1)** is satisfied. We also show that certain spectral transformations, such as the Trim transform (7) with $\tau = d_{\lfloor tn \rfloor}$, where $t \in (0,1)$ is an arbitrary constant, or the PCA adjustment satisfy the conditions **(A2)** and **(A3)** under certain conditions.

Remark (Fixed design)**.** *The results of Theorem 2 and Corollary 4.1 can be easily extended to the perturbed linear model with fixed design. One can even relax the assumption* ***(A1')*** *to a weaker condition*

$$\|V^T b\|_2^2 = \mathcal{O}\left(\frac{s\sigma^2 \log p}{p}\right).$$

It is worth noting that if the perturbation vector b has uniformly random direction, which is not the case with the confounding model (1), this becomes much weaker than the condition ***(A1')*** *above and we only require $\|b\|_2^2 = \mathcal{O}\left(\frac{s\sigma^2 \log p}{n}\right)$.*

4.4 Validity of the Assumptions

In this section we will justify the assumptions in Theorem 1 and Corollary 4.1 for certain spectral transformations F, with an emphasis on the Trim transform (7) and the PCA adjustment. We also discuss later the performance of other choices of spectral transformations.

Assumptions (A1) and (A1')

The assumption **(A1')** for the perturbed linear model says that the coefficient perturbation must not be too large. It can also be viewed as the condition which makes the perturbed linear model identifiable, since in general it is impossible to distinguish the true coefficient vector β from the perturbed coefficient vector $\beta + b$, unless b has some additional structure. The rate $\mathcal{O}(\sqrt{s\sigma^2 \log p/p})$ may seem too strict, but this is the rate with respect to the ℓ_2-norm, so if the perturbation vector is dense, this becomes approximately $\|b\|_1 = \mathcal{O}(\sqrt{s\sigma^2 \log p})$.

The following lemma shows that if the confounding is dense in the confounding model (the assumption **(A1)** holds), then the induced coefficient perturbation in the underlying perturbed linear model is small (the assumption **(A1')** holds). It is important to note that certain dense confounding assumption is necessary. The term Xb can be thought of as the part of the confounding $H\delta$ that can be explained by X and if, as an extreme example, the confounder H_i is correlated with only the predictor X_j, only the j-th component of X will be useful for describing the effect of H_i on Y and thus b_j will be very large and we will not be able to estimate β_j.

Lemma 4.1. Assume that the confounding model (1) satisfies $\lambda_{\min}(\Gamma) = \lambda_{\min}(\text{Cov}(H, X)) = \Omega\left(\sqrt{p}\right)$ and $\text{cond}(\Sigma_E) = \mathcal{O}(1)$. Then we have:

$$\|b\|_2^2 = \|\text{Cov}(X)^{-1}\text{Cov}(X, H)\delta\|_2^2 \leqslant \text{cond}(\Sigma_E) \cdot \frac{\|\delta\|_2^2}{\lambda_{\min}(\Gamma)^2} = \mathcal{O}\left(\frac{\|\delta\|_2^2}{p}\right) = \mathcal{O}\left(\frac{\sigma^2}{p}\right)$$

The condition $\lambda_{\min}(\Gamma) = \Omega_p(\sqrt{p})$ is satisfied, for example, if $\frac{q}{p} \to 0$ and Γ is drawn at random with either its rows or columns being independent, identically distributed sub-Gaussian random variables with expectation 0 and covariance matrix Σ_Γ, with $\lambda_{\min}(\Sigma_\Gamma)$ bounded away from zero.

From this we see that it is important that the effect of the latent variables is spread out over many predictors. If this is not true, $\lambda_{\min}(\Gamma)$ will be too small and thus $\|b\|_2$ will be too large.

Assumption (A2)

We investigate quickly the behaviour of singular values of X in order to see whether the assumption **(A2)** holds for the transformed matrix $\widetilde{X}$. This assumption says that after the transformation, the largest singular value is not too large.

In the confounding model we have $\Sigma = \Gamma^T\Gamma + \Sigma_E$, i.e. the covariance matrix of X has additional low-rank component $\Gamma^T\Gamma$, which causes the top several singular values of Σ to be very large. Since the rows of X are drawn from a distribution with covariance matrix Σ, the first few singular values of X will be large as well (Donoho et al., 2013). However, the following lemma shows that the bulk of the singular values will never be too large, i.e. they will be of order $\sqrt{p}$. The assumption **(A2)** requires the transformed singular values to be of this order.

Lemma 4.2. Assume that $X \in \mathbb{R}^{n\times p}$ is a random matrix whose rows are i.i.d. sub-Gaussian vectors with covariance matrix Σ. Let $d_1, \ldots, d_r \geqslant 0$ be its singular values. Assume also that $\text{Tr}(\Sigma) \asymp p$ and that $\sqrt{\log p/n} \to 0$. We have:

$$\frac{1}{n}\sum_{i=1}^{r} d_i^2 = \text{Tr}(\Sigma)(1 + o_p(1)).$$

Furthermore, when $p > n$, $d_{\lfloor tn \rfloor} = \mathcal{O}_p(\sqrt{p})$ for any $t \in (0, 1)$.

For the Trim transform the largest singular value after transformation $\widetilde{d}_{(1)}$ equals the trimming threshold τ and the above lemma shows that $\tau = d_{\lfloor tn \rfloor}$ for $t \in (0, 1)$, e.g. the median singular value when $t = 0.5$, is a good choice and the assumption **(A2)** holds.

If we further assume that Σ_E has bounded singular values, thus ensuring the gap between the q-th and $(q + 1)$-st eigenvalues of Σ, we get that all but the first q singular values of X will not be too large, thus justifying the assumption **(A2)** for the PCA adjustment, since there we have $\lambda_{\max}(\widetilde{X}) = \widetilde{d}_{(1)} = \widetilde{d}_{q+1} = d_{q+1}$.

Lemma 4.3. Assume that $p > n$ and that X has i.i.d. sub-Gaussian rows with covariance matrix $\Sigma = \Gamma^T \Gamma + \Sigma_E$, where $\Gamma \in \mathbb{R}^{q \times p}$ and $\lambda_{\max}(\Sigma_E) = \mathcal{O}(1)$, then we have $d_{q+1} = \mathcal{O}_p(\sqrt{p})$.

This lemma also shows that in this case the trimming threshold τ for the Trim transform can be chosen to be $\tau = d_{q+1}$, but $\tau = d_{\lfloor tn \rfloor}$ might be a better choice as the number of confounders q is unknown.

Assumption (A3)

This assumption says that the compatibility constant $\phi_{\widetilde{\Sigma}}$ does not substantially decrease after applying our transformation F. We want to show that by shrinking the singular values we have not shrunk our signal $X\beta$ too much. Intuitively, this means that the active set X_S is not too aligned with the directions along which we substantially shrink, which corresponds to the first several singular vectors in the case of Trim transform and PCA adjustment.

It is difficult to bound $\phi_{\widetilde{\Sigma}}$ for an arbitrary spectral transformation F, since the distribution of the singular vectors V of the design matrix X is complicated. However, one can directly exploit the results from the factor analysis literature (Bai, 2003) for the PCA adjustment, from which it follows that in a certain asymptotic regime the transformed design matrix $\widetilde{X}$ is close to the unconfounded design matrix E. Using this result, one can directly obtain the compatibility condition **(A3)** for the PCA adjustment by using the standard argument (Bühlmann and van de Geer, 2011).

Lemma 4.4. Let X be generated from the confounding model (1) and let F be a spectral transformation shrinking the first q singular values of X to 0. If q is fixed, $\frac{1}{p}\sum_{i,j=1}^{p} |(\Sigma_E)_{ij}|$ upper bounded and $\frac{s \log(pn)}{\min(n,p)} \to 0$, we have that, with probability converging to 1, the compatibility condition holds for the transformed design matrix $\widetilde{X} = FX$:

$$\phi^2_{\frac{1}{n}\widetilde{X}^T\widetilde{X}} \xrightarrow{p} \phi^2_{\frac{1}{n}E^TE} = \Omega_p\left(\lambda_{\min}(\Sigma_E)\right).$$

In the Appendix A.1 of Ćevid et al. (2020a) the analysis of the compatibility constant $\phi_{\tilde{\Sigma}}$ is also provided for arbitrary spectral transformation under the somewhat restrictive assumption that the singular vectors V have uniformly distributed direction.

Since the ratio of the transformed singular values for the Trim transform and PCA adjustment is bounded from below by $\frac{\tau}{d_{q+1}}$, the compatibility constant ϕ_{Trim} for the Trim transform can be bounded from below by the compatibility constant ϕ_{PCA} for the PCA adjustment:

$$\phi_{\text{Trim}} \geqslant \frac{\tau}{d_{q+1}} \phi_{\text{PCA}} = \frac{d_{\lfloor tn \rfloor}}{d_{q+1}} \phi_{\text{PCA}}$$

and thus the compatibility condition holds for the Trim transform as well if d_{q+1} and $\tau = d_{\lfloor tn \rfloor}$ are of comparable sizes, i.e. $\frac{d_{\lfloor tn \rfloor}}{d_{q+1}} = \Omega_p(1)$. By Lemma 4.3, we have $d_{q+1} = \mathcal{O}_p(\sqrt{p})$ and by the following lemma it holds that for quite a wide range of settings we also have that $d^2_{\lfloor tn \rfloor} = \Omega_p(\lambda_{\min}(\Sigma)p)$. Therefore, Lemma 4.4 can be used for showing the compatibility condition for the Trim transform as well.

Lemma 4.5. Assume that X is a random design matrix with i.i.d. rows with covariance matrix Σ and suppose $p > n$. Assume that any of the following conditions is satisfied:

i) the rows of X have a sub-Gaussian distribution and $\frac{p}{n} \to \infty$

ii) the rows of X have a $N(0, \Sigma)$ distribution and $\liminf \frac{p}{n} > 1$

iii) the rows of X have $N(0, \Sigma)$ distribution and $\limsup \frac{k}{n} < 1$

Then we have

$$d_k^2 = \Omega_p(\lambda_{\min}(\Sigma)p).$$

4.5 Performance of Various Spectral Transformations

The result of Theorem 1 can be applied to any spectral transformation that satisfies the assumptions (**A2**) and (**A3**). We discuss here which spectral transformations satisfy them and what are their possible advantages and disadvantages for the performance of the corresponding estimator $\widehat{\beta}$. The illustration of the spectral transformations discussed below is given in Figure 3.

PCA adjustment As shown above, under certain assumptions we get that the spectral transformation which maps first q singular values to 0 will satisfy assumptions **(A2)** and **(A3)**. Even though it might seem that one disadvantage of this method is that the number of confounding variables q needs to be estimated from the data, one can show that asymptotically it can be done accurately with high probability (Bai, 2003). PCA adjustment leaves most of the singular values intact, so the increase in the estimator variance will not be large.

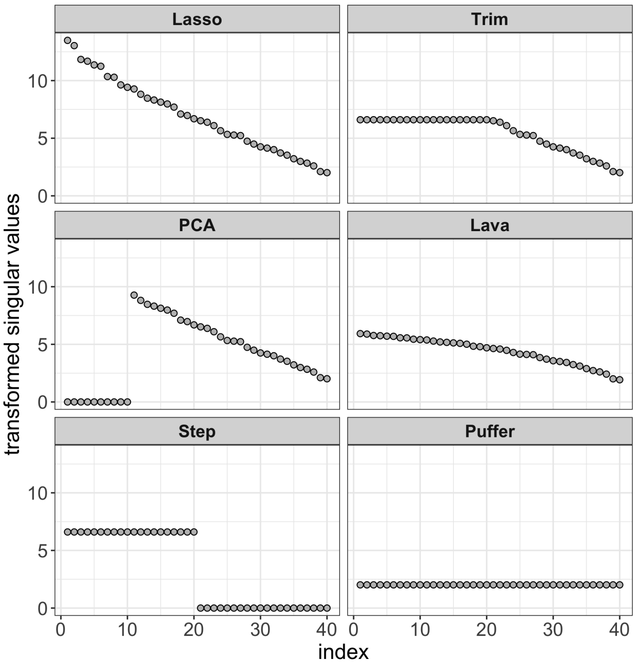


Figure 3: Singular values of $\widetilde{X}$ after applying spectral transformations corresponding to different methods to 40×60 matrix X with i.i.d. standard normal entries.

Lasso The simplest option is to take $\widetilde{d}_i = d_i$, i.e. the usual Lasso algorithm without any transformation. Standard Lasso theory shows that the assumption **(A3)** is satisfied (see Bühlmann and van de Geer (2011)). However, **(A2)** requires that the largest singular value of X is of order $\mathcal{O}(\sqrt{p})$, which typically does not hold in presence of confounding variables.

Trim transform As shown above, we have that the Trim transform satisfies assumptions **(A2)** and **(A3)** if we take the trimming threshold to be $\tau = d_{\lfloor tn \rfloor}$ for some $t \in (0, 1)$, e.g. the median singular value. Compared to the PCA adjustment, it has an advantage that one does not need to estimate the number of confounding variables from the data. Moreover, it does not shrink first several singular values to 0, but only to the necessary level. This more gradual shrinkage might lead to better performance especially if the signal $X\beta$ is more aligned with the first few singular vectors.

Lava The mapping $d_i \to \sqrt{n\lambda_2} d_i / \sqrt{n\lambda_2 + d_i^2}$ used in the Lava algorithm (Chernozhukov et al., 2017) satisfies the conditions **(A2)** and **(A3)** as well, since the transformed singular values $\widetilde{d}_i$ are quite close to the ones for the Trim transform $\widetilde{d}_i = \min(d_i, \tau)$, for an appropriate choice of τ:

$$\frac{1}{2} \min(d_i, \sqrt{n\lambda_2}) \leqslant \frac{\sqrt{n\lambda_2} d_i}{\sqrt{n\lambda_2 + d_i^2}} \leqslant \min(d_i, \sqrt{n\lambda_2}).$$

This also reveals how to choose the penalty λ_2 in Lava: $\lambda_2 = \frac{1}{n} d^2_{[\min(n,p)/2]}$ and λ_1 can be chosen by cross-validation. This transformation has the property that it is smoother than the Trim transform. We note that with this comment and Corollary 4.1, we have established the standard Lasso ℓ_1-error rate for Lava for estimating the sparse parameter β in a high-dimensional regression model; such result is not given in Chernozhukov et al. (2017).

Puffer transformation For the Puffer transform (Jia et al., 2015), where we map all singular values to a constant d_n (because of homogeneity it does not matter to which constant we map it, but we have assumed w.l.o.g. that $\widetilde{d}_i \leqslant d_i$, so we need to map them to d_n), the assumption **(A2)** is easily satisfied. However, for **(A3)** we need to have $d_n^2 = \Omega_p\left(\lambda_{\min}(\Sigma)\, p\right)$. From Vershynin (2012), we have that this holds only if $\liminf \frac{p}{n} > 1$, i.e. the Puffer transform will not work well if n and p are close.

Step function The justification of the assumptions **(A2)** and **(A3)** for Trim transform apply as well for the step function $\widetilde{d}_i = \tau \mathbb{1}(d_i > \tau)$ with the same threshold τ. However, unnecessarily shrinking singular values might cause worse performance than for the Trim transform.

4.6 Low-dimensional Case: $n > p$

The statement of Theorem 2 still holds in the low-dimensional case $n > p$. However, $\frac{1}{n}\|\widetilde{X}b\|_2^2$ will now be of larger order than λ. We have that $\lambda_{\max}(\widetilde{X}) = \mathcal{O}_p(\sqrt{n})$, compared to $\sqrt{p}$ before (see Lemma 4.2), which under the assumption **(A1')** gives us that $\frac{1}{n}\|\widetilde{X}b\|_2^2 = \mathcal{O}(\|b\|_2^2) = \mathcal{O}(\frac{s\sigma^2 \log p}{p})$. Therefore, the second term in the bound of Theorem 2 will be too large in comparison with the first term.

Fortunately, from the remark below Theorem 2, we see that by taking larger λ, we can decrease the rate of the second term. If the perturbation term $\frac{1}{n}\|\widetilde{X}b\|_2$ gets larger than the standard penalty rate, as it is the case when $n > p$, it is better to penalize more. One gets in this case:

$$\|\widehat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{s\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}} + \frac{\sqrt{s}\|b\|_2}{\sqrt{\lambda_{\min}(\Sigma)}}\right)$$

which by Lemma 4.1 in the confounding model, under the dense confounding assumption **(A1)**, becomes:

$$\|\widehat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{s\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}} + \frac{\sqrt{s}\sigma}{\sqrt{\lambda_{\min}(\Sigma)}\sqrt{p}}\right).$$

One can not expect the same error rate as in the high-dimensional setting, since this would imply that, for fixed p, the error converges to 0 as $n \to \infty$ which can not happen because the error is not only due to the randomness of the sample data, but also due to the coefficient perturbation b. The perturbation b only depends on how the confounding variables affect the predictors and not on the number of data points and thus one can not expect consistency for a fixed p. However, we see that the estimator is consistent when $n, p \to \infty$. The more predictors we have, the more is the effect of the confounding variables spread out.

This is also illustrated in Figure 4, where we can see that even though the error decreases as we increase the number of data points, it still seems to have a nonzero limit. However, the error is small, especially in comparison with the standard Lasso, and there is a benefit in using our method.

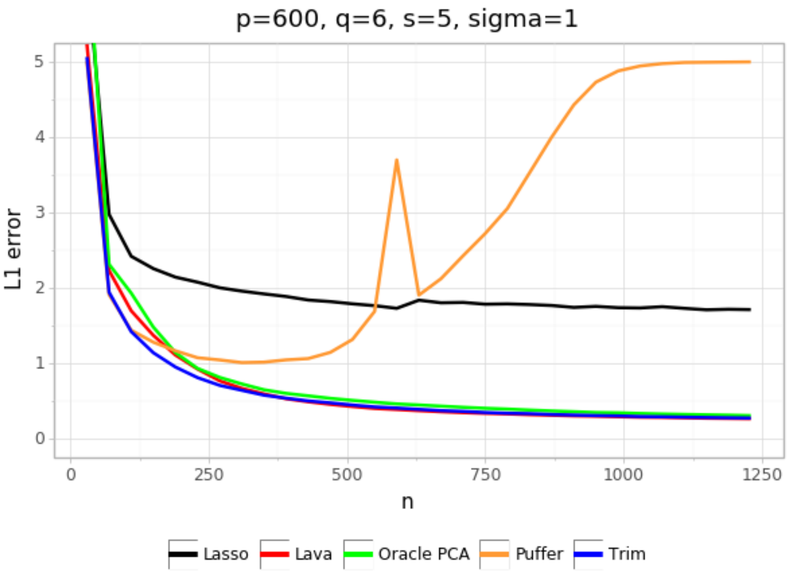


Figure 4: Dependence of the estimation error $\|\widehat{\beta} - \beta\|_1$ on the sample size n for different spectral transformations and data generated from the confounding model, including the case $p < n$, as described in Section 5.1.1.

5 Empirical Results

We present here some empirical results for simulated and real data.

5.1 Simulations

We demonstrate the performance of various spectral transformations for estimating the coefficient vector β with a subsequent use of the Lasso: Trim transform, Lava, Puffer and

PCA adjustment. We investigate the cases when the perturbation b arises from hidden confounding and when it is randomly sampled.

5.1.1 Setting

We generate the data from the confounding model (1). We take $\Sigma_E = \sigma_E^2 I_p$, where $\sigma_E = 2$ and $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$, so $s = 5$. For a fixed number q of hidden confounders, we sample the coefficients Γ_{ij} and δ_i independently as standard normal random variables. By default, we take $q = 6$. Unless stated otherwise, we use the noise level $\sigma = 1$ as the standard deviation of ϵ. Finally, the sample size is set to be $n = 200$ and the dimensionality of the predictors is $p = 600$ as the default value. All results are based on $N = 2^{12} = 4096$ independent simulations.

It is also interesting to consider the perturbed linear model (3). We do not generate data from this model directly, but we will modify the underlying perturbation term b which is implicit in the confounding model by formula (4). This way we can compare the results obtained for the confounding model and the perturbed linear model directly with each other. We replace b by Qb where Q is a random rotation matrix so that the new perturbation has the same size, but with uniformly random direction. We note that the resulting distribution is the same as of the perturbed linear model (3), where rows of X are drawn from $N(0, \Sigma)$, where $\Sigma = \Gamma^T\Gamma + I_p$, and b is drawn uniformly from a ball of radius $\|(\Gamma^T\Gamma + I_p)^{-1}\Gamma^T\delta\|_2$.

5.1.2 Choosing λ

In practice we encounter the problem of choosing the penalty level λ for the Lasso after applying a spectral transformation. The results of Theorem 1 and Corollary 4.1 give us that one can use the standard theoretical penalty rate $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$ to get the desired error rate of our estimator. In practice one often resorts to using cross-validation (CV) for choosing the penalty parameter rather than using the theoretical value, especially since σ is unknown.

However, one needs to be careful in presence of confounding variables; in this case the coefficient vector $\beta + b$ describes the data better than β, which we are trying to recover. Therefore, cross-validation tends to choose a smaller value of λ than the optimal for recovering β. This is illustrated in the Figure 5, where we see that, for example, the Puffer transform is significantly affected by this choice of λ. For recovering β in practice, it might be better to increase slightly the value of λ chosen by cross-validation (Janzing and Schölkopf, 2018). But on the other hand, smaller λ gives us a larger set of variables, which might be beneficial for variable screening.

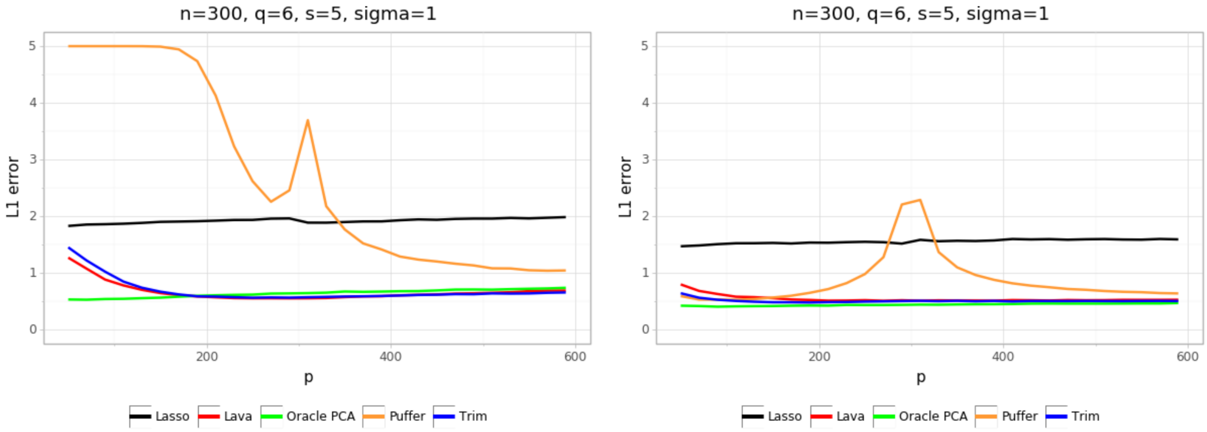


Figure 5: Dependence of the estimation error $\|\widehat{\beta} - \beta\|_1$ on the number of predictors p for different spectral transformations and data generated from the confounding model (1), as described in Section 5.1.1. In the left plot, the penalty is chosen by cross-validation, whereas in the right plot we use the oracle value for which the estimation error is minimal.

In all simulations, unless stated otherwise, the penalty level is chosen by cross-validation. This choice does not seem to worsen the performance of the Trim transform or Lava a lot, as one can see in Figure 5 and Figure 9, and it is of great practical importance since the oracle value of λ, i.e. the one for which $\|\widehat{\beta}_\lambda - \beta\|_1$ is smallest, can not be directly determined from the data.

5.1.3 Results

Here we present the results of the simulations for both the confounding model and the perturbed linear model. A fundamental difference between them is that the coefficient perturbation arising from the confounding model is pointing towards the singular vectors of X corresponding to the large singular values (see Figure 1). This makes $\|Xb\|_2$ larger for a fixed $\|b\|_2$, and in this case the estimation error will be larger. On the other hand, in this case we can improve our accuracy more compared to the plain Lasso by shrinking large singular values, as will be shown below.

Noise versus perturbation In the left plot in Figure 6 we can see how the estimation error changes depending on the size of the noise σ in the confounding model. When σ is small, the perturbation b has the biggest effect on the error. On the other hand, if σ is large, then the influence of the perturbation b becomes less pronounced.

We can see that the standard Lasso is affected a lot by the coefficient perturbation, whereas the Puffer transform and the PCA adjustment are affected more by the additive noise than the Lava and the Trim transform, since the slopes of the corresponding curves

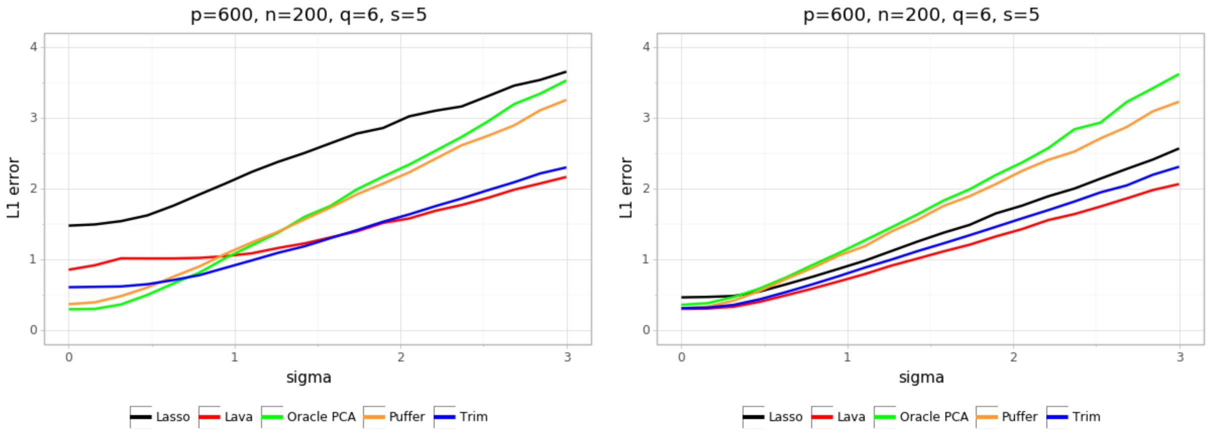


Figure 6: Dependence of the estimation error $\|\widehat{\beta} - \beta\|_1$ on the size of the noise for different spectral transformations for confounding model (left) and the perturbed linear model (right), as described in Section 5.1.1.

are steeper. The higher variance of the Puffer transform is most evident in Figure 4 and Figure 5; when n, p are close to each other, some of the singular values of X become quite small and thus mapping them to a constant can inflate the error ϵ in the corresponding directions by a lot. We can observe that the oracle PCA adjustment, which removes exactly the q largest singular values of X, works well, especially when σ is small. For larger σ, we see that Trim transform and Lava work slightly better since they do not remove that much of the signal.

In the right plot of Figure 6, we have randomized the direction of b while keeping everything else constant, as described in Section 5.1.1. This then corresponds to a model with random perturbation b, but no specific further structure in terms of confounding. We can see a substantial improvement of the standard Lasso: in hindsight this shows that the Lasso is very sensitive to confounding variables but much less so to perturbation of sparsity. Also, it is worth noting that the PCA adjustment method is now consistently worse than the Trim transform or Lava, since the projection of b onto the span of the first q singular vectors is not that large anymore.

We can see more clearly the bias-variance tradeoff for different spectral transformations in Figure 7, where we have taken the rotated coefficient perturbation b, as in the right plot of Figure 6 and then artificially scaled it by a chosen constant. For a very small b, we see that Puffer and PCA adjustment have somewhat worse performance. As b increases, Trim transform and Lava reduce the bias caused by b much better than the Lasso. We can also see that the PCA adjustment does not reduce the bias as much, but its performance would be significantly better if b was not rotated, but aligned with the top several principal components as in the confounding model, see Figure 6.

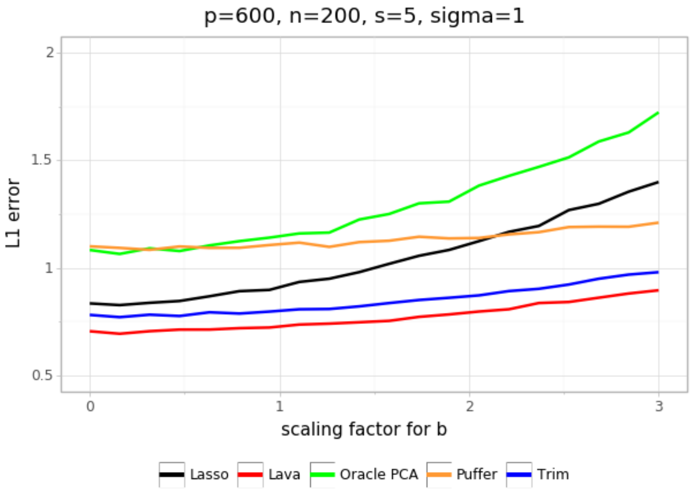


Figure 7: Dependence of the estimation error on the size of the perturbation vector b for different spectral transformation for the perturbed linear model, as described in Section 5.1.1.

Number of confounding variables In Figure 8 we can see how the estimation error depends on the number q of confounding variables. As above, we see that the Lasso is severely affected by the presence of confounding variables. The Puffer transform performs reasonably well since n and p are different enough and the Trim transform and Lava exhibit similar and good performance in all cases.

PCA adjustment works well for the confounding model if we correctly guess the number of confounding variables. In the left plot in Figure 8 we can clearly see how the estimation error is affected by the misspecification of the number of the principal components we remove. The oracle PCA method, which removes exactly q principal components, performs reasonably well, particularly for smaller values of q. However, if we overestimate or especially if we underestimate the number of confounding variables, the estimation error will become significantly worse compared to the Trim transform or Lava.

Method robustness We are interested in whether there are any disadvantages in using the spectral transformations if we wrongly think that there is some hidden confounding or that the sparse coefficient has been perturbed.

In Figure 9 we display the estimation error for the confounding model as in Figure 8, but where the coefficient bias b has been set to 0, i.e. this is a standard sparse linear model with X being generated from the spiked covariance model.

There is no indication for relevant differences between the performances of the Trim transform, Lava and the Lasso. The Lasso performs slightly better for larger values of q and slightly worse for smaller q. It is worth noting that on this plot the estimation error starts to decrease as q increases, which is due to a scaling issue. This happens because

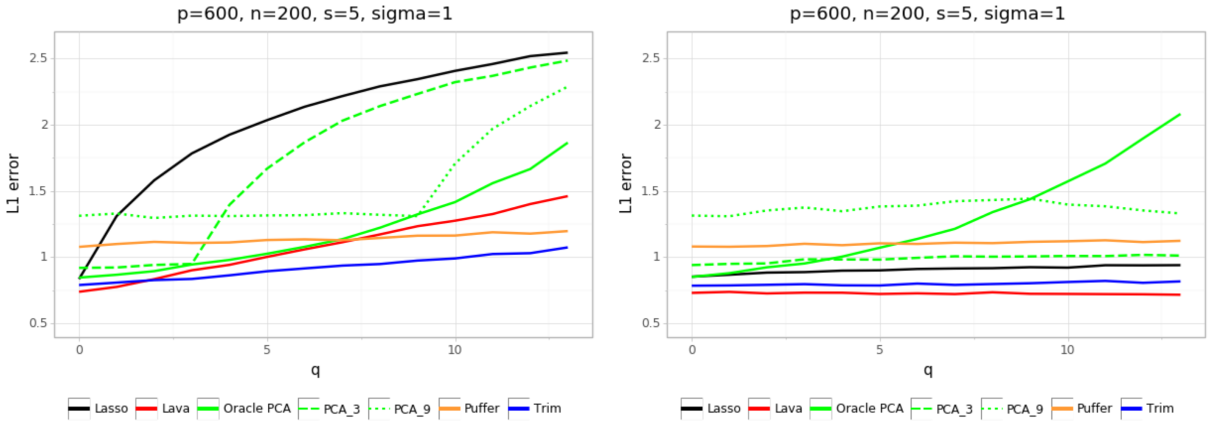


Figure 8: Dependence of the estimation error $\|\widehat{\beta} - \beta\|_1$ on the number of confounding variables for different spectral transformation for confounding model (left) and the perturbed linear model (right) as described in Section 5.1.1.

the variance of X increases as q increases, since $\Sigma = \Gamma^T\Gamma + \Sigma_E$, thus effectively increasing the signal to noise ratio. PCA adjustment seems to be affected most by the choice of λ, especially for larger q since its shrinkage is larger in this case, see Figure 9. With the oracle choice of the penalty level, its performance is very similar to the performance of the Lasso.

Our empirical results support theoretical evidence, which showed that it is safe to use wisely chosen spectral transformations such as the Trim transform or the Lava. If there are any confounding variables present, there is a large improvement over the standard Lasso. On the other hand, if there are no confounding variables, the Trim transform or Lava will have about the same performance as the Lasso. Therefore, our method can be thought of as an easy to use modification of the Lasso which is robust to hidden confounding.

5.2 Application to a Genomic Dataset

In this section we demonstrate the robustness of our method against hidden confounders on a real genomic dataset where we have certain knowledge about the confounding variables. We inspect various spectral transformations in combination with the Lasso and evaluate the differences between the estimates for the original data set and the one where the confounding variables have been adjusted for.

5.2.1 Gene Expression Dataset

We have obtained data from the GTEx Portal (`http://gtexportal.org`). The GTEx project provides large-scale data with an aim to help the scientific community to study gene expression, gene regulation and their relationship to genetic variation. It provides

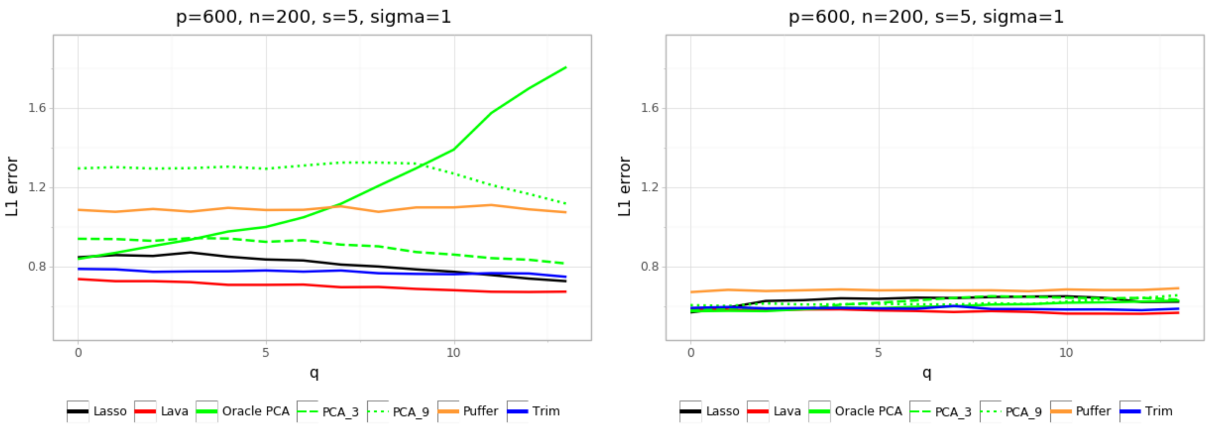


Figure 9: Size of the estimation error $\|\widehat{\beta} - \beta\|_1$ for a sparse linear model where $\Sigma = \Gamma^T\Gamma + I_p$, i.e. the confounding model with the induced perturbation b set to $b = 0$. The penalty level λ is either chosen by cross-validation (left) or taken to be the oracle value, which minimizes the ℓ_1-error (right).

gene expression data from 11,688 samples collected postmortem from 53 different tissues of 714 human donors.

Gene expression is a process in the cell in which the information stored in a certain gene is used for the synthesis of gene products such as proteins. In the GTEx Project it was quantified by the amount of the mRNA in the cell which was created from this gene. Gene expression differs among different people and among different cells within the human body. The type of the cells is determined by the gene expression within them; even though the DNA in all cell nuclei is the same, cells in different tissues behave and look differently and perform significantly different tasks. Gene expression is also affected by the genetic variation and determining the expression quantitative trait loci (eQTL), which are parts of genome which explain the variation in the gene expression, is a very important problem which will help to understand the relationship between genetic variation and different phenotypes.

5.2.2 Setting

We use the fully processed, filtered and normalized gene expression matrix for the skeletal muscle tissue. We consider the gene expression of $p = 14'713$ protein-coding genes measured from $n = 491$ samples. For our purpose, an important aspect of this dataset is that there are also $q = 65$ different covariates provided, which are proxies for the hidden confounding variables. They include genotyping principal components and PEER factors. We can thus obtain the deconfounded data by regressing out these given covariates.

The left panel of Figure 10 displays the singular values of the initial data matrix. We see that the first several singular values are substantially larger than the rest which suggests a possible existence of hidden confounders. In the right part of Figure 10 we can see the singular values of the deconfounded data matrix where we have regressed out all of the $q = 65$ covariates which are provided as confounding proxies.

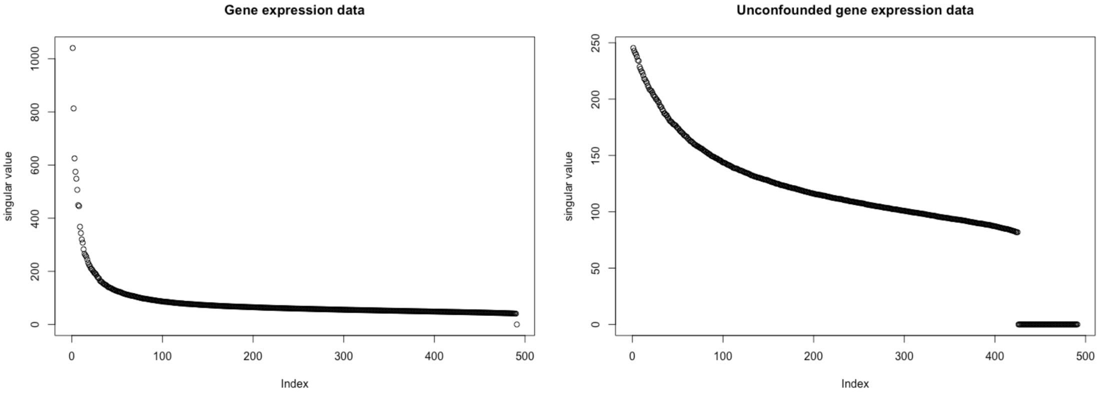


Figure 10: Singular values of the gene expression data matrix for skeletal muscle tissue before (left) and after (right) regressing out the provided $q = 65$ confounding covariates.

We are going to explore now the robustness of the Lasso, Trim transform, and Lava against hidden confounders by comparing the estimates based on the original and the deconfounded data. For a fixed value of k, we regress out first k given confounder proxies from the original gene expression data matrix X in order to get the matrix $X^{(k)}$ and we randomly choose one column to represent the response Y. We are thus trying to explain the expression of one gene by the expressions of other genes.

For every $s = 1, \dots, 20$, we apply the given method on X and $X^{(k)}$ with the regularization λ chosen as the largest value such that the support size of $\widehat{\beta}$ equals a prespecified value s. This leads to estimates $\widehat{\beta}_s$ and $\widehat{\beta}_s^{(k)}$. We measure the dissimilarity of the corresponding supports by $J(\operatorname{supp} \widehat{\beta}_s, \operatorname{supp} \widehat{\beta}_s^{(k)})$, where J is the Jaccard distance:

$$J(A, B) = \frac{A \triangle B}{A \cup B}.$$

5.2.3 Results

In the top left image in Figure 11, we can see the difference of the estimates for the original and the deconfounded data, where 5 randomly chosen confounding variables have been removed and the response Y is the expression of a randomly chosen gene. We can see that the Jaccard distance for the Lasso is closer to 1, indicating that the estimated support sets are very different and almost disjoint; The Trim transform and Lava are much more robust to the hidden confounders and we see that the Jaccard distance between the estimates based on confounded and deconfounded data is much smaller.

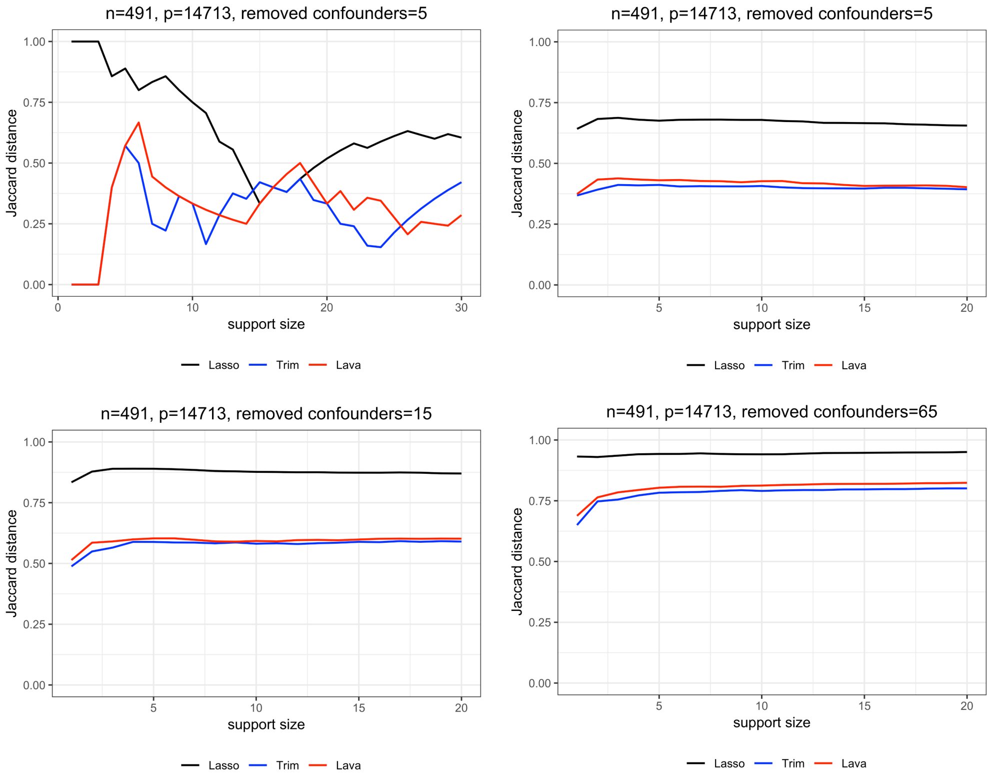


Figure 11: Jaccard distance of the supports of the estimates based on the original and deconfounded data for one randomly chosen response (top left). Jaccard distance, averaged over 500 randomly chosen responses, of the supports of estimates based on the original data and data with 5 (top right), 15 (bottom left) and 65 (bottom right) confounder proxies removed.

In order to make sure that the choice of response Y did not affect the results, we have repeated this experiment for 500 randomly chosen genes and averaged the obtained results. The results are also displayed in Figure 11. We can see that, as we increase the number k of confounding variables which we regress out, the Jaccard distance for all methods is increasing. This is to be expected since $X^{(k)}$ and X are becoming more different as we increase k. However, we can infer that the Trim transform and Lava are consistently better than the Lasso, exhibiting also in this real dataset the robustness against confounding variables.

6 Discussion

We propose to add robustness against hidden confounding variables by employing a wisely chosen spectral transformation before using the Lasso or other high-dimensional sparse regression techniques. There is essentially nothing to lose but much to be gained which is in line with the typical argument of robustness (Huber, 2011) We can also take directly the viewpoint of deconfounding before performing further analysis: this is the more common thinking in many applications where hidden confounding is expected to happen, a prime example being genetics (Novembre and Stephens, 2008).

The confounding issue in the context of linear models can be represented and analyzed as a regression problem with coefficient $\beta + b$; the coefficient β is the true underlying parameter in absence of confounding variables, while the perturbation b is due to the confounding. We develop theory for a linear model with regression parameters $\beta + b$ where β is sparse and the perturbation b sufficiently small, a condition satisfied when the confounding is sufficiently 'dense' in the sense that each confounding variable affects many predictors. We show that certain spectral transformations, such as the Trim transform or the PCA adjustment, in conjunction with using the Lasso afterwards, achieve the same ℓ_1-convergence rate of the $\|\widehat{\beta} - \beta\|_1$ as the Lasso for the linear model without confounding; see Section 4 and Theorem 1. Such a theoretical result is entirely new and covers also the Lava method (Chernozhukov et al., 2017). As a consequence, the theoretical result also establishes spectral deconfounding as an excellent method for removing the effect of dense hidden confounders in high-dimensional settings.

Another advantage of our approach is its simplicity: it consists of just one simple pre-transformation step before using the Lasso. It requires the computation of the SVD of the design matrix which has computational complexity of $\mathcal{O}(\min(n^2p, np^2))$ and can be done in a few lines of code.

The topic of deconfounding has not received too much attention, despite its practical importance (Greenland et al., 1999; Brookhart et al., 2010). Here we have shown that it is possible and easy to protect against hidden dense confounding in the case of linear regression. Similar ideas might be powerful as well for more complicated models.

Acknowledgements

The research of D. Ćevid and P. Bühlmann was supported by the European Research Council under the Grant Agreement No 786461 (CausalStats - ERC-2017-ADG). We would like to thank Gian Thanei and Benjamin Frot for providing processed data for our biological applications.

Paper

B

Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding.

Guo, Z., Ćevid, D., Bühlmann, P.

Submitted to Annals of Statistics.

Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding

Zijian Guo[†], Domagoj Ćevid[*], Peter Bühlmann[*]

[†]RUTGERS UNIVERSITY, PISCATAWAY, USA
[*]SEMINAR FÜR STATISTIK, ETH ZÜRICH, 8092 ZÜRICH, SWITZERLAND.

Abstract

Inferring causal relationships or related associations from observational data can be invalidated by the existence of hidden confounding. We focus on a high-dimensional linear regression setting, where the measured covariates are affected by hidden confounding and propose the *Doubly Debiased Lasso* estimator for individual components of the regression coefficient vector. Our advocated method simultaneously corrects both the bias due to estimation of high-dimensional parameters as well as the bias caused by the hidden confounding. We establish its asymptotic normality and also prove that it is efficient in the Gauss-Markov sense. The validity of our methodology relies on a dense confounding assumption, i.e. that every confounding variable affects many covariates. The finite sample performance is illustrated with an extensive simulation study and a genomic application.

Keywords. Causal Inference; Structural Equation Model; Dense Confounding; Linear Model; Spectral Deconfounding

1 Introduction

Observational studies are often used to infer causal relationship in fields such as genetics, medicine, economics or finance. A major concern for confirmatory conclusions is the existence of hidden confounding (Guertin et al., 2016; Manghnani et al., 2018). In this case, standard statistical methods can be severely biased, particularly for large-scale observational studies, where many measured covariates are possibly confounded.

To better address this problem, let us consider first the following linear Structural Equation Model (SEM) with a response Y_i, high-dimensional measured covariates $X_{i,\cdot} \in \mathbb{R}^p$ and hidden confounders $H_{i,\cdot} \in \mathbb{R}^q$:

$$Y_i \leftarrow \beta^\intercal X_{i,\cdot} + \phi^\intercal H_{i,\cdot} + e_i, \quad \text{and} \quad X_{i,\cdot} \leftarrow \Psi^\intercal H_{i,\cdot} + E_{i,\cdot} \quad \text{for } 1 \leqslant i \leqslant n, \tag{1}$$

where the random error $e_i \in \mathbb{R}$ is independent of $X_{i,\cdot} \in \mathbb{R}^p$, $H_{i,\cdot} \in \mathbb{R}^q$ and $E_{i,\cdot} \in \mathbb{R}^p$ and the components of $E_{i,\cdot} \in \mathbb{R}^p$ are uncorrelated with the components of $H_{i,\cdot} \in \mathbb{R}^q$. The focus on a SEM as in (1) is not necessary and we relax this restriction in model (2) below. Such kind of models are used for e.g. biological studies to explore the effects of measured genetic variants on the disease risk factor, and the hidden confounders can be geographic information (Novembre et al., 2008), data sources in mental analysis (Price et al., 2006) or general population stratification in GWAS (McCarthy et al., 2008).

Our aim is to perform statistical inference for individual components β_j, $1 \leqslant j \leqslant p$, of the coefficient vector, where p can be large, in terms of obtaining confidence intervals or statistical tests. This inference problem is challenging due to high dimensionality of the model and the existence of hidden confounders. As a side remark, we mention that our proposed methodology can also be used for certain measurement error models, an important general topic in statistics and economics (Carroll et al., 2006; Wooldridge, 2010).

1.1 Our Results and Contributions

We focus on a dense confounding model, where the hidden confounders $H_{i,\cdot}$ in (1) are associated with many measured covariates $X_{i,\cdot}$. Such dense confounding model seems reasonable in quite many practical applications, e.g. for addressing the problem of batch effects in biological studies (Haghverdi et al., 2018; Johnson et al., 2007; Leek et al., 2010).

We propose a two-step estimator for the regression coefficient β_j for $1 \leqslant j \leqslant p$ in the high-dimensional dense confounding setting, where a large number of covariates has possibly been affected by hidden confounding. In the first step, we construct a penalized spectral deconfounding estimator $\widehat{\beta}^{init}$ as in (Ćevid et al., 2018), where the standard squared error loss is replaced by a squared error loss after applying a certain spectral transformation to the design matrix X and the response Y. In the second step, for the regression coefficient of interest β_j, we estimate the high-dimensional nuisance parameters $\beta_{-j} = \{\beta_l;\ l \neq j\}$ by $\widehat{\beta}^{init}_{-j}$ and construct an approximately unbiased estimator $\widehat{\beta}_j$.

The main idea of the second step is to correct the bias from two sources, one from estimating the high-dimensional nuisance vector β_{-j} by $\widehat{\beta}^{init}_{-j}$ and the other arising from hidden confounding. In the standard high-dimensional regression setting with no hidden confounding, debiasing, desparsifying or Neyman's Orthogonalization were proposed for inference for β_j (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and

Montanari, 2014; Belloni et al., 2014; Chernozhukov et al., 2015; Farrell, 2015; Chernozhukov et al., 2018). However, these methods, or some of its direct extensions, do not account for the bias arising from hidden confounding. In order to address this issue, we introduce a *Doubly Debiased Lasso* estimator which corrects both biases simultaneously. Specifically, we construct a spectral transformation $\mathcal{P}^{(j)} \in \mathbb{R}^{n \times n}$, which is applied to the nuisance design matrix X_{-j} when the parameter of interest is β_j. This spectral transformation is crucial to simultaneously correcting the two sources of bias.

We establish the asymptotic normality of the proposed *Doubly Debiased Lasso* estimator in Theorem 1. An efficiency result is also provided in Theorem 2 of Section 4.2.1, showing that the *Doubly Debiased Lasso* estimator retains the same Gauss-Markov efficiency bound as in standard high-dimensional linear regression with no hidden confounding (van de Geer et al., 2014; Jankova and van de Geer, 2018). Our result is in sharp contrast to Instrumental Variables (IV) based methods, see Section 1.2, whose inflated variance is often of concern, especially with a limited amount of data (Wooldridge, 2010; Boef et al., 2014). This remarkable efficiency result is possible by assuming denseness of confounding. Various intermediary results of independent interest are also derived in the supplementary material of Guo et al. (2020). Finally, the performance of the proposed estimator is illustrated on simulated and real genomic data in Section 5.

To summarize, our main contribution is two-fold:

1. We propose a novel Doubly Debiased Lasso estimator for individual coefficients β_j and estimation of the corresponding standard error in a high-dimensional linear SEM with hidden confounding.

2. We show that the proposed estimator is asymptotically Gaussian and efficient in the Gauss-Markov sense. This implies the construction of asymptotically optimal confidence intervals for individual coefficients β_j.

1.2 Related Work

In econometrics, hidden confounding and measurement errors are unified under the framework of endogenous variables. Inference for treatment effects or corresponding regression parameters in presence of hidden confounders or measurement errors has been extensively studied in the literature with Instrumental Variables (IV) regression. The construction of IVs typically requires a lot of domain knowledge, and obtained IVs are often suspected of violating the main underlying assumptions (Han, 2008; Wooldridge, 2010; Kang et al., 2016; Burgess et al., 2017; Guo et al., 2018; Windmeijer et al., 2019). In high dimensions, the construction of IVs is even more challenging, since for identification of the causal effect, one has to construct as many IVs as the number of confounded covariates, which is the so-called “rank condition” (Wooldridge, 2010). Some recent work on the high-dimensional hidden confounding problem relying on the construction of IVs includes (Gautier and Rose, 2011; Fan and Liao, 2014; Lin et al., 2015; Belloni et al., 2017; Zhu, 2018; Neykov et al., 2018; Gold et al., 2020). Another approach builds on directly estimating and adjusting with respect to latent factors (Wang and Blei, 2019).

A major distinction of the current work from the contributions above is that we consider a confounding model with a denseness assumption (Chandrasekaran et al., 2012; Ćevid et al., 2018; Shah et al., 2020). (Ćevid et al., 2018) consider point estimation of β in the high-dimensional hidden confounding model (1), whereas (Shah et al., 2020) deal with point estimation of the precision and covariance matrix of high-dimensional covariates, which are possibly confounded. The current paper is different in that it considers the challenging problem of confidence interval construction, which requires novel ideas for both methodology and theory.

The dense confounding model is also connected to the high-dimensional factor models (Fan et al., 2008; Lam et al., 2011; Lam and Yao, 2012; Fan et al., 2016; Wang et al., 2017b). The main difference is that the factor model literature focuses on accurately extracting the factors, while our method is essentially filtering them out in order to provide consistent estimators of regression coefficients, under much weaker requirements than for the identification of factors.

Another line of research (Gagnon-Bartsch and Speed, 2012; Sun et al., 2012; Wang et al., 2017a) studies the latent confounder adjustment models but focuses on a different setting where many outcome variables can be possibly associated with a small number of observed covariates and several hidden confounders.

Notation. We use $X_j \in \mathbb{R}^n$ and $X_{-j} \in \mathbb{R}^{n\times(p-1)}$ to denote the $j-$th column of the matrix X and the sub-matrix of X excluding the $j-$th column, respectively; $X_{i,\cdot} \in \mathbb{R}^p$ is used to denote the $i-$th row of the matrix X (as a column vector); $X_{i,j}$ and $X_{i,-j}$ denote respectively the (i,j) entry of the matrix X and the sub-row of $X_{i,\cdot}$ excluding the j-th entry. Let $[p] = \{1, 2, \ldots, p\}$. For a subset $J \subseteq [p]$ and a vector $x \in \mathbb{R}^p$, x_J is the sub-vector of x with indices in J and x_{-J} is the sub-vector with indices in J^c. For a set S, $|S|$ denotes the cardinality of S. For a vector $x \in \mathbb{R}^p$, the ℓ_q norm of x is defined as $\|x\|_q = (\sum_{l=1}^{p} |x_l|^q)^{\frac{1}{q}}$ for $q \geqslant 0$ with $\|x\|_0 = |\{1 \leqslant l \leqslant p : x_l \neq 0\}|$ and $\|x\|_\infty = \max_{1\leqslant l\leqslant p} |x_l|$. We use e_i to denote the i-th standard basis vector in $\mathbb{R}^p$ and I_p to denote the identity matrix of size $p \times p$. We use c and C to denote generic positive constants that may vary from place to place. For a sub-Gaussian random variable X, we use $\|X\|_{\psi_2}$ to denote its sub-Gaussian norm; see definitions 5.7 and 5.22 in (Vershynin, 2012). For a sequence of random variables X_n indexed by n, we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. For a sequence of random variables X_n and numbers a_n, we define $X_n = o_p(a_n)$ if X_n/a_n converges to zero in probability. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means that $\exists C > 0$ such that $a_n \leqslant C b_n$ for all n; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n \to \infty} a_n/b_n = 0$. For a matrix M, we use $\|M\|_F$, $\|M\|_2$ and $\|M\|_\infty$ to denote its Frobenius norm, spectral norm and element-wise maximum norm, respectively. We use $\lambda_j(M)$ to denote the j-th largest singular value of some matrix M, that is, $\lambda_1(M) \geqslant \lambda_2(M) \geqslant \ldots \geqslant \lambda_q(M) \geqslant 0$. For a symmetric matrix A, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its maximum and minimum eigenvalues, respectively.

2 Hidden Confounding Model

We consider the Hidden Confounding Model for i.i.d. data $\{X_{i,\cdot}, Y_i\}_{1 \leqslant i \leqslant n}$ and unobserved i.i.d. confounders $\{H_{i,\cdot}\}_{1 \leqslant i \leqslant n}$, given by:

$$Y_i = \beta^\intercal X_{i,\cdot} + \phi^\intercal H_{i,\cdot} + e_i \quad \text{and} \quad X_{i,\cdot} = \Psi^\intercal H_{i,\cdot} + E_{i,\cdot}, \tag{2}$$

where $Y_i \in \mathbb{R}$ and $X_{i,\cdot} \in \mathbb{R}^p$ respectively denote the response and the measured covariates and $H_{i,\cdot} \in \mathbb{R}^q$ represents the hidden confounders. We assume that the random error $e_i \in \mathbb{R}$ is independent of $X_{i,\cdot} \in \mathbb{R}^p$, $H_{i,\cdot} \in \mathbb{R}^q$ and $E_{i,\cdot} \in \mathbb{R}^p$ and the components of $E_{i,\cdot} \in \mathbb{R}^p$ are uncorrelated with the components of $H_{i,\cdot} \in \mathbb{R}^q$.

The coefficient matrices $\Psi \in \mathbb{R}^{q \times p}$ and $\phi \in \mathbb{R}^{q \times 1}$ encode the linear effect of the hidden confounders $H_{i,\cdot}$ on the measured covariates $X_{i,\cdot}$ and the response Y_i. We consider the high-dimensional setting where p might be much larger than n. Throughout the paper it is assumed that the regression vector $\beta \in \mathbb{R}^p$ is sparse, with a small number k of nonzero components, and that the number q of confounding variables is a small positive integer. However, both k and q are allowed to grow with n and p. We write Σ_E or Σ_X for the covariance matrices of $E_{i,\cdot}$ or $X_{i,\cdot}$, respectively. Without loss of generality, it is assumed that $\mathbb{E} X_{i,\cdot} = 0$, $\mathbb{E} H_{i,\cdot} = 0$, $\mathrm{Cov}(H_{i,\cdot}) = \mathrm{I}_q$ and hence $\Sigma_X = \Psi^\intercal \Psi + \Sigma_E$.

The probability model (2) is more general than the Structural Equation Model in (1). It only describes the observational distribution of the latent variable $H_{i,\cdot}$ and the observed data $(X_{i,\cdot}, Y_i)$, which possibly may be generated from the hidden confounding SEM (1).

Our goal is to construct confidence intervals for the components of β, which in the model (1) describes the causal effect of X on the response Y. The problem is challenging due to the presence of unobserved confounding. In fact, the regression parameter β can not even be identified without additional assumptions. Our main condition addressing this issue is a denseness assumption that the rows $\Psi_{j,\cdot} \in \mathbb{R}^p$ are dense in a certain sense (see Condition (A2) in Section 4), i.e., many covariates of $X_{i,\cdot} \in \mathbb{R}^p$ are simultaneously affected by hidden confounders $H_{i,\cdot} \in \mathbb{R}^q$.

2.1 Representation as a Linear Model

The Hidden Confounding Model (2) can be represented as a linear model for the observed data $\{X_{i,\cdot}, Y_i\}_{1 \leqslant i \leqslant n}$:

$$Y_i = (\beta + b)^\intercal X_{i,\cdot} + \epsilon_i \quad \text{and} \quad X_{i,\cdot} = \Psi^\intercal H_{i,\cdot} + E_{i,\cdot}, \tag{3}$$

by writing

$$\epsilon_i = e_i + \phi^\intercal H_{i,\cdot} - b^\intercal X_{i,\cdot} \quad \text{and} \quad b = \Sigma_X^{-1} \Psi^\intercal \phi.$$

As in (2) we assume that $E_{i,\cdot}$ is uncorrelated with $H_{i,\cdot}$ and, by construction of b, ϵ_i is uncorrelated with $X_{i,\cdot}$. With σ_e^2 denoting the variance of e_i, the variance of the error ϵ_i equals $\sigma_\epsilon^2 = \sigma_e^2 + \phi^\intercal \left(\mathrm{I}_q - \Psi \Sigma_X^{-1} \Psi^\intercal\right) \phi$. In model (3), the response is generated from a linear model where the sparse coefficient vector β has been perturbed by some perturbation vector $b \in \mathbb{R}^p$. This representation reveals how the parameter of interest β is not in general identifiable from observational data, where one can not easily differentiate it from the perturbed coefficient vector $\beta + b$, where the perturbation vector b is induced by hidden confounding. However, as shown in the supplement of Guo et al. (2020), b is dense and $\|b\|_2$ is small for large p under the assumption of dense confounding, which enables us to identify β asymptotically. It is important to note that the term $b^\intercal X_{i,\cdot}$ induced by hidden confounders $H_{i,\cdot}$ is not necessarily small and hence cannot be simply ignored in model (3), but requires novel methodological approach.

Connection to measurement errors We briefly relate certain measurement error models to the Hidden Confounding Model (2). Consider a linear model for the outcome Y_i and covariates $X_{i\cdot}^0 \in \mathbb{R}^p$, where we only observe $X_{i,\cdot} \in \mathbb{R}^p$ with measurement error $W_{i,\cdot} \in \mathbb{R}^p$:

$$Y_i = \beta^\intercal X_{i\cdot}^0 + e_i \quad \text{and} \quad X_{i,\cdot} = X_{i,\cdot}^0 + W_{i,\cdot} \quad \text{for } 1 \leqslant i \leqslant n. \tag{4}$$

Here, e_i is a random error independent of $X_{i,\cdot}^0$ and $W_{i,\cdot}$, and $W_{i,\cdot}$ is the measurement error independent of X_i^0. We can then express a linear dependence of Y_i on the observed $X_{i,\cdot}$,

$$Y_i = \beta^\intercal X_{i,\cdot} + (e_i - \beta^\intercal W_{i,\cdot}) \quad \text{and} \quad X_{i,\cdot} = W_{i,\cdot} + X_{i,\cdot}^0$$

We further assume the following structure of the measurement error:

$$W_{i,\cdot} = \Psi^\intercal H_{i,\cdot},$$

i.e. there exist certain latent variables $H_{i,\cdot} \in \mathbb{R}^q$ that contribute independently and linearly to the measurement error, a conceivable assumption in some practical applications. Combining this with the equation above we get

$$Y_i = \beta^\intercal X_{i,\cdot} + (e_i - \phi^\intercal H_{i,\cdot}) \quad \text{and} \quad X_{i,\cdot} = \Psi^\intercal H_{i,\cdot} + X_{i\cdot}^0, \tag{5}$$

where $\phi = \Psi \beta \in \mathbb{R}^q$. Therefore, the model (5) can be seen as a special case of the model (2), by identifying $X_{i\cdot}^0$ in (5) with $E_{i,\cdot}$ in (2).

3 Doubly Debiased Lasso Estimator

In this section, for a fixed index $j \in \{1, \ldots, p\}$, we propose an inference method for the regression coefficient β_j of the Hidden Confounding Model (2). The validity of the method is demonstrated by considering the equivalent model (3).

3.1 Double Debiasing

We denote by $\widehat{\beta}^{init}$ an initial estimator of β. We will use the spectral deconfounding estimator proposed in (Ćevid et al., 2018), described in detail in Section 3.4. We start from the following decomposition:

$$Y - X_{-j}\widehat{\beta}^{init}_{-j} = X_j\left(\beta_j + b_j\right) + X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j}) + X_{-j}b_{-j} + \epsilon \quad \text{for} \quad j \in \{1, \ldots, p\}. \tag{6}$$

The above decomposition reveals two sources of bias: the bias $X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j})$ due to the error of the initial estimator $\widehat{\beta}^{init}$ and the bias $X_{-j}b_{-j}$ induced by the perturbation vector b in the model (3), arising by marginalizing out the hidden confounding in (2). Note that the bias b_j is negligible in the dense confounding setting, see the supplement of Guo et al. (2020). The first bias, due to penalization, appears in the standard high-dimensional linear regression as well, and can be corrected with the debiasing methods proposed in (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014) when assuming no hidden confounding. However, in presence of hidden confounders, methodological innovation is required for correcting both bias terms and conducting the resulting statistical inference. We propose a novel Doubly Debiased Lasso estimator for correcting both sources of bias simultaneously.

Denote by $\mathcal{P}^{(j)} \in \mathbb{R}^{n \times n}$ a symmetric spectral transformation matrix, which shrinks the singular values of the sub-design $X_{-j} \in \mathbb{R}^{n \times (p-1)}$. The detailed construction, together with some examples, is given in Section 3.3. We shall point out that the construction of the transformation matrix $\mathcal{P}^{(j)}$ depends on which coefficient β_j is our target and hence refer to $\mathcal{P}^{(j)}$ as the nuisance spectral transformation with respect to the coefficient β_j. Multiplying both sides of the decomposition (6) with the transformation $\mathcal{P}^{(j)}$ gives:

$$\mathcal{P}^{(j)}(Y - X_{-j}\widehat{\beta}^{init}_{-j}) = \mathcal{P}^{(j)}X_j\left(\beta_j + b_j\right) + \mathcal{P}^{(j)}X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j}) + \mathcal{P}^{(j)}X_{-j}b_{-j} + \mathcal{P}^{(j)}\epsilon. \tag{7}$$

The quantity of interest β_j appears on the RHS of the equation (7) next to the vector $\mathcal{P}^{(j)}X_j$, whereas the additional bias lies in the span of the columns of $\mathcal{P}^{(j)}X_{-j}$. For this reason, we construct a projection direction vector $\mathcal{P}^{(j)}Z_j \in \mathbb{R}^n$ as the transformed residuals of regressing X_j on X_{-j}:

$$Z_j = X_j - X_{-j}\widehat{\gamma}, \tag{8}$$

where the coefficients $\widehat{\gamma}$ are estimated with the Lasso for the transformed covariates using $\mathcal{P}^{(j)}$:

$$\widehat{\gamma} = \arg\min_{\gamma\in\mathbb{R}^{p-1}} \left\{ \frac{1}{2n}\|\mathcal{P}^{(j)}X_j - \mathcal{P}^{(j)}X_{-j}\gamma\|_2^2 + \lambda_j \sum_{l\neq j} \frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}}|\gamma_l| \right\}, \tag{9}$$

with $\lambda_j = A\sigma_j\sqrt{\log p/n}$ for some positive constant $A > \sqrt{2}$ (for σ_j, see Section 4.1).

Finally, motivated by the equation (7), we propose the following estimator for β_j:

$$\widehat{\beta}_j = \frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}(Y - X_{-j}\widehat{\beta}^{init}_{-j})}{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j}. \tag{10}$$

We refer to this estimator as the Doubly Debiased Lasso estimator as it simultaneously corrects the bias induced by $\widehat{\beta}^{init}$ and the confounding bias $X_{-j}b_{-j}$ by using the spectral transformation $\mathcal{P}^{(j)}$.

In the following, we briefly explain why the proposed estimator estimates β_j well. We start with the following error decomposition of $\widehat{\beta}_j$, derived from (7)

$$\widehat{\beta}_j - \beta_j = \underbrace{\frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}\epsilon}{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j}}_{\text{Variance}} + \underbrace{\frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j})}{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j} + \frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}b_{-j}}{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j}}_{\text{Remaining Bias}} + b_j\,. \tag{11}$$

In the above equation, the bias after correction consists of two components: the remaining bias due to the estimation error of $\widehat{\beta}^{init}_{-j}$ and the remaining confounding bias due to $X_{-j}b_{-j}$ and b_j. These two components can be shown to be negligible in comparison to the variance component under certain model assumptions, see Theorem 1 and its proof for details. Intuitively, the construction of the spectral transformation matrix $\mathcal{P}^{(j)}$ is essential for reducing the bias due to the hidden confounding. The term $\frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}b_{-j}}{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j}$ in equation (11) is of a small order because $\mathcal{P}^{(j)}$ shrinks the leading singular values of X_{-j} and hence $\mathcal{P}^{(j)}X_{-j}b_{-j}$ is significantly smaller than $X_{-j}b_{-j}$. The induced bias $X_{-j}b_{-j}$ is not negligible since b_{-j} points in the direction of leading right singular vectors of X_{-j}, thus leading to $\|\frac{1}{\sqrt{n}}X_{-j}b_{-j}\|_2$ being of constant order. By applying a spectral transformation to shrink the leading singular values, one can show that $\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}b_{-j}\|_2 = O_p(1/\sqrt{\min\{n,p\}})$.

Furthermore, the other remaining bias term $\frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j})}{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j}$ in (11) is small since the initial estimator $\widehat{\beta}^{init}$ is close to β in ℓ_1 norm and $\mathcal{P}^{(j)}Z_j$ and $\mathcal{P}^{(j)}X_{-j}$ are nearly orthogonal due to the construction of $\widehat{\gamma}$ in (9). This bias correction idea is analogous to the Debiased Lasso estimator introduced in (Zhang and Zhang, 2014) for the standard high-dimensional linear regression:

$$\widehat{\beta}^{DB}_j = \frac{(Z^{DB}_j)^\intercal (Y - X_{-j}\widehat{\beta}^{init}_{-j})}{(Z^{DB}_j)^\intercal X_j}, \tag{12}$$

where Z_j^{DB} is constructed similarly as in (8) and (9), but where $\mathcal{P}^{(j)}$ is the identity matrix. Therefore, the main difference between the estimator in (12) and our proposed estimator (10) is that for its construction we additionally apply the nuisance spectral transformation $\mathcal{P}^{(j)}$.

We emphasize that the additional spectral transformation $\mathcal{P}^{(j)}$ is necessary even for just correcting the bias of $\widehat{\beta}_{-j}^{init}$ in presence of confounding (i.e., it is also needed for the first besides the second bias term in (11)). To see this, we define the best linear projection of $X_{1,j}$ to all other variables $X_{1,-j} \in \mathbb{R}^{p-1}$ with the coefficient vector $\gamma = [\mathbb{E}(X_{i,-j}X_{i,-j}^{\intercal})]^{-1}\mathbb{E}(X_{i,-j}X_{i,j}) \in \mathbb{R}^{p-1}$ (which is then estimated by the Lasso in the standard construction of Z_j^{DB}). We notice that γ need not be sparse due to the fact that all covariates are affected by a common set of hidden confounders yielding spurious associations. Hence, the standard construction of Z_j^{DB} in (12) is not favorable in the current setting. In contrast, the proposed method with $\mathcal{P}^{(j)}$ works for two reasons: first, the application of $\mathcal{P}^{(j)}$ in (9) leads to a consistent estimator of the sparse component of γ, denoted as γ^E (see the expression of γ^E given in the supplementary material of Guo et al. (2020)); second, the application of $\mathcal{P}^{(j)}$ leads to a small prediction error $\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma} - \gamma^E)$. We illustrate in Section 5 how the application of $\mathcal{P}^{(j)}$ corrects the bias due to $\widehat{\beta}_{-j}^{init}$ and observe a better empirical coverage after applying $\mathcal{P}^{(j)}$ in comparison to the standard debiased Lasso in (12); see Figure 7.

3.2 Confidence Interval Construction

In Section 4, we establish the asymptotic normal limiting distribution of the proposed estimator $\widehat{\beta}_j$ under certain regularity conditions. Its standard deviation can be estimated by $\sqrt{\frac{\widehat{\sigma}_e^2 \cdot Z_j^{\intercal}(\mathcal{P}^{(j)})^4 Z_j}{[Z_j^{\intercal}(\mathcal{P}^{(j)})^2 X_j]^2}}$ with $\widehat{\sigma}_e$ denoting a consistent estimator of σ_e. The detailed construction of $\widehat{\sigma}_e$ is described in Section 3.5. Therefore, a confidence interval (CI) with asymptotic coverage $1-\alpha$ can be obtained as

$$\mathrm{CI}(\beta_j) = \left(\widehat{\beta}_j - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{\sigma}_e^2 \cdot Z_j^{\intercal}(\mathcal{P}^{(j)})^4 Z_j}{[Z_j^{\intercal}(\mathcal{P}^{(j)})^2 X_j]^2}}, \widehat{\beta}_j + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{\sigma}_e^2 \cdot Z_j^{\intercal}(\mathcal{P}^{(j)})^4 Z_j}{[Z_j^{\intercal}(\mathcal{P}^{(j)})^2 X_j]^2}}\right), \tag{13}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a standard normal random variable.

3.3 Construction of Spectral Transformations

Construction of the spectral transformation $\mathcal{P}^{(j)} \in \mathbb{R}^{n \times n}$ is an essential step for the Doubly Debiased Lasso estimator (10). The transformation $\mathcal{P}^{(j)} \in \mathbb{R}^{n \times n}$ is a symmetric matrix shrinking the leading singular values of the design matrix $X_{-j} \in \mathbb{R}^{n \times (p-1)}$. Denote by $m = \min\{n, p-1\}$ and the SVD of the matrix X_{-j} by $X_{-j} = U(X_{-j})\Lambda(X_{-j})[V(X_{-j})]^{\intercal}$, where $U(X_{-j}) \in \mathbb{R}^{n\times m}$ and $V(X_{-j}) \in \mathbb{R}^{(p-1)\times m}$ have orthonormal columns and $\Lambda(X_{-j}) \in \mathbb{R}^{m\times m}$ is a diagonal matrix of singular values which are sorted in a decreasing order $\Lambda_{1,1}(X_{-j}) \geqslant \Lambda_{2,2}(X_{-j}) \geqslant \ldots \geqslant \Lambda_{m,m}(X_{-j}) \geqslant 0$. We then define the spectral transformation $\mathcal{P}^{(j)}$ for X_{-j} as $\mathcal{P}^{(j)} = U(X_{-j})S(X_{-j})[U(X_{-j})]^\intercal$, where $S(X_{-j}) \in \mathbb{R}^{m\times m}$ is a diagonal shrinkage matrix with $0 \leqslant S_{l,l}(X_{-j}) \leqslant 1$ for $1 \leqslant l \leqslant m$. The SVD for the complete design matrix X is defined analogously. We highlight the dependence of the SVD decomposition on X_{-j}, but for simplicity it will be omitted when there is no confusion. Note that $\mathcal{P}^{(j)}X_{-j} = U\,(S\Lambda)\,V^\intercal$, so the spectral transformation shrinks the singular values $\{\Lambda_{l,l}\}_{1\leqslant l\leqslant m}$ to $\{S_{l,l}\Lambda_{l,l}\}_{1\leqslant l\leqslant m}$, where $\Lambda_{l,l} = \Lambda_{l,l}(X_{-j})$.

Trim transform For the rest of this paper, the spectral transformation that is used is the Trim transform (Ćevid et al., 2018). It limits any singular value to be at most some threshold τ. This means that the shrinkage matrix S is given as: for $1 \leqslant l \leqslant m$,

$$S_{l,l} = \begin{cases} \tau/\Lambda_{l,l} & \text{if} \quad \Lambda_{l,l} > \tau \\ 1 & \text{otherwise} \end{cases}.$$

A good default choice for the threshold τ is the median singular value $\Lambda_{\lfloor m/2\rfloor,\lfloor m/2\rfloor}$, so only the top half of the singular values is shrunk to the bulk value $\Lambda_{\lfloor m/2\rfloor,\lfloor m/2\rfloor}$ and the bottom half is left intact. More generally, one can use any percentile $\rho_j \in (0,1)$ to shrink the top $(100\rho_j)\%$ singular values to the corresponding ρ_j-quantile $\Lambda_{\lfloor\rho_j m\rfloor,\lfloor\rho_j m\rfloor}$. We define the ρ_j-Trim transform $\mathcal{P}^{(j)}$ as

$$\mathcal{P}^{(j)} = U(X_{-j})S(X_{-j})[U(X_{-j})]^\intercal \text{ with } S_{l,l}(X_{-j}) = \begin{cases} \frac{\Lambda_{\lfloor\rho_j m\rfloor,\lfloor\rho_j m\rfloor}(X_{-j})}{\Lambda_{l,l}(X_{-j})} & \text{if } l \leqslant \lfloor\rho_j m\rfloor \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

In Section 4 we investigate the dependence of the asymptotic efficiency of the resulting Doubly Debiased Lasso $\widehat{\beta}_j$ on the percentile choice $\rho_j = \rho_j(n)$. There is a certain trade-off in choosing ρ_j: a smaller value of ρ_j leads to a more efficient estimator, but one needs to be careful to keep $\rho_j m$ sufficiently large compared to the number of hidden confounders q, in order to ensure reduction of the confounding bias. In the supplementary material of Guo et al. (2020), the general conditions that the used spectral transformations need to satisfy in order to ensure good performance of the resulting estimator are described.

3.4 Initial Estimator $\widehat{\beta}^{init}$

For the Doubly Debiased Lasso (10), we use the spectral deconfounding estimator proposed in (Ćevid et al., 2018) as our initial estimator $\widehat{\beta}^{init}$. It uses a spectral transformation $\mathcal{Q} = \mathcal{Q}(X)$, constructed similarly as the transformation $\mathcal{P}^{(j)}$ described in Section

3.3, with the difference that instead of shrinking the singular values of X_{-j}, $\mathcal{Q}$ shrinks the leading singular values of the whole design matrix $X \in \mathbb{R}^{n\times p}$. Specifically, for any percentile $\rho \in (0,1)$, the ρ-Trim transform $\mathcal{Q}$ is given by

$$\mathcal{Q} = U(X)S(X)[U(X)]^{\intercal} \text{ with } S_{l,l}(X) = \begin{cases} \frac{\Lambda_{\lfloor \rho m \rfloor, \lfloor \rho m \rfloor}(X)}{\Lambda_{l,l}(X)} & \text{if } l \leqslant \lfloor \rho m \rfloor \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

The estimator $\widehat{\beta}^{init}$ is computed by applying the Lasso to the transformed data $\mathcal{Q}X$ and $\mathcal{Q}Y$:

$$\widehat{\beta}^{init} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|\mathcal{Q}\left(y - X\beta\right)\|_2^2 + \lambda \sum_{j=1}^{p} \frac{\|\mathcal{Q}X_{\cdot j}\|_2}{\sqrt{n}}|\beta_j|, \tag{16}$$

where $\lambda = A\sigma_e\sqrt{\log p/n}$ is a tuning parameter with $A > \sqrt{2}$.

The transformation $\mathcal{Q}$ reduces the effect of the confounding and thus helps for estimation of β. In the supplementary material of Guo et al. (2020), the ℓ_1 and ℓ_2-error rates of $\widehat{\beta}^{init}$ are given, thereby extending the results of (Ćevid et al., 2018).

3.5 Noise Level Estimator

In addition to an initial estimator of β, we also require a consistent estimator $\widehat{\sigma}_e^2$ of the error variance $\sigma_e^2 = \mathbb{E}(e_i^2)$ for construction of confidence intervals. Choosing a noise level estimator which performs well for a wide range of settings is not easy to do in practice (Reid et al., 2016). We propose using the following estimator:

$$\widehat{\sigma}_e^2 = \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\|\mathcal{Q}y - \mathcal{Q}X\widehat{\beta}^{init}\|_2^2, \tag{17}$$

where $\mathcal{Q}$ is the same spectral transformation as in (16).

The motivation for this estimator is based on the expression

$$\mathcal{Q}y - \mathcal{Q}X\widehat{\beta}^{init} = \mathcal{Q}\epsilon + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb, \tag{18}$$

which follows from the model (3). The consistency of the proposed noise level estimator, formally shown in Proposition 2, follows from the following observations: the initial spectral deconfounding estimator $\widehat{\beta}^{init}$ has a good rate of convergence for estimating β; the spectral transformation $\mathcal{Q}$ significantly reduces the additional error Xb induced by the hidden confounders; $\|\mathcal{Q}\epsilon\|_2^2/\mathrm{Tr}(\mathcal{Q}^2)$ consistently estimates σ_ϵ^2. Additionally, the dense confounding model is shown to lead to a small difference between the noise levels σ_ϵ^2 and σ_e^2, see the supplement of Guo et al. (2020). In Section 4 we show that variance estimator $\widehat{\sigma}_e^2$ defined in (17) is a consistent estimator of σ_e^2.

3.6 Method Overview and Choice of the Tuning Parameters

The Doubly Debiased Lasso needs specification of various tuning parameters. A good and theoretically justified rule of thumb is to use the Trim transform with $\rho = \rho_j = 1/2$, which shrinks the large singular values to the median singular value, see (14). Furthermore, similarly to the standard Debiased Lasso (Zhang and Zhang, 2014), our proposed method involves the regularization parameters λ in the Lasso regression for the initial estimator $\widehat{\beta}^{init}$ (see equation (16)) and λ_j in the Lasso regression for the projection direction $\mathcal{P}^{(j)} Z_j$ (see equation (9)). For choosing λ we use 10-fold cross-validation, whereas for λ_j, we increase slightly the penalty chosen by the 10-fold cross-validation, so that the variance of our estimator, which can be determined from the data up to a proportionality factor σ_e^2, increases by 25%, as proposed in (Dezeure et al., 2017).

4 Theoretical Justification

This section provides theoretical justifications of the proposed method for the Hidden Confounding Model (2). The proofs of the main results together with several other technical results of independent interest can be found in the supplementary material of Guo et al. (2020).

4.1 Model assumptions

In the following, we fix the index $1 \leqslant j \leqslant p$ and introduce the model assumptions for establishing the asymptotic normality of our proposed estimator $\widehat{\beta}_j$ defined in (10). For the coefficient matrix $\Psi \in \mathbb{R}^{q \times p}$ in (3), we use $\Psi_j \in \mathbb{R}^q$ to denote the j-th column and $\Psi_{-j} \in \mathbb{R}^{q \times (p-1)}$ denotes the sub-matrix with the remaining $p-1$ columns. Furthermore, we write γ for the best linear approximation of $X_{1,j} \in \mathbb{R}$ by $X_{1,-j} \in \mathbb{R}^{p-1}$, that is $\gamma = \arg\min_{\gamma' \in \mathbb{R}^{p-1}} \mathbb{E}(X_{1,j} - X_{1,-j}\gamma')^2$, whose explicit expression is:

$$\gamma = [\mathbb{E}(X_{1,-j} X_{1,-j}^{\intercal})]^{-1} \mathbb{E}(X_{1,-j} X_{1,j}).$$

For ease of notation, we do not explicitly express the dependence of γ on j. Similarly, define

$$\gamma^E = [\mathbb{E}(E_{1,-j} E_{1,-j}^{\intercal})]^{-1} \mathbb{E}(E_{1,-j} E_{1,j}).$$

We denote the corresponding residuals by $\eta_{i,j} = X_{i,j} - X_{i,-j}^{\intercal}\gamma$ and $\nu_{i,j} = E_{i,j} - E_{i,-j}^{\intercal}\gamma^E$ for $1 \leqslant i \leqslant n$. We use σ_j to denote the standard deviation of $\nu_{i,j}$.

The first assumption is on the precision matrix of $E_{i,\cdot} \in \mathbb{R}^p$ in (2):

(A1) The precision matrix $\Omega_E = [\mathbb{E}(E_{i,\cdot}E_{i,\cdot}^{\intercal})]^{-1}$ satisfies $c_0 \leqslant \lambda_{\min}(\Omega_E) \leqslant \lambda_{\max}(\Omega_E) \leqslant C_0$ and $\|(\Omega_E)_{\cdot,j}\|_0 \leqslant s$ where $C_0 > 0$ and $c_0 > 0$ are some positive constants and s denotes the sparsity level which can grow with n and p.

Such assumptions on well-posedness and sparsity are commonly required for estimation of the precision matrix (Meinshausen and Bühlmann, 2006; Lam et al., 2009; Yuan, 2010; Cai et al., 2011) and are also used for confidence interval construction in the standard high-dimensional regression model without unmeasured confounding (van de Geer et al., 2014). Here, the conditions are not directly imposed on the covariates $X_{i,\cdot}$, but rather on their unconfounded part $E_{i,\cdot}$.

The second assumption is about the coefficient matrix Ψ in (3), which describes the effect of the hidden confounding variables $H_{i,\cdot} \in \mathbb{R}^q$ on the measured variables $X_{i,\cdot} \in \mathbb{R}^p$:

(A2) The q-th singular value of the coefficient matrix $\Psi_{-j} \in \mathbb{R}^{q\times(p-1)}$ satisfies

$$\lambda_q(\Psi_{-j}) \gg l(n,p,q) := \max\left\{M\sqrt{\frac{qp}{n}}(\log p)^{3/4}, \sqrt{Mq}p^{1/4}(\log p)^{3/8}, \sqrt{qn\log p}\right\} \tag{19}$$

where M is the sub-Gaussian norm for components of $X_{i,\cdot}$, as defined in Assumption (A3). Furthermore, we have

$$\max\{\|\Psi(\Omega_E)_{\cdot,j}\|_2, \|\Psi_j\|_2, \|\Psi_{-j}(\Omega_E)_{-j,j}\|_2, \|\phi\|_2\} \lesssim \sqrt{q}(\log p)^c, \tag{20}$$

where Ψ and ϕ are defined in (2) and $0 < c \leqslant 1/4$ is some positive constant.

The condition (A2) is crucial for identifying the coefficient β_j in the high-dimensional Hidden Confounding Model (2). Condition (A2) is referred to as the dense confounding assumption. A few remarks are in order regarding when this identifiability condition holds.

Since all vectors $\Psi(\Omega_E)_{\cdot,j}$, Ψ_j, $\Psi_{-j}(\Omega_E)_{-j,j}$ and ϕ are q-dimensional, the upper bound condition (20) on their ℓ_2 norms is mild. If the vector $\phi \in \mathbb{R}^q$ has bounded entries and the vectors $\{\Psi_{\cdot,l}\}_{1\leqslant l\leqslant p} \in \mathbb{R}^q$ are independently generated with zero mean and bounded second moments, then the condition (20) holds with probability larger than $1-(\log p)^{-2c}$, where c is defined in (20).

In the factor model literature (Fan et al., 2013; Wang et al., 2017b) the spiked singular value condition $\lambda_q(\Psi) \asymp \sqrt{p}$ is quite common and holds under mild conditions. The Hidden Confounding Model is closely related to the factor model, where the hidden confounders $H_{i,\cdot}$ are the factors and the matrix Ψ describes how these factors affect the observed variables $X_{i,\cdot}$. However, for our analysis, our assumption on $\lambda_q(\Psi_{-j})$ in (19) can be much weaker than the classical factor assumption $\lambda_q(\Psi_{-j}) \asymp \sqrt{p}$, especially for a range of dimensionality where $p \gg n$. In certain dense confounding settings, we can show that condition (19) holds with high probability. Consider first the special case with a single hidden confounder, that is, $q = 1$ and the effect matrix is reduced to a vector $\Psi \in \mathbb{R}^p$. In this case, $\lambda_1(\Psi_{-j}) = \|\Psi_{-j}\|_2$ and the denseness of the effect vector Ψ_{-j} leads to a large $\lambda_1(\Psi_{-j})$. The condition (19) can be satisfied even if only a certain proportion of covariates is affected by hidden confounding. When $q = 1$, if we assume that there exists a set $A \subseteq \{1, 2, \ldots, p\}$ such that $\{\Psi_l\}_{l \in A}$ are i.i.d. and $|A| \gg l(n, p, q)^2$, where $l(n, p, q)$ is defined in (19), then with high probability $\lambda_q(\Psi) \gtrsim \sqrt{|A|} \gg l(n, p, q)$. In the multiple hidden confounders setting, if the vectors $\{\Psi_l\}_{l \in A}$ are generated as i.i.d. sub-Gaussian random vectors, which has an interpretation that all covariates are analogously affected by the confounders, then the spiked singular value condition (19) is satisfied with high probability as well. See the supplementary material of Guo et al. (2020) for the exact statement. In Section 5.1, we also explore the numerical performance of the method when different proportions of the covariates are affected and observe that the proposed method works well even if the hidden confounders only affect a small percentage of the covariates, say 5%.

Under the model (2), if the entries of Ψ are assumed to be i.i.d. sub-Gaussian with zero mean and variance σ_Ψ^2, then we have $\lambda_q(\Psi_{-j}) \asymp \sqrt{p}\sigma_\Psi$ with high probability. Together with (19), this requires

$$\sigma_\Psi \gg \max\left\{ M\sqrt{\frac{q}{n}}(\log p)^{3/4}, \sqrt{\frac{qn\log p}{p}}, \frac{\sqrt{qM(\log p)^{3/4}}}{p^{1/4}} \right\},$$

So if $p \gg qn\log p$ and $\min\{n, p\} \gg q^3(\log p)^{3/2}M^2$, then the required effect size σ_Ψ of the hidden confounder $H_{i,\cdot}$ on an individual covariate $X_{i,j}$ can diminish to zero fairly quickly.

The condition (19) can in fact be empirically checked using the sample covariance matrix $\widehat{\Sigma}_X$. Since $\Sigma_X = \Psi^\intercal\Psi + \Sigma_E$, then the condition (19) implies that Σ_X has at least q spiked eigenvalues. If the population covariance matrix Σ_X has a few spikes, the corresponding sample covariance matrix will also have spiked eigenvalue structure with a high probability (Wang et al., 2017b). Hence, we can inspect the spectrum of the sample covariance matrix $\widehat{\Sigma}_X$ and informally check whether it has spiked singular values. See the left panel of Figure 2 for an illustration.

The third assumption is imposed on the distribution of various terms:

(A3) The random error e_i in (2) is assumed to be independent of $(X_{i,\cdot}^\intercal, H_{i,\cdot}^\intercal)^\intercal$, the error vector $E_{i,\cdot}$ is assumed to be independent of the hidden confounder $H_{i,\cdot}$, and the noise term $\nu_{i,j} = E_{i,j} - E_{i,-j}^\intercal\gamma^E$ is assumed to be independent of $E_{i,-j}$. Furthermore, $E_{i,\cdot}$ is a sub-Gaussian random vector and e_i and $\nu_{i,j}$ are sub-Gaussian random variables, whose sub-Gaussian norms satisfy $\max\{\|E_{i,\cdot}\|_{\psi_2}, \|e_i\|_{\psi_2}, \max_{1 \leqslant l \leqslant p}\|\nu_{i,l}\|_2\} \leqslant C$, where $C > 0$ is a positive constant independent of n and p. For $1 \leqslant l \leqslant p$, $X_{i,l}$ are sub-Gaussian random variables whose sub-Gaussian norms satisfy $\max_{1 \leqslant l \leqslant p}\|X_{i,l}\|_{\psi_2} \leqslant M$, where $1 \lesssim M \lesssim \sqrt{n/\log p}$.

The independence assumption between the random error e_i and $(X_{i,\cdot}^{\intercal}, H_{i,\cdot}^{\intercal})^{\intercal}$ is commonly assumed for the SEM (1) and thus it holds in the induced Hidden Confounding Model (2) as well, see for example (Pearl, 2009). Analogously, when modelling $X_{i,\cdot}$ as a SEM where the hidden variables $H_{i,\cdot}$ are directly influencing $X_{i,\cdot}$, that is, they are parents of the $X_{i,\cdot}$'s, the independence of $E_{i,\cdot}$ from $H_{i,\cdot}$ is a standard assumption. The independence assumption between $\nu_{i,j}$ and $E_{i,-j}$ holds automatically if $E_{i,\cdot}$ has a multivariate Gaussian distribution (but $X_{i,\cdot}$ is still allowed to be non-Gaussian, e.g. due to non-Gaussian confounders).

We emphasize that the individual components $X_{i,j}$ are assumed to be sub-Gaussian, instead of the whole vector $X_{i,\cdot} \in \mathbb{R}^p$. The sub-Gaussian norm M is allowed to grow with q and p. Particularly, if we assume $H_{i,\cdot}$ to be a sub-Gaussian vector, then condition (20) implies that $M \lesssim \sqrt{q}(\log p)^c \|H_{i,\cdot}\|_{\psi_2}$. Furthermore, our theoretical analysis also covers the case when the sub-Gaussian norm M is of constant order. This happens, for example, when the entries of Ψ are of order $1/\sqrt{q}$, since $M \asymp \max_{l=1,\dots,p} \|\Psi_l\|_2$.

The final assumption is that the restricted eigenvalue condition (Bickel et al., 2009) for the transformed design matrices $\mathcal{Q}X$ and $\mathcal{P}^{(j)}X_{-j}$ is satisfied with high probability.

(A4) With probability at least $1 - \exp(-cn)$, we have

$$\mathrm{RE}\left(\tfrac{1}{n} X^{\intercal}\mathcal{Q}^2 X\right) = \inf_{\substack{\mathcal{T}\subseteq[p] \\ |\mathcal{T}|\leqslant k}} \min_{\substack{\omega\in\mathbb{R}^p \\ \|\omega_{\mathcal{T}^c}\|_1 \leqslant CM\|\omega_{\mathcal{T}}\|_1}} \frac{\omega^{\intercal}\left(\frac{1}{n}X^{\intercal}\mathcal{Q}^2 X\right)\omega}{\|\omega\|_2^2} \geqslant \tau_*; \tag{21}$$

$$\mathrm{RE}\left(\tfrac{1}{n} X_{-j}^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}\right) = \inf_{\substack{\mathcal{T}\subseteq[p]\backslash\{j\} \\ |\mathcal{T}|\leqslant s}} \min_{\substack{\omega\in\mathbb{R}^{p-1} \\ \|\omega_{\mathcal{T}^c}\|_1 \leqslant CM\|\omega_{\mathcal{T}}\|_1}} \frac{\omega^{\intercal}(\frac{1}{n}X_{-j}^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j})\omega}{\|\omega\|_2^2} \geqslant \tau_* \tag{22}$$

where $c, C, \tau_* > 0$ are positive constants independent of n and p and M is the sub-Gaussian norm for components of $X_{i,\cdot}$, as defined in Assumption (A3). For ease of notation, the same constants τ_* and C are used in (21) and (22).

Such assumptions are common in the high-dimensional statistics literature, see (Bühlmann and van de Geer, 2011). The restricted eigenvalue condition (A4) is similar, but more complicated than the standard restricted eigenvalue condition introduced in (Bickel et al., 2009). The main complexity is that, rather than for the original design matrix, the restricted eigenvalue condition is imposed on the transformed design matrices $\mathcal{P}^{(j)}X_{-j}$ and $\mathcal{Q}X$, after applying the Trim transforms $\mathcal{P}^{(j)}$ and $\mathcal{Q}$, described in detail in Sections 3.3 and 3.4, respectively. In the following, we verify the restricted eigenvalue condition (A4) for $\frac{1}{n}X^{\intercal}\mathcal{Q}^2 X$ and the argument can be extended to $\frac{1}{n}X_{-j}^{\intercal}(\mathcal{P}^{(j)})^2 X_{-j}$.

Proposition 1. *Suppose that assumptions (A1) and (A3) hold, $H_{i,\cdot}$ is a sub-Gaussian random vector, $q + \log p \lesssim \sqrt{n}$ and $k = \|\beta\|_0$ satisfies $M^2 k q^2 \log p \log n / n \to 0$. Assume further that the loading matrix $\Psi \in \mathbb{R}^{q \times p}$ satisfies $\|\Psi\|_\infty \lesssim \sqrt{\log(qp)}$, $\lambda_1(\Psi)/\lambda_q(\Psi) \lesssim 1$ and that*

$$\lambda_q(\Psi) \gg \frac{\sqrt{Mp}\max\{k^{1/4}q^{5/4}, 1\}\log(np)}{\min\{n, p\}^{1/4}}. \tag{23}$$

If $\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n}XX^\intercal) \geqslant c\max\{1, p/n\}$ for ρ defined in (15) and some positive constant $c > 0$ independent of n and p, then there exist positive constants $c_1, c_2 > 0$ such that, with probability larger than $1 - p^{-c_2} - \exp(-c_2 n)$, we have $\mathrm{RE}\left(\frac{1}{n}X^\intercal \mathcal{Q}^2 X\right) \geqslant c_1 \lambda_{\min}(\Sigma_X)$.

An important condition for establishing Proposition 1 is the condition (23). Under the commonly assumed spiked singular value condition $\lambda_q(\Psi) \asymp \sqrt{p}$ (Fan et al., 2013; Wang et al., 2017b; Bai, 2003; Bai and Ng, 2002), the condition (23) is reduced to $k \ll \min\{n, p\}/(M^2 q^5 \log(np)^4)$. As a comparison, for the standard high-dimensional regression model with no hidden confounders, (Zhou, 2009; Raskutti et al., 2010) verified the restricted eigenvalue condition under the sparsity condition $k \ll n/\log p$. That is, if $\lambda_q(\Psi) \asymp \sqrt{p}$, then the sparsity requirement in Proposition 1 is the same as that for the high-dimensional regression model with no hidden confounders, up to a polynomial order of q and $\log(np)$,

In comparison to the condition (19) in (A2), (23) can be slightly stronger for a range of dimensionality where $p \gg n^{3/2}$. However, Proposition 1 does not require the strong spiked singular value condition $\lambda_q(\Psi) \asymp \sqrt{p}$. The proof of Proposition 1 is presented in the supplement of Guo et al. (2020). The condition $\lambda_{\lfloor \rho m \rfloor}(\frac{1}{n}XX^\intercal) \geqslant c\max\{1, p/n\}$ can be empirically verified from the data. In the supplementary material of Guo et al. (2020), further theoretical justification for this condition is provided, under mild assumptions.

4.2 Main Results

In this section we present the most important properties of the proposed estimator (10). We always consider asymptotic expressions in the limit where both $n, p \to \infty$ and focus on the high-dimensional regime with $c^* = \lim p/n \in (0, \infty]$. We mention here that some new results on point estimation of the initial estimator $\widehat{\beta}^{init}$ defined in (16) are given in the supplementary material of Guo et al. (2020), as they are established under more general conditions than in (Ćevid et al., 2018).

4.2.1 Asymptotic normality

We first present the limiting distribution of the proposed Doubly Debiased Lasso estimator. The proof of Theorem 1 and important intermediary results for establishing

Theorem 1 are presented in the supplementary material of Guo et al. (2020)

Theorem 1. *Consider the Hidden Confounding Model* (2)*. Suppose that conditions* (A1)-(A4) *hold and further assume that* $c^* = \lim p/n \in (0, \infty]$, $k := \|\beta\|_0 \ll \sqrt{n}/(M^3 \log p)$, $s := \|(\Omega_E)_{\cdot,j}\|_0 \ll n/(M^2 \log p)$ *and* $e_i \sim N(0, \sigma_e^2)$*. Let the tuning parameters for* $\widehat{\beta}^{init}$ *in* (16) *and* $\widehat{\gamma}$ *in* (9) *respectively be* $\lambda \asymp \sigma_e \sqrt{\log p/n} + \sqrt{q \log p/\lambda_q^2(\Psi)}$ *and* $\lambda_j \asymp \sigma_j \sqrt{\log p/n} + \sqrt{q \log p/\lambda_q^2(\Psi_{-j})}$*. Furthermore, let* $\mathcal{Q}$ *and* $\mathcal{P}^{(j)}$ *be the Trim transform* (14) *with* $\min\{\rho, \rho_j\} \geqslant (q+1)/\min\{n, p-1\}$ *and* $\max\{\rho, \rho_j\} < 1$*. Then the Doubly Debiased Lasso estimator* (10) *satisfies*

$$\frac{1}{\sqrt{V}} \left(\widehat{\beta}_j - \beta_j\right) \xrightarrow{d} N(0,1), \tag{24}$$

where

$$\mathrm{V} = \frac{\sigma_e^2 Z_j^{\intercal} (\mathcal{P}^{(j)})^4 Z_j}{[Z_j^{\intercal} (\mathcal{P}^{(j)})^2 X_j]^2} \quad \textit{and} \quad V^{-1} \frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} \xrightarrow{p} 1. \tag{25}$$

Remark 1. The Gaussianity of the random error e_i is mainly imposed to simplify the proof of asymptotic normality. We believe that this assumption is a technical condition and can be removed by applying more refined probability arguments as in (Götze and Tikhomirov, 2002), where the asymptotic normality of quadratic forms $(\mathcal{P}^{(j)} e)^{\intercal} \mathcal{P}^{(j)} e$ is established for the general sub-Gaussian case. The argument could be extended to obtain the asymptotic normality for $(\mathcal{P}^{(j)} \eta_j)^{\intercal} \mathcal{P}^{(j)} e$, which is essentially needed for the current result.

Remark 2. For constructing $\mathcal{Q}$ and $\mathcal{P}^{(j)}$, the main requirement is to trim the singular values enough in both cases, that is, $\min\{\rho, \rho_j\} \geqslant (q+1)/\min\{n, p-1\}$. This condition is mild in the high-dimensional setting with a small number of hidden confounders. Our results are not limited to the proposed estimator which uses the Trim transform $\mathcal{P}^{(j)}$ in (14) and the penalized estimators $\widehat{\gamma}$ and $\widehat{\beta}^{init}$ in (9) and (16), but hold for any any initial estimator and transformation that satisfy the conditions given in the supplementary material of Guo et al. (2020).

Remark 3. If we further assume the error ϵ_i in the model (3) to be independent of $X_{i,\cdot}$, then the requirement (19) of the condition (A2) can be relaxed to

$$\lambda_q(\Psi_{-j}) \gg \max\left\{ M\sqrt{\frac{qp}{n}} (\log p)^{3/4}, \sqrt{qM} p^{1/4} (\log p)^{3/8}, \sqrt{(sM^2 + k\sqrt{n} M^3) q \log p} \right\}.$$

Note that the factor model implies the upper bound $\lambda_q(\Psi_{-j}) \lesssim \sqrt{p}$. Even if $n \geqslant p$, the above condition on $\lambda_q(\Psi_{-j})$ can still hold if $p \gg kqM^3 \log p \sqrt{n}$. On the other hand, the condition (19) together with $\lambda_q(\Psi_{-j}) \lesssim \sqrt{p}$ imply that $p \gg qn \log p$, which excludes the setting $n \geqslant p$.

There are three conditions on the parameters s, q, k imposed in the Theorem 1 above. The most stringent one is the sparsity assumption $k \ll \sqrt{n}/[M^3 \log p]$. In standard high-dimensional sparse linear regression, a related sparsity assumption $k \ll \sqrt{n}/\log p$ has also been used for confidence interval construction (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014) and has been established in (Cai and Guo, 2017) as a necessary condition for constructing adaptive confidence intervals. In the high-dimensional Hidden Confounding Model with $M \asymp 1$, the condition on the sparsity of β is then of the same asymptotic order as in the standard high-dimensional regression with no hidden confounding. The condition on the sparsity of the precision matrix, $s = \|(\Omega_E)_{\cdot,j}\|_0 \ll n/(M^2 \log p)$, is mild in the sense that, for $M \asymp 1$, it is the maximal sparsity level for identifying $(\Omega_E)_{\cdot,j}$. Implied by (19), the condition that the number of hidden confounders q is small is fundamental for all reasonable factor or confounding models.

4.2.2 Efficiency

We investigate now the dependence of the asymptotic variance V in (25) on the choice of the spectral transformation $\mathcal{P}^{(j)}$. We further show that the proposed Doubly Debiased Lasso estimator (10) is efficient in the Gauss-Markov sense, with a careful construction of the transformation $\mathcal{P}^{(j)}$.

The Gauss-Markov theorem states that the smallest variance of any unbiased linear estimator of β_j in the standard low-dimensional regression setting (with no hidden confounding) is $\sigma_e^2/(n\sigma_j^2)$, which we use as a benchmark. The corresponding discussion on efficiency of the standard high-dimensional regression can be found in Section 2.3.3 of (van de Geer et al., 2014). The expression for the asymptotic variance V of our proposed estimator (10) is given by $\frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]}$ (see Theorem 1). For the Trim transform defined in (14), which trims top $(100\rho_j)\%$ of the singular values, we have that

$$\frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} = \frac{\sigma_e^2}{\sigma_j^2} \cdot \frac{\sum_{l=1}^m S_{l,l}^4}{(\sum_{l=1}^m S_{l,l}^2)^2},$$

where we write $m = \min\{n, p-1\}$ and $S_{l,l} = S_{l,l}(X_{-j}) \in [0,1]$. Since $S_{l,l}^4 \leqslant S_{l,l}^2$ for every l, $\sum_{l=1}^m S_{l,l}^2 \geqslant (1-\rho_j)m$ and $(\sum_{l=1}^m S_{l,l}^2)^2 \leqslant m \cdot \sum_{l=1}^m S_{l,l}^4$, we obtain

$$\frac{\sigma_e^2}{\sigma_j^2 m} \leqslant \frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} \leqslant \frac{1}{1-\rho_j} \cdot \frac{\sigma_e^2}{\sigma_j^2 m}.$$

In the high-dimensional setting where $p - 1 \geqslant n$, we have $m = n$ and then

$$\frac{\sigma_e^2}{\sigma_j^2 n} \leqslant \frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} \leqslant \frac{1}{1-\rho_j} \cdot \frac{\sigma_e^2}{\sigma_j^2 n}. \tag{26}$$

Theorem 2. *Suppose that the assumptions of Theorem 1 hold. If* $p \geqslant n + 1$ *and* $\rho_j = \rho_j(n) \to 0$*, then the Doubly Debiased Lasso estimator in* (10) *has asymptotic variance* $\frac{\sigma_e^2}{\sigma_j^2 n}$*, that is, it achieves the Gauss-Markov efficiency bound.*

The above theorem shows that in the $q \ll n$ regime, the Doubly Debiased Lasso achieves the Gauss-Markov efficiency bound if $\rho_j = \rho_j(n) \to 0$ and $\min\{\rho, \rho_j\} \geqslant (q+1)/n$ (which is also a condition of Theorem 1). When using the median Trim transform, i.e. $\rho_j = 1/2$, the bound in (26) implies that the variance of the resulting estimator is at most twice the size of the Gauss-Markov bound. In Section 5, we illustrate the finite-sample performance of the Doubly Debiased Lasso estimator for different values of ρ_j; see Figure 6.

In general for the high-dimensional setting $p/n \to c^* \in (0, \infty]$, the Asymptotic Relative Efficiency (ARE) of the proposed Doubly Debiased Lasso estimator with respect to the Gauss-Markov efficiency bound satisfies the following:

$$\text{ARE} \in \left[\frac{1}{\min\{c^*, 1\}}, \frac{1}{(1-\rho^*)\min\{c^*, 1\}}\right], \tag{27}$$

where $\rho^* = \lim_{n\to\infty} \rho_j(n) \in [0, 1)$. The equation (27) reveals how the efficiency of the Doubly Debiased Lasso is affected by the choice of the percentile $\rho_j = \rho_j(n)$ in transformation $\mathcal{P}^{(j)}$ and the dimensionality of the problem. Smaller ρ_j leads to a more efficient estimator, as long as the top few singular values are properly shrunk. Intuitively, a smaller percentile ρ_j means that less information in X_{-j} is trimmed out and hence the proposed estimator is more efficient. In addition, for the case $\rho^* = 0$, we have $\text{ARE} = \max\{1/c^*, 1\}$. With $\rho^* = 0$, a plot of ARE with respect to the ratio $c^* = \lim p/n$ is given in Figure 1. We see

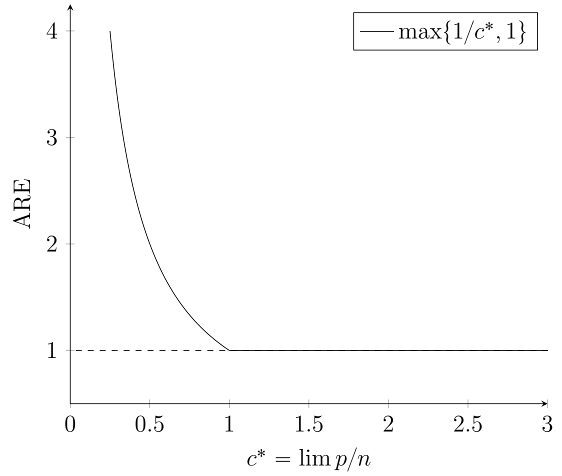


Figure 1: The plot of ARE versus $c^* = \lim p/n$, for the setting of $\rho^* = 0$.

that for $c^* < 1$ (that is $p < n$), the relative efficiency of the proposed estimator increases as the dimension p increases and when $c^* \geqslant 1$ (that is $p \geqslant n$), we have that $\text{ARE} = 1$, saying that the Doubly Debiased Lasso achieves the efficiency bound in the Gauss-Markov sense.

The phenomenon that the efficiency is retained even in presence of hidden confounding is quite remarkable. For comparison, even in the classical low-dimensional setting, the most commonly used approach assumes availability of sufficiently many instrumental variables (IV) satisfying certain stringent conditions under which one can consistently estimate the effects in presence of hidden confounding. In Theorem 5.2 of (Wooldridge, 2010), the popular IV estimator, two-stage-least-squares (2SLS), is shown to have variance strictly larger than the efficiency bound in the Gauss-Markov setting (with no unmeasured confounding). It has been also shown in Theorem 5.3 of (Wooldridge, 2010) that the 2SLS estimator is efficient in the class of all linear instrumental variable estimators and thus, all linear instrumental variable estimators are strictly less efficient than our Doubly Debiased Lasso. On the other hand, our proposed method not only avoids the difficult step of coming up with a large number of valid instrumental variables, but also achieves the efficiency bound with a careful construction of the spectral transformation $\mathcal{P}^{(j)}$. This occurs due to a blessing of dimensionality and the assumption of dense confounding, where a large number of covariates are assumed to be affected by a small number of hidden confounders.

4.2.3 Asymptotic validity of confidence intervals

The asymptotic normal limiting distribution in Theorem 1 can be used for construction of confidence intervals for β_j. Consistently estimating the variance V of our estimator, defined in (25), requires a consistent estimator of the error variance σ_e^2. The following proposition establishes the rate of convergence of the estimator $\widehat{\sigma}_e^2$ proposed in (17):

Proposition 2. *Consider the Hidden Confounding Model* (2). *Suppose that conditions* (A1)-(A4) *hold. Suppose further that* $c^* = \lim p/n \in (0,\infty]$, $k \lesssim n/\log p$ *and* $q \ll \min\{n, p/\log p\}$. *Then with probability larger than* $1 - \exp(-ct^2) - \frac{1}{t^2} - c(\log p)^{-1/2} - n^{-c}$ *for some positive constant* $c > 0$ *and for any* $0 < t \leqslant \sqrt{n}$, *we have*

$$|\widehat{\sigma}_e^2 - \sigma_e^2| \lesssim \frac{t}{\sqrt{n}} + M^2 k \frac{\log p}{n} + \frac{q \log p}{p} + \frac{pq\sqrt{\log p}/n + M^2 kq \log p}{\lambda_q^2(\Psi)},$$

where M *is the sub-Gaussian norm for components of* $X_{i,\cdot}$ *defined in Assumption (A3).*

Together with (19) of the condition (A2), we apply the above proposition and establish $\widehat{\sigma}_e^2 - \sigma_e^2 \xrightarrow{p} 0$. As a remark, the estimation error $|\widehat{\sigma}_e^2 - \sigma_\epsilon^2|$ is of the same order of magnitude as $|\widehat{\sigma}_e^2 - \sigma_e^2|$ since the difference $\sigma_\epsilon^2 - \sigma_e^2$ is small in the dense confounding model.

Proposition 2, together with Theorem 1, imply the asymptotic coverage and precision properties of the proposed confidence interval $\mathrm{CI}(\beta_j)$, described in (13):

Corollary 1. Suppose that the conditions of Theorem 1 hold, then the confidence interval defined in (13) satisfies the following properties:

$$\liminf_{n,p\to\infty} \mathbb{P}\left(\beta_j \in \mathrm{CI}(\beta_j)\right) \geqslant 1-\alpha, \tag{28}$$

$$\limsup_{n,p\to\infty} \mathbb{P}\left(\mathbf{L}\left(\mathrm{CI}(\beta_j)\right) \geqslant (2+c) z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]}}\right) = 0, \tag{29}$$

for any positive constant $c > 0$, where $\mathbf{L}\left(\mathrm{CI}(\beta_j)\right)$ denotes the length of the proposed confidence interval.

Similarly to the efficiency results in Section 4.2.2, the exact length depends on the construction of the spectral transformation $\mathcal{P}^{(j)}$. Together with (26), the above proposition shows that the length of constructed confidence interval is shrinking at the rate of $n^{-1/2}$ for the Trim transform in the high-dimensional setting. Specifically, for the setting $p \geqslant n+1$, if we choose $\rho_j = \rho_j(n) \geqslant (q+1)/n$ and $\rho_j(n) \to 0$, the constructed confidence interval has asymptotically optimal length.

5 Empirical results

In this section we consider the practical aspects of Doubly Debiased Lasso methodology and illustrate its empirical performance on both real and simulated data. The overview of the method and the tuning parameters selection can be found in Section 3.6.

In order to investigate whether the given data set is potentially confounded, one can inspect the principal components of the design matrix X, or equivalently consider its SVD. Spiked singular value structure (see Figure 2) indicates the existence of hidden confounding, as much of the variance of our data can be explained by a small number of latent factors. This also serves as an informal check of the spiked singular value condition in the assumption (A2).

The scree plot can also be used for choosing the trimming thresholds, if one wants to depart from the default median rule (see Section 3.6). We have seen from the theoretical considerations in Section 4 that we can reduce the estimator variance by decreasing the trimming thresholds for the spectral transformation $\mathcal{P}^{(j)}$. On the other hand, it is crucial to choose them so that the number of shrunk singular values is still sufficiently large compared to the number of confounders. However, exactly estimating the number of confounders, e.g. by detecting the elbow in the scree plot (Wang et al., 2017b), is not necessary with our method, since the efficiency of our estimator decreases relatively slowly as we decrease the trimming threshold.

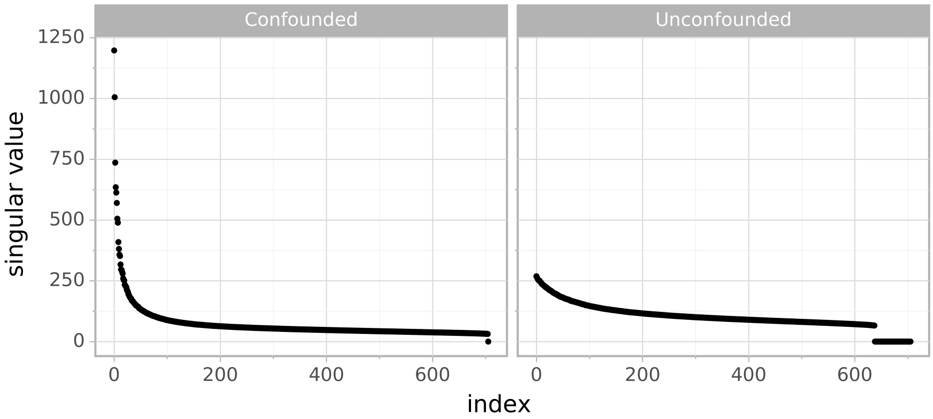


Figure 2: Left: Spiked singular values of the standardized gene expression matrix (see Section 5.2) indicate possible confounding. Right: Singular values after regressing out the $q = 65$ confounding proxies given in the dataset (thus labeled as "unconfounded"). The singular values in both plots are sorted decreasingly.

In what follows, we illustrate the empirical performance of the Doubly Debiased Lasso in practice. We compare the performance with the standard Debiased Lasso (Zhang and Zhang, 2014), even though it is not really a competitor for dealing with hidden confounding. Our goal is to illustrate and quantify the error and bias when using the naive and popular approach which ignores potential hidden confounding. We first investigate the performance of our method on simulated data for a range of data generating mechanisms and then investigate its behaviour on a gene expression dataset from the GTEx project (Lonsdale et al., 2013).

5.1 Simulations

In this section, we compare the Doubly Debiased Lasso with the standard Debiased Lasso in several different simulation settings for estimation of β_j and construction of the corresponding confidence intervals.

In order to make comparisons with the standard Debiased Lasso as fair as possible, we use the same procedure for constructing the standard Debiased Lasso, but with $\mathcal{Q} = \mathrm{I}_p$, $\mathcal{P}^{(j)} = \mathrm{I}_{p-1}$, whereas for the Doubly Debiased Lasso, $\mathcal{P}^{(j)}$, $\mathcal{Q}$ are taken to be median Trim transform matrices, unless specified otherwise. Finally, to investigate the usefulness of double debiasing, we additionally include the standard Debiased Lasso estimator with the same initial estimator $\widehat{\beta}^{init}$ as our proposed method, see Section 3.4. Therefore, this corresponds to the case where $\mathcal{Q}$ is the median Trim transform, whereas $\mathcal{P}^{(j)} = \mathrm{I}_{p-1}$.

We will compare the (scaled) bias and variance of the corresponding estimators. For a fixed index j, from the equation (11) we have

$$V^{-1/2}(\widehat{\beta}_j - \beta_j) = N(0,1) + B_\beta + B_b,$$

where the estimator variance V is defined in (25) and the bias terms B_β and B_b are given by

$$B_\beta = V^{-1/2}\frac{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}(\widehat{\beta}_{-j}^{init} - \beta_{-j})}{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j}, \qquad B_b = V^{-1/2}\frac{Z_j^\intercal(\mathcal{P}^{(j)})^2 X b}{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j}.$$

Larger estimator variance makes the confidence intervals wider. However, large bias makes the confidence intervals inaccurate. We quantify this with the scaled bias terms B_β, which is due to the error in estimation of β, and B_b, which is due to the perturbation b arising from the hidden confounding. Having small $|B_\beta|$ and $|B_b|$ is essential for having a correct coverage, since the construction of confidence intervals is based on the approximation $V^{-1/2}(\widehat{\beta}_j - \beta_j) \approx N(0,1)$. We investigate the validity of the confidence interval construction by measuring the coverage of the nominal 95% confidence interval. We present here a wide range of simulations settings and further simulations can be found in the Section 7.

Simulation parameters Unless specified otherwise, in all simulations we fix $q = 3$, $s = 5$ and $\beta = (1,1,1,1,1,0,\dots 0)^\intercal$ and we target the coefficient $\beta_1 = 1$. The rows of the unconfounded design matrix E are generated from $N(0, \Sigma_E)$ distribution, where $\Sigma_E = \mathrm{I}_p$, as a default. The matrix of confounding variables H, the additive error e and the coefficient matrices Ψ and ϕ all have i.i.d. $N(0,1)$ entries, unless stated otherwise. Each simulation is averaged over $5,000$ independent repetitions.

Varying dimensions n and p In this simulation setting we investigate how the performance of our estimator depends on the dimensionality of the problem. The results can be seen in Figure 3. In the first scenario, shown in the top row, we have $p = 500$ and n varying from 50 to $2,000$, thus covering both low-dimensional and high-dimensional cases. In the second scenario, shown in the bottom row, the sample size is fixed at $n = 500$ and the number of covariates p varies from 100 to $2,000$. We provide analogous simulations in Section 7, where both the random variables and the model parameters are generated from non-Gaussian distributions.

We see that the absolute bias term $|B_b|$ due to confounding is substantially smaller for Doubly Debiased Lasso compared to the standard Debiased Lasso, regardless of which initial estimator is used. This is because $\mathcal{P}^{(j)}$ additionally removes bias by shrinking large principal components of X_{-j}. This spectral transformation helps also to make the absolute bias term $|B_\beta|$ smaller for the Doubly Debiased Lasso compared to the Debiased Lasso, even when using the same initial estimator $\widehat{\beta}^{init}$. This comes however at the expense of slightly larger variance, but we can see that the decrease in bias reflects positively on the validity of the constructed confidence intervals. Their coverage is significantly more accurate for Doubly Debiased Lasso, over a large range of n and p.

There are two challenging regimes for estimation under confounding. Firstly, when the dimension p is much larger than the sample size n, the coverage can be lower than 95%, since in this regime it is difficult to estimate β accurately and thus the term $|B_\beta|$ is fairly large, even after the bias correction step. We see that the absolute bias $|B_\beta|$ grows with p, but it is much smaller for the Doubly Debiased Lasso which positively impacts the coverage. Secondly, in the regime where p is relatively small compared to n, $|B_b|$ begins to dominate and leads to undercoverage of confidence intervals. B_b is caused by the hidden confounding and does not disappear when $n \to \infty$, while keeping p constant. The simulation results agree with the asymptotic analysis of the bias term in the supplementary material of Guo et al. (2020), where the term $|B_b|$ vanishes as $\lambda_q(\Psi)$ increases, in addition to increasing the sample size n. In the regime considered in this simulation, $|B_b|$ can even grow, since the bias becomes increasingly large compared to the estimator's variance. However, it is important to note that even in these difficult regimes, Doubly Debiased Lasso performs significantly better than the standard Debiased Lasso (irrespective of the initial estimator) as it manages to additionally decrease the estimator's bias.

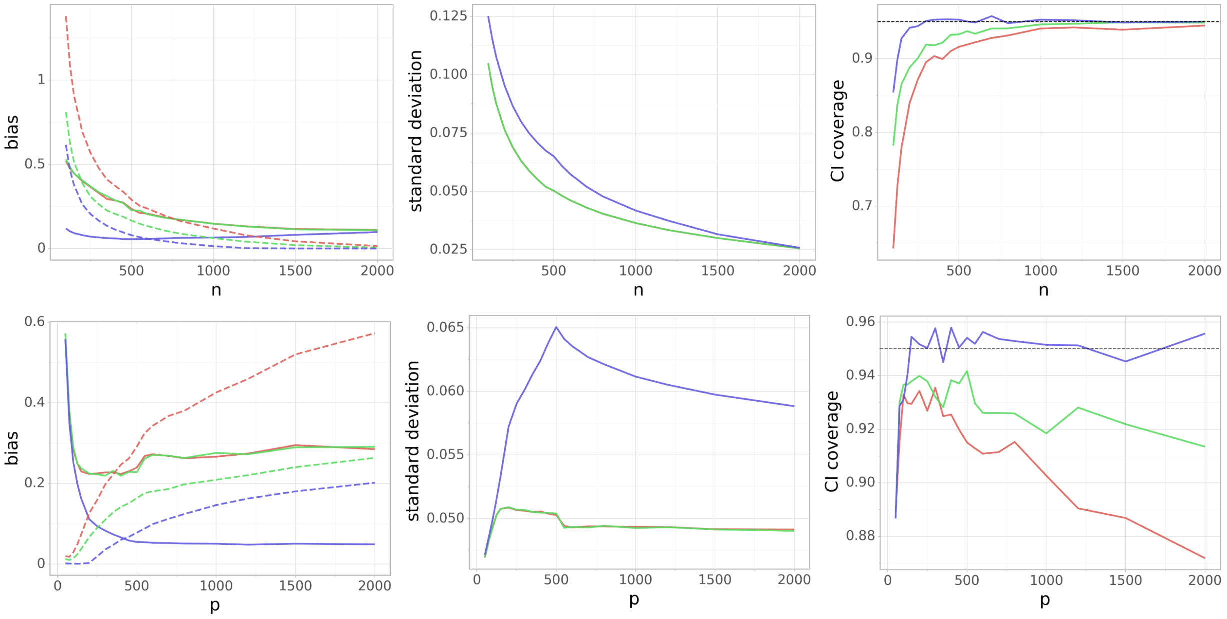


Figure 3: *(Varying dimensions)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of data points n (top row) and the number of covariates p (bottom row). On the left side, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. In the top row we fix $p = 500$, whereas in the bottom row we have $n = 500$. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

Toeplitz covariance structure for Σ_E Now we fix $n = 300, p = 1,000$, but we generate the covariance matrix Σ_E of the unconfounded part of the design matrix X to have Toeplitz covariance structure: $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$, where we vary κ across the interval $[0, 0.97]$. As we increase κ, the covariates $X_1, \ldots, X_5$ in the active set get more correlated, so it gets harder to distinguish their effects on the response and therefore to estimate β. Similarly, it gets as well harder to estimate γ in the regression of X_j on X_{-j}, since X_j can be explained well by many linear combinations of the other covariates that are correlated with X_j. In Figure 4 we can see that Doubly Debiased Lasso is much less affected by correlated covariates. The (scaled) absolute bias terms $|B_b|$ and $|B_\beta|$ are much larger for standard Debiased Lasso, which causes the coverage to worsen significantly for values of κ that are closer to 1.

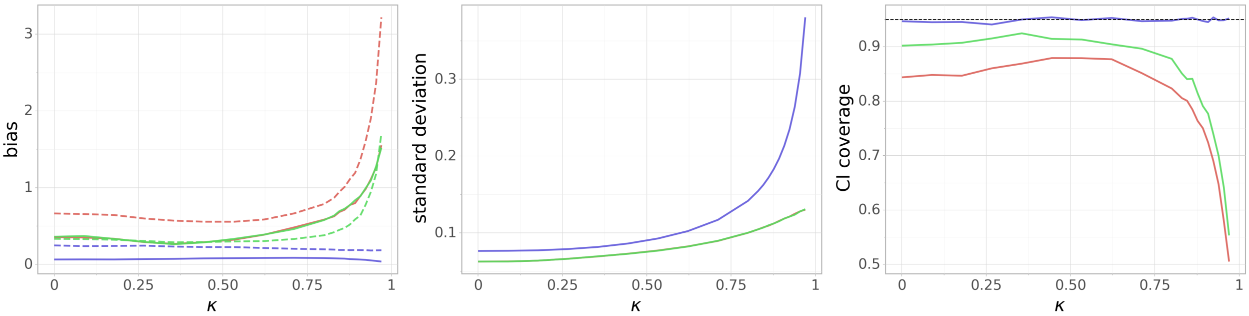


Figure 4: *(Toeplitz covariance for Σ_E)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the parameter κ of the Toeplitz covariance structure. $n = 300$ and $p = 1,000$ are fixed. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

Proportion of confounded covariates In order to investigate how the confounding denseness affects the performance of our method, we now again fix $n = 300$ and $p = 1,000$, but we change the proportion of covariates X_i that are affected by each confounding variable. We do this by setting to zero a desired proportion of entries in each row of the matrix $\Psi \in \mathbb{R}^{q\times p}$, which describes the effect of the confounding variables on each predictor. Its non-zero entries are still generated as $N(0, 1)$. We set once again $\Sigma_E = \mathrm{I}_p$ and we vary the proportion of nonzero entries of Ψ from 5% to 100%. The results can be seen in Figure 5. We can see that Doubly Debiased Lasso performs well even when only a very small number (5%) of the covariates are affected by the confounding variables, which agrees with our theoretical discussion for assumption **(A2)**. We can also see that the coverage of the standard Debiased Lasso is poor even for a small number of affected variables and it worsens as the confounding variables affect more and more covariates. The coverage improves to some extent when we use a better initial estimator, but is still worse than our proposed method.

In Section 7 we also show how the performance changes with the strength of confounding, by gradually decreasing the size of the entries of the loading matrix Ψ.

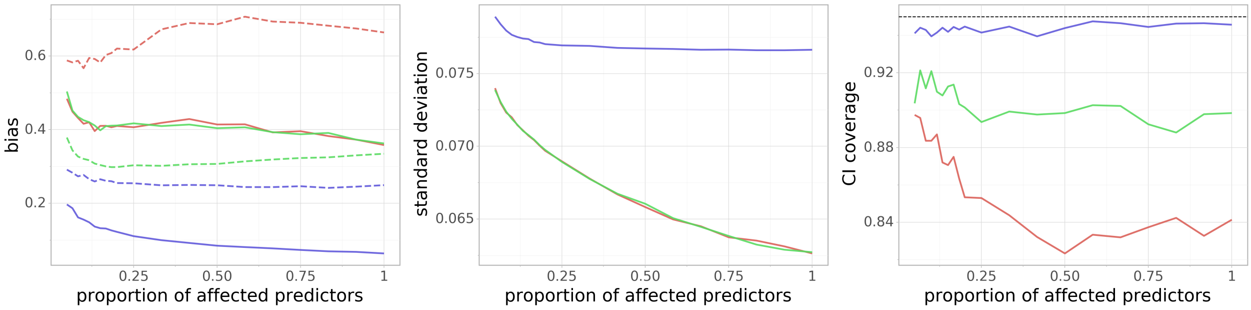


Figure 5: *(Proportion confounded)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on proportion of confounded covariates. $n = 300$ and $p = 1,000$ are fixed. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

Trimming level We investigate here the dependence of the performance on the choice of the trimming threshold for the Trim transform (14), parametrized by the proportion of singular values ρ_j which we shrink. The spectral transformation $\mathcal{Q}$ used for the initial estimator $\widehat{\beta}^{init}$ is fixed to be the default choice of Trim transform with median rule. We fix $n = 300$ and $p = 1,000$ and consider the same setup as in Figure 3. We take $\tau = \Lambda_{\lfloor \rho_j m \rfloor, \lfloor \rho_j m \rfloor}$ to be the ρ_j-quantile of the set of singular values of the design matrix X, where we vary ρ_j across the interval $[0, 0.9]$. When $\rho_j = 0$, τ is the maximal singular value, so there is no shrinkage and our estimator reduces to the standard Debiased Lasso (with the initial estimator $\widehat{\beta}^{init}$). The results are displayed in Figure 6. We can see that Doubly Debiased Lasso is quite insensitive to the trimming level, as long as the number of shrunken singular values is large enough compared to the number of confounding variables q. In the simulation $q = 3$ and the (scaled) absolute bias terms $|B_b|$ and $|B_\beta|$ are still small when $\rho_j \approx 0.02$, corresponding to shrinking 6 largest singular values. We see that the standard deviation decreases as ρ_j decreases, i.e. as the trimming level τ increases, which matches our efficiency analysis in Section 4.2.1. However, we see that the default choice $\tau = \Lambda_{\lfloor m/2 \rfloor, \lfloor m/2 \rfloor}$ has decent performance as well. In Section 7 we also explore whether the choice of spectral transformation significantly affects the performance, with a focus on the PCA adjustment, which maps first several singular values to 0, while keeping the others intact.

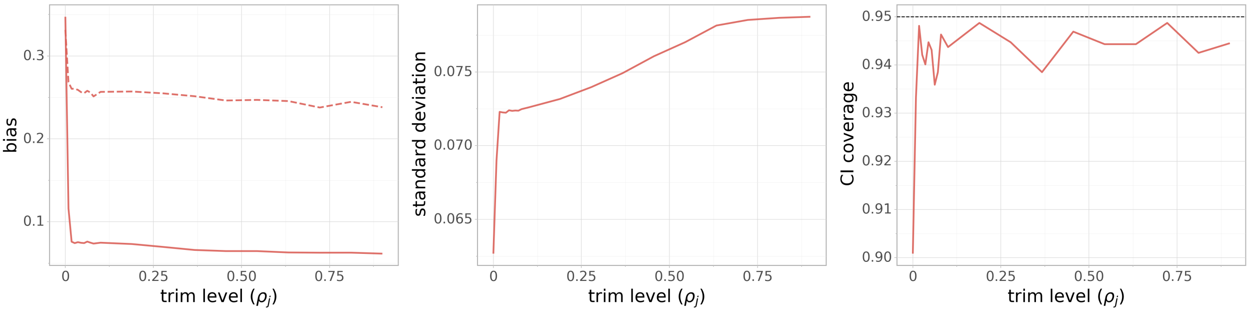


Figure 6: *(Trimming level)* Dependence of the (scaled) absolyte bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the trimming level ρ_j of the Trim transform (see Equation (14)). The sample size is fixed at $n = 300$ and the dimension at $p = 1,000$. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. The case $\rho_j = 0$ corresponds to Debiased Lasso with the spectral deconfounding initial estimator $\widehat{\beta}^{init}$, described in (16).

No confounding bias We consider now the same simulation setting as in Figure 3, where we fix $n = 500$ and vary p, but where in addition we remove the effect of the perturbation b that arises due to the confounding. We generate from the model (2), but then adjust for the confounding bias: $Y \leftarrow (Y - Xb)$, where b is the induced coefficient perturbation, as in Equation (3). In this way we still have a perturbed linear model, but where we have enforced $b = 0$ while keeping the same spiked covariance structure of X: $\Sigma_X = \Sigma_E + \Psi^\intercal\Psi$ as in (2). The results can be seen in the top row of Figure 7. We see that Doubly Debiased Lasso still has smaller absolute bias $|B_\beta|$, slightly higher variance and better coverage than the standard Debiased Lasso, even in absence of confounding. The bias term B_b equals 0, since we have put $b = 0$. We can even observe a decrease in estimation bias for large p, and thus an improvement in the confidence interval coverage. This is due to the fact that X has a spiked covariance structure and trimming the large singular values reduces the correlations between the predictors. This phenomenon is also illustrated in the additional simulations in the Section 7, where we set $q = 0$ and put E to have either Toeplitz or equicorrelation covariance structure with varying degree of spikiness (by varying the correlation parameters).

In the bottom row of Figure 7 we repeat the same simulation, but where we set $q = 0$ and take $\Sigma_X = \Sigma_E = I$ in order to investigate the performance of the method in the setting without confounding, but where the covariance matrix of the predictors is not spiked. We see that there is not much difference in the bias and only a slight increase in the variance of our estimator and thus also there is not much difference in the coverage of the confidence intervals. We conclude that our method can provide certain robustness against dense confounding: if there is such confounding, our proposed method is able to significantly reduce the bias caused by it; on the other hand, if there is no confounding, in comparison to the standard Debiased Lasso, our proposed method still has essentially as good performance, with a small increase in variance.

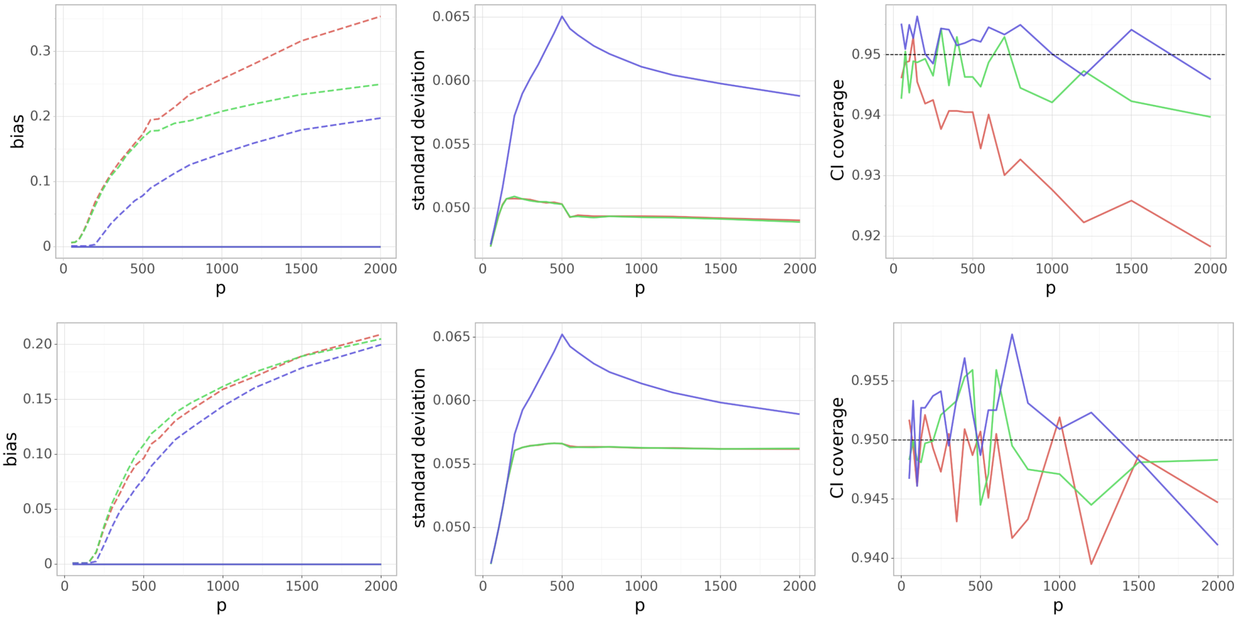


Figure 7: *(No confounding bias)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of covariates p, while keeping $n = 500$ fixed. In the plots on the left, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively, but $B_b = 0$ since we have enforced $b = 0$. Top row corresponds to the spiked covariance case $\Sigma_X = \Psi^T\Psi + I$, whereas for the bottom row we set $\Sigma_X = I$. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable V.

Measurement error We now generate from the measurement error model (4), which can be viewed as a special case of our model (2). The measurement error $W = \Psi^\intercal H$ is generated by $q = 3$ latent variables $H_{i,\cdot} \in \mathbb{R}^q$ for $1 \leqslant i \leqslant n$. We fix the number of data points to be $n = 500$ and vary the number of covariates p from 50 to $1,000$, as in Figure 3. The results are displayed in Figure 8, where we can see a similar pattern as before: Doubly Debiased Lasso decreases the bias at the expense of a slightly inflated variance, which in turn makes the inference much more accurate and the confidence intervals have significantly better coverage.

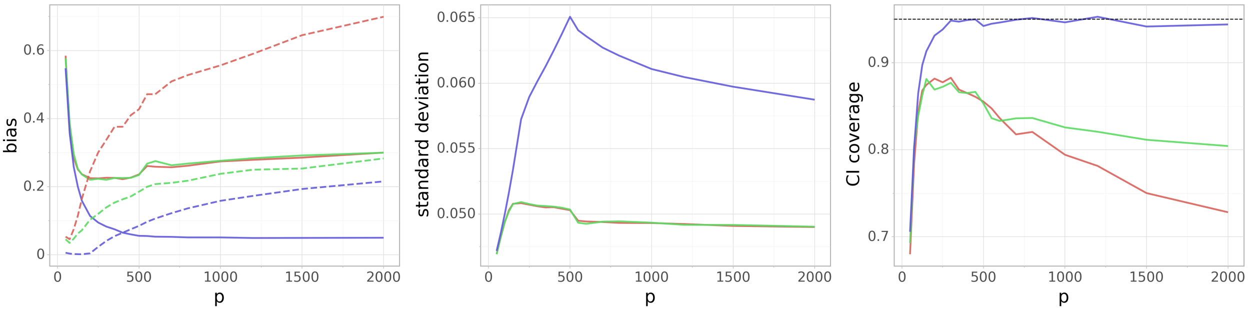


Figure 8: *(Measurement error)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of covariates p in the measurement error model (4). The sample size is fixed at $n = 500$. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

5.2 Real data

We investigate here the performance of Doubly Debiased Lasso on a genomic dataset. The data are obtained from the GTEx project (Lonsdale et al., 2013), where the gene expression has been measured postmortem on samples coming from various tissue types. For our purposes, we use fully processed and normalized gene expression data for the skeletal muscle tissue. The gene expression matrix X consists of measurements of expressions of $p = 12,646$ protein-coding genes for $n = 706$ individuals. Genomic datasets are particularly prone to confounding (Leek and Storey, 2007; Gagnon-Bartsch and Speed, 2012; Gerard and Stephens, 2020), and for our analysis we are provided with $q = 65$ proxies for hidden confounding, computed with genotyping principal components and PEER factors.

We investigate the associations between the expressions of different genes by regressing one target gene expression X_i on the expression of other genes X_{-i}. Since the expression of many genes is very correlated, researchers often use just $\sim 1,000$ carefully chosen landmark genes as representatives of the whole gene expression (Subramanian et al., 2017). We will use several such landmark genes as the responses in our analysis.

In Figure 9 we can see a comparison of 95%-confidence intervals that are obtained from Doubly Debiased Lasso and standard Debiased Lasso. For a fixed response landmark gene X_i, we choose 25 predictor genes X_j where $j \neq i$ such that their corresponding coefficients of the Lasso estimator for regressing X_i on X_{-i} are non-zero. The covariates are ordered according to decreasing absolute values of their estimated Lasso coefficients. We notice that the confidence intervals follow a similar pattern, but that the Doubly Debiased Lasso, besides removing bias due to confounding, is more conservative as the resulting confidence

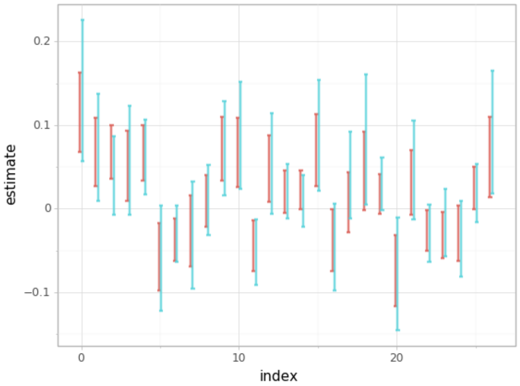


Figure 9: Comparison of 95% confidence intervals obtained by Doubly Debiased Lasso (blue) and Doubly Debiased Lasso (red) for regression of the expression of one target landmark gene on the other gene expressions.

intervals are wider.

This behavior becomes even more apparent in Figure 10, where we compare all p-values for a fixed response landmark gene. We see that Doubly Debiased Lasso is more conservative and it declares significantly less covariates significant than the standard Debiased Lasso. Even though the p-values of the two methods are correlated (see also Figure 12), we see that it can happen that one method declares a predictor significant, whereas the other does not.

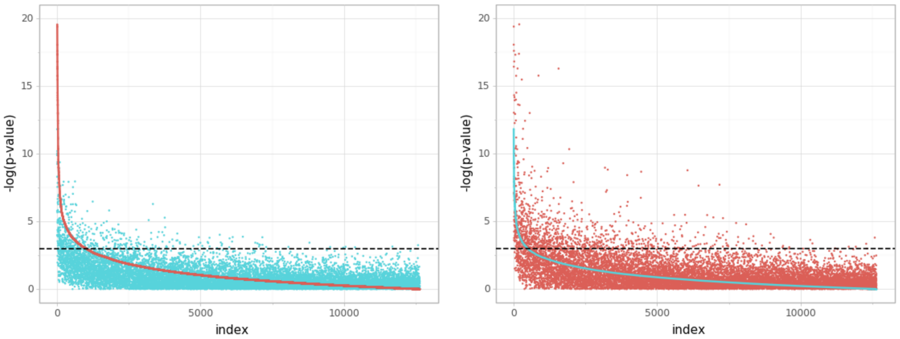


Figure 10: Comparison of p-values for two-sided test of the hypothesis $\beta_j = 0$, obtained by Doubly Debiased Lasso (red) and Doubly Debiased Lasso (blue) for regression of the expression of one target gene on the other gene expressions. The covariates are ordered by decreasing significance, either estimated by the Debiased Lasso (left) or by the Doubly Debiased Lasso (right). Black dotted line indicates the 5% significance level.

Robustness against hidden confounding We now adjust the data matrix X by regressing out the $q = 65$ provided hidden confounding proxies. By regressing out these covariates, we obtain an estimate of the unconfounded gene expression matrix $\tilde{X}$. We compare the estimates for the original gene expression matrix with the estimates obtained from the adjusted matrix.

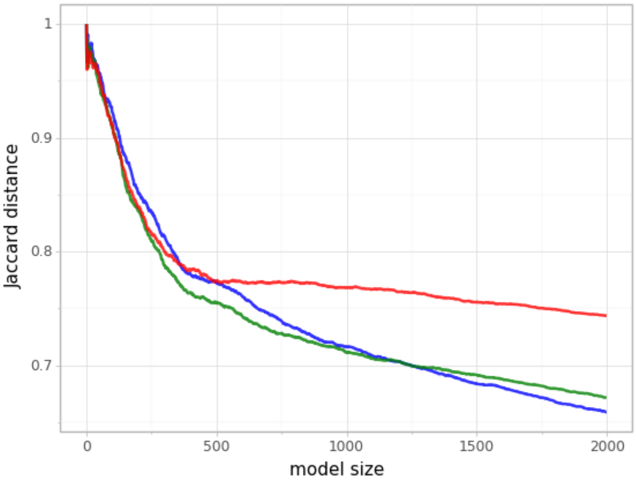


Figure 11: Comparison of the sets of the most significant covariates chosen based on the original expression matrix X and the deconfounded gene expression matrix $\tilde{X}$, for different cardinalities of the sets (model size). The set differences are measured by Jaccard distance. Red line represents the standard Debiased Lasso method, whereas the blue and green lines denote the Doubly Debiased Lasso that uses $\rho = 0.5$ and $\rho = 0.1$ for obtaining the trimming threshold, respectively; see Equation (14).

For a fixed response landmark gene expression X_i, we can determine significance of the predictor genes by considering the p-values. One can perform variable screening by considering the set of most significant genes. For Doubly Debiased Lasso and the standard Lasso we compare the sets of most significant variables determined from the gene expression matrix X and the deconfounded matrix $\tilde{X}$. The difference of the chosen sets is measured by the Jaccard distance. A larger Jaccard distance indicates a larger difference between the chosen sets. The results can be seen in Figure 11. The results are averaged over 10 different response landmark genes. We see that the Doubly Debiased Lasso gives more similar sets for the large model size, indicating that the analysis conclusions obtained by using Doubly Debiased Lasso are more robust in presence of confounding variables. However, for small model size we do not see large gains. In this case the sets produced by any method are quite different, i.e. the Jaccard distance is very large. This indicates that the problem of determining the most significant covariates is quite difficult, since X and $\tilde{X}$ differ a lot.

In Figure 12 we can see the relationship between the p-values obtained by Doubly Debiased Lasso and the standard Debiased Lasso for the original gene expression matrix X and the deconfounded matrix $\tilde{X}$. The p-values are aggregated over 10 response landmark genes and are computed for all possible predictor genes. We can see from the left plot that the Doubly Debiased Lasso is much more conservative for the confounded data. The cloud of points is skewed upwards showing that the standard Debiased Lasso declares many

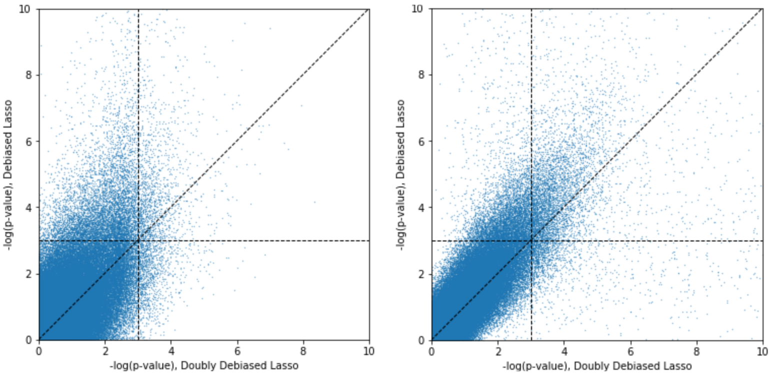


Figure 12: Comparison of p-values for two-sided test of the hypothesis $\beta_j = 0$, obtained by Doubly Debiased Lasso and standard Debiased Lasso for regression of the expression of one target gene on the other gene expressions. The points are aggregated over 10 landmark response genes. The p-values are either determined using the original gene expression matrix (left) or the matrix where we have regressed out the given $q = 65$ confounding proxies (right). Horizontal and vertical black dashed lines indicate the 5% significance level.

more covariates significant in presence of the hidden confounding. On the other hand, in the right plot we can see that the p-values obtained by the two methods are much more similar for the unconfounded data and the point cloud is significantly less skewed upwards. The remaining deviation from the $y = x$ line might be due to the remaining confounding, not accounted for by regressing out the given confounder proxies.

6 Discussion

We propose the Doubly Debiased Lasso estimator for hypothesis testing and confidence interval construction for single regression coefficients in high-dimensional settings with "dense" confounding. We present theoretical and empirical justifications and argue that our double debiasing leads to robustness against hidden confounding. In case of no confounding, the price to be paid is (typically) small, with a small increase in variance but even a decrease in estimation bias, in comparison to the standard Debiased Lasso (Zhang and Zhang, 2014); but there can be substantial gain when "dense" confounding is present.

It is ambitious to claim significance based on observational data. One always needs to make additional assumptions to guard against confounding. We believe that our robust Doubly Debiased Lasso is a clear improvement over the use of standard inferential high-dimensional techniques, yet it is simple and easy to implement, requiring two additional SVDs only, with no additional tuning parameters when using our default choice of trimming $\rho = \rho_j = 50\%$ of the singular values in Equations (14) and (15).

7 Additional Simulations

We present here some additional simulations to the ones presented in the Section 5.1. We use the same simulation setup where we further vary certain aspects of the data generating distribution or we vary the tuning parameters of the proposed Doubly Debiased Lasso method.

No confounding - Toeplitz and Equicorrelation covariance Here we explore further the scenarios where there is no confounding at all, i.e. $q = 0$, similarly as in the bottom part of Figure 7, but with different covariance structure of $X = E$. We fix $n = 300, p = 1,000$, and take the covariance matrix Σ_E to be either a Toeplitz matrix, with $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$ for $\kappa \in [0, 1)$, or we take it to be equicorrelation matrix where $(\Sigma_E)_{i,j} = \kappa \in [0, 1)$ when $i \neq j$ and 1 otherwise. In both cases, as the correlation parameter κ approaches 1, the singular values become more spiked and the predictors become more correlated. The results can be seen in Figure 13. We see that Doubly Debiased Lasso seems to have much smaller bias $|B_\beta|$ and thus better coverage even in the case when $q = 0$, because Trimming large singular values reduces the correlations between the predictors. This difference in bias and the coverage is even more clearly pronounced for the equicorrelation covariance structure, since for the Toeplitz covariance structure $\text{Cor}(X_i, X_j)$ decays as $|i - j|$ gets bigger, whereas for equicorrelation case it is constant and equal to κ.

Non-Gaussian distribution The Assumption (A3) in Section 3 requires that the noise term $\nu_{i,j} = E_{i,j} - E_{i,-j}^{\intercal}\gamma^E$ is is independent of $E_{i,-j}$. This condition will automatically hold if $E_{i,\cdot}$ is multivariate Gaussian or $E_{i,\cdot}$ has independent entries. We now test the robustness of Doubly Debiased Lasso method when this assumption is violated. In order to examine that, we repeat the simulation setting displayed in Figure 3, where $n = 500$ and p varies from 1 to $2,000$. We change the distribution as follows: Let $\mathbb{P}$ be some real distribution with zero mean and unit variance. The entries of the matrix of the confounders H are generated i.i.d. from $\mathbb{P}$. Furthermore, the unconfounded part of the predictors E is generated as $Z\Sigma_E^{1/2}$, where Z is a $n \times p$ matrix with i.i.d. entries coming from the distribution $\mathbb{P}$ and Σ_E is a Toeplitz matrix with $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$ for $\kappa = 0.7$. Finally, the noise variables e_i used for generating Y (see Equation 2) are also generated from $\mathbb{P}$. The results can be seen in Figure 14. We take $\mathbb{P}$ to be the following distributions: standardized chi-squared with 1 degree of freedom, standardized t-distribution with 5 degrees of freedom and standardized $\text{Bin}(16, 0.5)$. For comparisons of the performance, we also include $N(0, 1)$ distribution, but one needs to keep in mind that the obtained plot

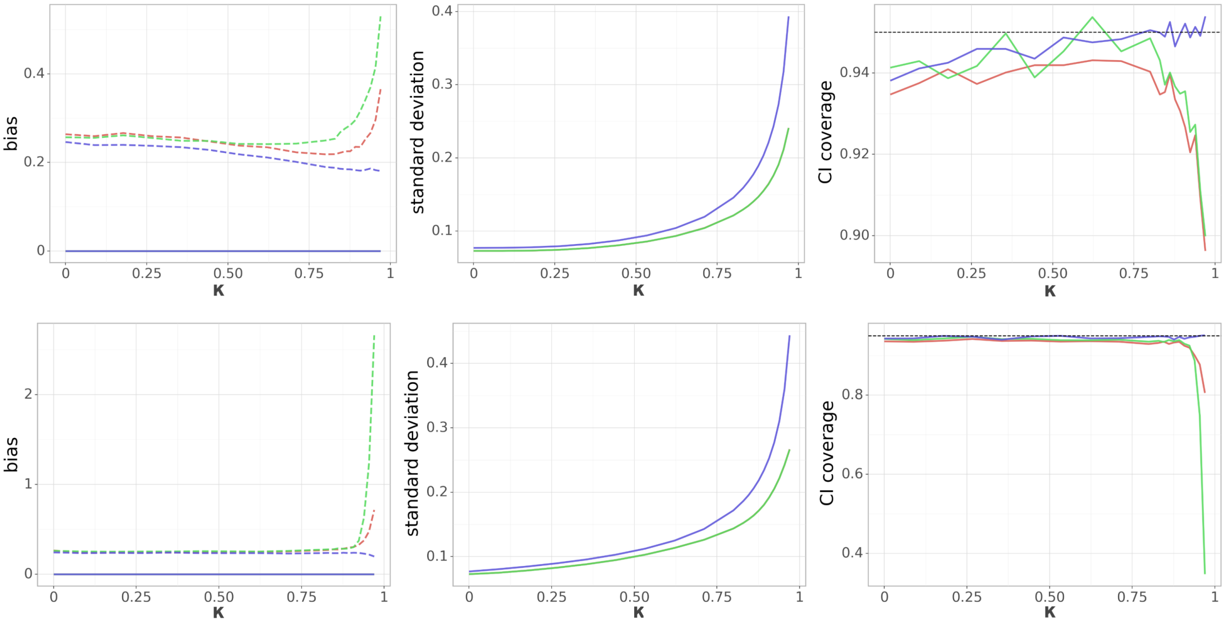
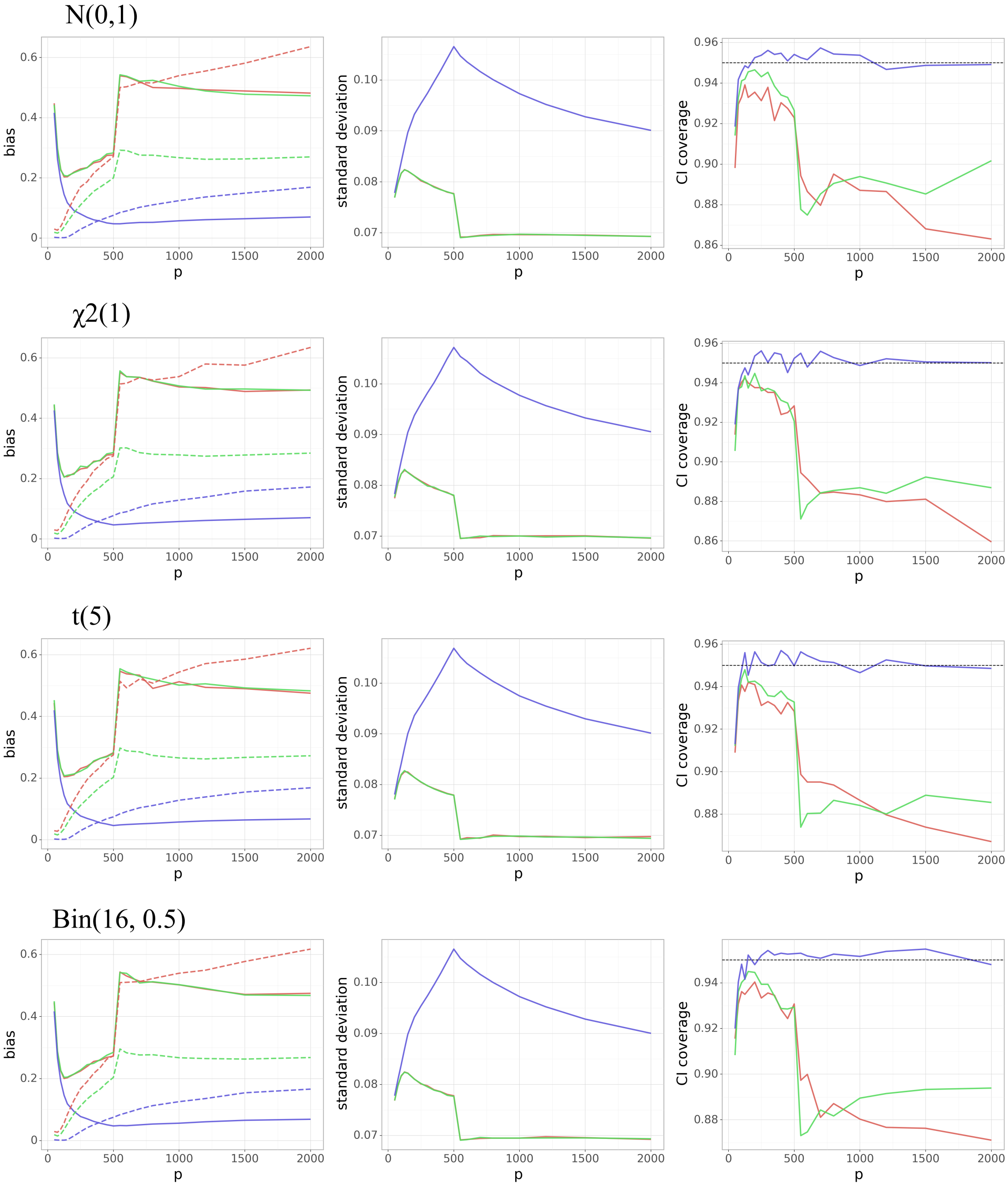


Figure 13: *(No confounding - Toeplitz and Equicorrelation covariance)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the correlation parameter κ, while keeping $p = 1,000, n = 300, q = 0$ fixed. In the plots on the left, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively, but $B_b = 0$ since we zero confounders $q = 0$. Top row corresponds to the Toeplitz covariance structure $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$, whereas for the bottom row we have equicorrelation covariance matrix where the off-diagonal elements equal κ. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable V.

differs from the one in Figure 3 because of different correlation structure of E. We can see that there is very little change in the performance of the proposed estimator, thus showing that Doubly Debiased Lasso can be used for a wide range of models.

Comparison to PCA adjustment Here we investigate how the choice of the spectral transformation can affect the performance of the Doubly Debiased Lasso estimator. We focus on the PCA adjustment which maps first $\hat{q}$ singular values to 0, for some tuning parameter $\hat{q}$, while keeping the remaining singular values unchanged. This transformation is used frequently in the literature because it arises by regressing out the top $\hat{q}$ principal components from every predictor.

We fix $n = 300, p = 1,000, q = 5$ and vary the parameter $\hat{q}$. We compare the estimator using the PCA adjustment for both $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ with the estimator using the Trim transform with the median rule for both $\mathcal{P}^{(j)}$ and $\mathcal{Q}$. Finally, we also consider the estimator using the Trim transform for $\mathcal{Q}$ and PCA adjustment for $\mathcal{P}^{(j)}$, in order to separate the effects of



N(0,1)

χ2(1)

t(5)

Bin(16, 0.5)

Figure 14: *(Non-Gaussian distribution)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of predictors p, while keeping $n = 500, q = 3$ fixed. On the left side, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. We change the distribution of H, E, e in (1) as described in the text. Each row in the plot corresponds to a different distribution $\mathbb{P}$. We set Σ_E to have Toeplitz structure with parameter $\kappa = 0.7$. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

changing the spectral transformation for the initial estimator $\widehat{\beta}^{init}$ and the overall estimator construction. The results can be seen in Figure 15.

We see that the performance is very sensitive to the choice of the tuning parameter $\hat{q}$. On one hand, if $\hat{q} < q$, we do not manage to remove enough of the confounding bias B_b, which has as a consequence that there is certain undercoverage of the confidence intervals. On the other hand, if $\hat{q} \leqslant q$, the bias B_b becomes very small, but the variance of our estimator increases slowly as $\hat{q}$ grows. Also, removing too many principal components when computing $\widehat{\beta}^{init}$ can remove too much signal, resulting in the higher bias B_β. Trim transform has an advantage that we do not need to estimate the number of latent confounders q from the data, which might be a quite difficult task. This is done by trimming many principal components, but not removing them completely. However, this can result in a small increase of the estimator variance compared to the PCA adjustment with the optimal tuning $\hat{q} = q$.

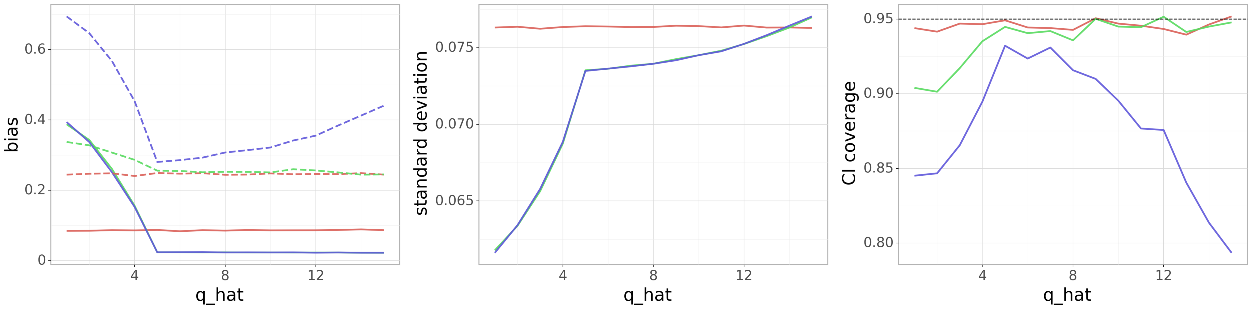


Figure 15: *(Comparison to PCA adjustment)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the correlation parameter κ, while keeping $p = 1,000, n = 300, q = 3$ fixed. In the left plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. We vary the parameter $\hat{q}$ of the PCA adjustment, which maps the first $\hat{q}$ to zero. Red color corresponds to the Doubly Debiased Lasso using Trim transform for both $\mathcal{P}^{(j)}$ and Q, blue color represents the Doubly Debiased Lasso using PCA adjustment for both $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ and green color corresponds to the Doubly Debiased Lasso estimator using the same default $\widehat{\beta}^{init}$ with $\mathcal{Q}$ being the median Trim transform, but uses PCA adjustment for $\mathcal{P}^{(j)}$. Note that the last two methods have almost indistinguishable V.

Weak confounding Here, we explore how the performance of our estimator depends on the strength of the confounding, i.e. how H affects X. In Figure 5, we have already explored how the performance of our method depends on the number of affected predictors by each confounder. Here we allow all predictors to be affected, but with decaying strength. This we achieve by generating the entries of the loading matrix Ψ as $\Psi_{ij} \sim N(0, 1/\sigma_i(j)^a)$, where for each of the q rows we take a random permutation $\sigma_i : \{1, \ldots, p\} \to \{1, \ldots, p\}$, and $a \geqslant 1$ is a tuning parameter describing the decay of the loading coefficients. The values

$n = 300, p = 1,000$ and $q = 3$ are kept fixed. The results can be seen in the Figure 16. We see that when a is close to 1 and the confounding is strong that our proposed estimator is much better that the standard Debiased Lasso estimator. On the other hand, when a is larger, meaning that the confounding gets much weaker, the difference in performance decreases, but Doubly Debiased Lasso still has smaller bias and thus better coverage.

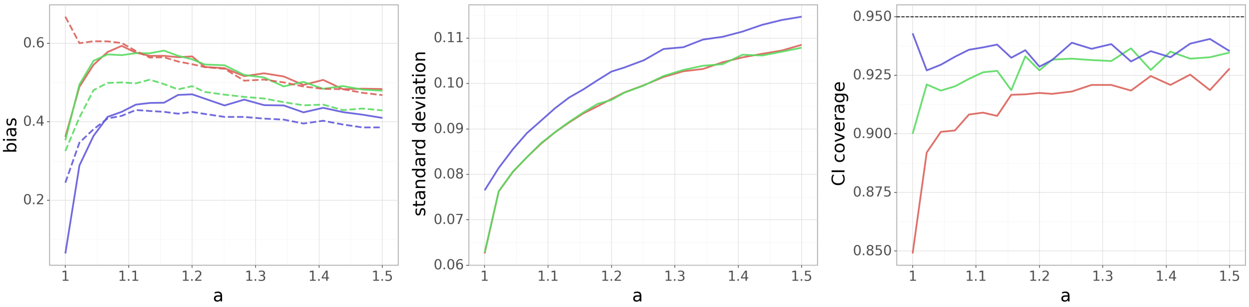


Figure 16: *(Weak confounding)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the loadings decay parameter a, while keeping $p = 1,000, n = 300, q = 3$ fixed. In the left plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable V.

Acknowledgements

We thank Yuansi Chen for providing the code to preprocess the raw data from the GTEx project. We also thank Matthias Löffler for his help and useful discussions about random matrix theory.

Paper

C

Identifying cancer pathway dysregulations using differential causal effects.

Jablonski, K. P., Pirkl, M., Ćevid, D., Bühlmann, P., Beerenwinkel, N.

Submitted to Bioinformatics.

Identifying cancer pathway dysregulations using differential causal effects

Kim Philipp Jablonski†, Martin Pirkl†, Domagoj Ćevid*, Peter Bühlmann*, Niko Beerenwinkel†

† DEPARTMENT OF BIOSYSTEMS SCIENCE AND ENGINEERING, ETH ZÜRICH,
BASEL, 4058, SWITZERLAND
*SEMINAR FÜR STATISTIK, ETH ZÜRICH, 8092 ZÜRICH, SWITZERLAND.

Abstract

Signaling pathways control cellular behavior. Dysregulated pathways, for example due to mutations that cause genes and proteins to be expressed abnormally, can lead to diseases, such as cancer. We introduce a novel computational approach, called Differential Causal Effects (*dce*), which compares normal to cancerous cells using the statistical framework of causality. The method allows to detect individual edges in a signaling pathway that are dysregulated in cancer cells, while accounting for confounding. Hence, artificial signals from, for example, batch effects have less influence on the result and *dce* has a higher chance to detect the biological signals. We show that *dce* outperforms competing methods on synthetic data sets and on CRISPR knockout screens. In an exploratory analysis on breast cancer data from TCGA, we recover known and discover new genes involved in breast cancer progression.

1 Introduction

The complexity of cancer makes finding reliable diagnosis and treatment options a difficult task. Decades of research made the intractable disease better understood. However, many challenges remain due to its high variability and context specificity, e.g., regarding tissue and cell type. Patients with common cancer types in early stages show promising survival rates, even though rare subtypes still show low survival rates due to different traits like a more aggressive disease progression (Hawkes, 2019; Miller et al., 2019; Troester and Swift-Scanlan, 2009).

It has been hypothesized that cancer diversity can at least in part be explained by heterogeneous mutational patterns. These patterns influence the activity of biological pathways at the cellular level (Khakabimamaghani et al., 2019; Hanahan and Weinberg, 2011). For example, signaling pathways consist of several genes, which regulate certain cell programs, such as growth or apoptosis. The programs are driven by the causal interaction between the genes, e.g., the up-regulation of one causes the up-regulation of another gene. The causal effect (CE) determines the strength of this causal interaction, e.g., by increasing the expression of gene X two-fold, the expression of its child Y increases four-fold. Thus, X has a causal effect on Y of 2 (Pearl, 2000). Understanding how these causal networks are perturbed in tumors is necessary for prioritizing drug targets, understanding inter-patient heterogeneity, and detecting driver mutations (Vogelstein et al., 2013).

Traditionally, perturbed pathways are detected by assessing whether differentially expressed genes are members of the respective pathway more often than expected by chance. More sophisticated methods measure whether genes belonging to a pathway are localized at certain positions of a rank-ordered set of differentially expressed genes (Subramanian et al., 2005). In such cases, a pathway is interpreted as a simple set of genes and all topological information concerning the functional interconnectivity of genes is ignored. It has been recognized that interactions among genes can have a significant effect on the computation of pathway enrichments. Some tools consider, for example, gene expression correlations to account for confounding effects and control the type I error while retaining good statistical power (Wu and Smyth, 2012). The underlying structure of gene interactions can thus be either estimated from the data (P. Spirtes, 2000; Sedgewick et al., 2016) used for the enrichment analysis, or obtained from existing databases. Canonical pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al., 1999) can then be incorporated as prior knowledge to guide the enrichment analysis using topological information of gene connectivity (Liu et al., 2019; Dutta et al., 2012; Tarca et al., 2009; Saez-Rodriguez et al., 2009).

While such enrichment methods go beyond treating pathways as plain gene sets and incorporate topological information of molecular interactions, they often only report a global pathway dysregulation score (Tarca et al., 2009). An exception is PARADIGM, which records an inferred activity for each entity in the pathway under consideration for a given patient sample (Vaske et al., 2010). It does, however, not model causal effects, but only quantifies whether there is some general association among the genes like correlation. Differential causal effects (DCEs) on biological pathways have already been investigated in a formal setting (Wang et al., 2018; He et al., 2019; Tian et al., 2016), where a DCE is modeled as the difference between CEs for the same edge under two conditions. These methods infer the gene network from observational data, which is a difficult task due to the combination of typically low sample size and noise of real data. An incorrect network can result in biased estimation of CEs and DCEs. Additionally, none of these methods make use of the DCEs to compute a pathway enrichment score.

Here, we separate the problem of estimating the causal network and the CEs by replacing the former with the addition of prior knowledge in the form of biological pathways readily available in public databases (Ogata et al., 1999; Nishimura, 2001; Whirl-Carrillo et al., 2012; Mi et al., 2021; Schaefer et al., 2009). We make use of the general concept of causal effects in order to define differential CEs. Specifically, we estimate the CE of gene X on gene Y in normal samples and cancer samples and define the DCE as their difference. In particular, we compare the causal effects between two conditions, such as a malignant tissue from a tumor and a healthy tissue, to detect differences in the gene interactions. We propose Differential Causal Effects (*dce*), a new method which computes the DCE for every edge (i.e., molecular interaction) of a pathway for two given conditions based on gene expression data (fig. 1).

This allows us to identify pathway perturbations at the individual edge level while controlling for confounding factors using the statistical framework of causality. By including the additional covariates constructed from the principal components of the design matrix, we also provide a methodological extension of our method to handle potential unobserved confounding that is 'dense', i.e., where the confounding variable affects many covariates. For example, batch effects from different experimental laboratories or cell cycle information are not necessarily known, but are accounted for automatically. Our approach allows for computing pathway enrichments in order to rank all networks in large pathway databases to identify cancer specific dysregulated pathways. In this manner, we can detect pathways which play a prominent role in tumorigenesis and pinpoint specific interactions in the pathway that make a large contribution to its dysregulation and the disease phenotype.

We show that *dce* can recover significant DCEs and outperforms competitors in simulations. In a validation on real data we apply *dce* to a public CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) data set to recover differential effects in the network. In an exploratory study, we apply *dce* to breast cancer samples and compare the DCEs among different cancer stages. We identify dysregulated edges common across stages as well as stage-specific edges.

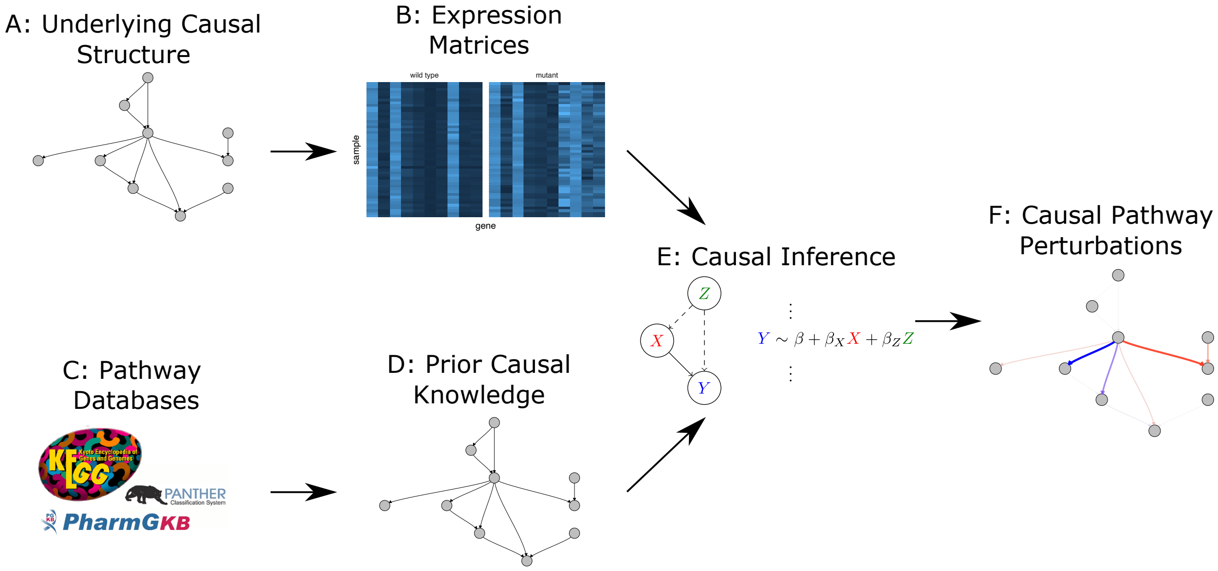

A: Underlying Causal Structure
B: Expression Matrices
wild type
mutant
sample
gene
C: Pathway Databases
KEGG
Kyoto Encyclopedia of Genes and Genomes
PANTHER
Classification System
PharmGKB
D: Prior Causal Knowledge
E: Causal Inference
Z
X
Y
$Y \sim \beta + \beta_X X + \beta_Z Z$
F: Causal Pathway Perturbations

Figure 1: A causal network of genetic interactions in a biological pathway (A) is responsible for the observed wild type expression levels in a cell (B: wild type). A disease can lead to perturbations of these pathways and in turn generate altered expression levels (B: mutant). Pathway databases such as KEGG (Ogata et al., 1999), PharmGKB (Whirl-Carrillo et al., 2012) and Panther (Mi et al., 2021) curate genetic interaction data (C) and thus provide networks of putative causal interactions (D). Given the observed wild type and disease expression levels as well as the causal structure, *dce* fits a generalized linear model (GLM) for each edge to estimate differential causal effects (E). In the given example, the differential causal effect from X on Y (solid edge) is estimated using the valid adjustment set $\{Z\}$ (as determined from the dashed edges). These differential causal effects correspond to causal perturbations (i.e., differential causal effects), e.g., an increase of causal effect strength from wild type to mutant is marked in blue. Negative differential causal effects are marked in red. The transparency of an edge corresponds to the magnitude of the associated effect) of the biological pathway caused by the disease and are important for diagnosis and treatment design (F).

2 Methods

In this section, we describe the Differential Causal Effects (*dce*) method. We briefly review the causality framework and then introduce the model and computation of DCEs, including under potential latent 'dense' confounding. We provide implementation details for obtaining both the estimates and their significance levels. Then, we describe the generating mechanism for synthetic data used throughout the paper. Finally, we explain the setup of our Perturb-seq validation as well as exploratory TCGA analysis.

Causality of biological pathways. First, we give a quick review of causality in the context of biological pathways. A gene pathway can be represented as a structural equation model (SEM) consisting of a directed acyclic graph (DAG) $\mathcal{G}$ with nodes $X = (X_i)_{i=1}^p$ describing the expression of genes, a set of directed edges $E = (E_i)_{i=1}^m$ representing the causal structure and the structural equations $(f_i)_{i=1}^n$ describing how each variable X_i is generated from its parents $X_{pa(i)}$ in $\mathcal{G}$, $X_i \leftarrow f_i(X_{pa(i)}, \epsilon_i)$, where $(\epsilon_i)_{i=1}^p$ are jointly independent noise variables. The causal interpretation of an edge between any two nodes is as follows: changing the expression of a parent X_i affects the expression of the child node X_j, which is propagated further to all descendants. The parental sets are given by the edge set E. Of particular interest are the interventional distributions for the SEM, in particular their expectations $\mathbb{E}[X_i \mid do(X_j = x)]$, which describe how the expected value of the variable X_i changes when we intervene and set the variable X_j to some fixed value x. We define the causal effect (CE) of a variable X_j on its descendant X_i as

$$CE[X_i \mid do(X_j = x)] = \frac{d}{dx}\mathbb{E}[X_i \mid do(X_j = x)]. \tag{1}$$

This derivative equals β_x if, by changing the value of X_j from x to $x + \Delta x$, for some small value Δx, the value of X_i changes on average by $\beta_x \cdot \Delta x$. In the literature, the CE is often also referred to as the total causal effect, because it quantifies the overall effect of an intervention at variable X_j on all of its descendants. We are interested in differential causal effects (DCE) defined as the differences between the causal effects of two conditions of interest, such as, e.g., two different cancer stages or healthy and cancerous samples.

Linearity of the conditional mean. We model the relationship between the mean of any gene expression X_i and its parents $X_{pa(i)}$ by a linear function:

$$X_i \leftarrow \gamma_0^{(i)} + \sum_{j \in pa(i)} \gamma_j^{(i)} X_j + \epsilon_i(X_{pa(i)}), \tag{2}$$

where, conditionally on $X_{pa(i)}$, the error term $\epsilon_i(X_{pa(i)})$ has mean zero and variance depending on $X_{pa(i)}$. A prime example is any generalized linear model (GLM) with identity link function. The coefficients $\gamma_j^{(i)}$ correspond to the direct causal effects, whereas the total causal effects (1) measure the aggregate effect over all directed paths from a certain variable X_j to X_i in $\mathcal{G}$.

Let us consider two arbitrary genes X_i and X_j in the pathway. Under the linearity assumption (2), the causal effect $CE[X_i \mid do(X_j = x)]$ does not depend on x. Furthermore, it can be computed as the coefficient β in the linear regression of X_i on X_j and an adjustment set $Z = (Z_k)_{k=1}^{|Z|}$,

$$X_i = \beta_0 + \beta X_j + \sum_{k=1}^{|Z|} \beta_k Z_k + \eta, \tag{3}$$

where β_0 denotes the intercept and η is random noise with mean zero (Goldszmidt and Pearl, 1992; Pearl, 1995). The adjustment set Z is a set of nodes in the pathway $\mathcal{G}$ which fulfills the Back-door criterion (Pearl, 2000). Hence, it holds that no element of Z is a descendant of X_j, and Z blocks every path between X_i and X_j that contains an edge with X_j as the child. For example, the parent set $X_{pa(j)}$ always fulfills the Back-door criterion and we always use it as adjustment set.

If the causal effects of the gene expression X_j on the gene expression X_i are respectively denoted as β^A and β^B under different conditions A and B, then the differential causal effect (DCE) δ is obtained as the difference

$$\delta = \beta^B - \beta^A. \tag{4}$$

Given a graph $\mathcal{G}$ describing a biological pathway and observations of the variables, we can compute all differential causal effects and identify interactions between any such two variables X_j and X_i that are different between the two conditions (fig. 1).

Testing for significance. We can compute the DCE δ for the edge $X_j \to X_i$ by fitting a joint model for both conditions, which also allows us to easily compute the significance of the estimates. Let I be an indicator random variable, which is equal to 1, if the observation comes from condition A, and 0, if it comes from condition B. The DCE δ can be computed from all samples jointly by fitting the following linear model

$$X_i = (\beta_0^A + (\beta_0^B - \beta_0^A)I) + (\beta^A + (\beta^B - \beta^A)I)X_j + \sum_{k=1}^{|Z|} \left(\beta_k^A + (\beta_k^B - \beta_k^A)I\right) Z_k + \eta \tag{5}$$

with interaction terms $I \cdot X_j$ and $I \cdot Z_i$. The differential causal effect $\delta = \beta^B - \beta^A$ can be estimated by using the coefficient estimate corresponding to the interaction term IX_j in (5).

Testing the significance of the estimated DCEs now corresponds to the well-known task of testing the significance of coefficient estimates in a linear model. However, some care is needed if the variances of the error terms $\epsilon_i(X_{pa(i)})$ in our structural equations (2) indeed depend on the values of the predictors $X_{pa(i)}$, i.e., if there is a certain mean-variance relationship for the gene expression levels, as has been described for RNA-seq data (Robinson and Smyth, 2007). In this case, the linear model (5) is heteroscedastic and the usual formulae for standard errors of the coefficient estimates, that result in t-tests for the significance, do not apply. We therefore use heteroscedasticity-consistent standard errors that yield asymptotically valid confidence intervals and p-values regardless of the dependence of the noise level on predictor values (Eicker, 1967; Huber et al., 1967; White, 1980).

Besides assessing significance of DCEs for single edges, we can also calculate a global p-value measuring the overall dysregulation of a given pathway $\mathcal{G}$: we combine the p-values corresponding to different differential causal effects $\delta = (\delta_i)_{i=1}^m$ by taking their harmonic mean (Good, 1958).

Adjusting for latent confounding. A fundamental assumption for most of causal inference methods is that there is no unobserved confounding, i.e., that there are no factors affecting both the cause and the effect (Leek et al., 2012; Gagnon-Bartsch et al., 2013). For example, batch effects due to varying laboratory conditions could act as such unobserved confounders. Presence of latent confounding can result in spurious correlations and false causal conclusions. Therefore, adjusting for potential latent confounding is crucial for making the method robust in applications to biological data (Ćevid et al., 2020a).

Some information about latent factors can often be obtained from the principal components of the data (Novembre and Stephens, 2008). This can be made rigorous under the linearity assumption (2) for our structural equation model $\mathcal{G}$, as follows. We assume that there are q latent variables $H_1, \ldots, H_q$ affecting our data. We extend the model (2) to include the latent confounding as follows:

$$X_i \leftarrow \gamma_0^{(i)} + \sum_{j \in pa(i)} \gamma_j^{(i)} X_j + \sum_{j=1}^{q} \delta_j^{(i)} H_j + \epsilon_i(X_{pa(i)}, H), \tag{6}$$

i.e., the latent confounders $H_1, \ldots, H_q$ are additional source nodes in the DAG $\mathcal{G}$ and affect genes in the pathway linearly, analogously to (2). Not every gene needs to be affected ($\delta_j^{(i)}$ could be zero), but the methodology works better the more genes are affected, see discussion below. By writing the structural equations (6) in matrix form, where we define $\Gamma_{ji}^0 = \gamma_0^{(i)}$, $\Gamma_{ji} = \gamma_j^{(i)}$, $\Delta_{ji} = \delta_j^{(i)}$ and $E(X,H)_{ji} = \epsilon_i(X_{pa(i)}, H)_j$, we obtain

$$X_{n\times p} \leftarrow \Gamma_{n\times p}^0 + X_{n\times p}\Gamma_{p\times p} + H_{n\times q}\Delta_{q\times p} + E(X,H)_{n\times p}, \tag{7}$$

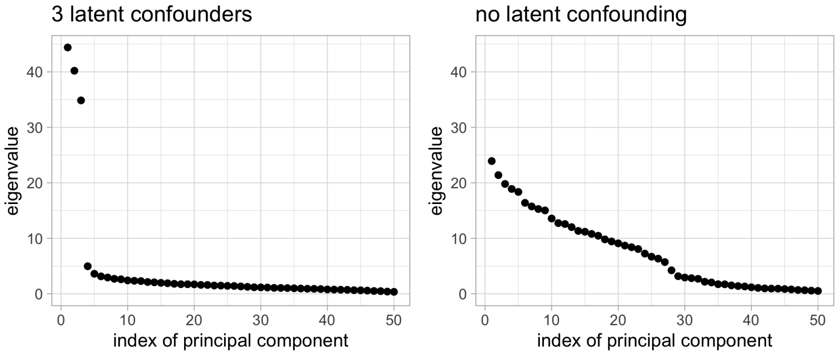


Figure 2: The scree plot (of synthetic data generated as described in the Methods section) shows that in presence of latent confounding as in (6), the first q principal components explain much more variability of the data, which we exploit for confounding adjustment.

which gives

$$X = \underbrace{\Gamma^0}_{\text{intercepts}} + H \underbrace{\Delta(I-\Gamma)^{-1}}_{\text{loadings} \in \mathbb{R}^{q\times p}} + \underbrace{E(X,H)(I-\Gamma)^{-1}}_{\text{random noise with mean} = 0} , \tag{8}$$

which is the standard linear factor model with heteroscedastic errors. From this representation, one can see that H can be determined from the principal components of X (fig. 2). The scree plot for a toy example visualizes the effect of latent variables having a global effect on the data. The first principal components are clearly separated from the rest, if latent factors are present (fig. 2, left). Therefore we obtain the confounding proxies $\hat{H}$ as the scores of the first $\hat{q}$ principal components of the design matrix combining the data from both conditions.

The confounding proxies $\hat{H}$ are then simply added to the adjustment set Z, see equations (3) and (5). In this way, the Back-door adjustment not only adjusts for the confounding variables observed in the DAG $\mathcal{G}$ as before, but also helps reducing the bias induced by latent confounding.

The deconfounding methodology relies on the assumption that every confounding variable affects many variables in the dataset, i.e., the confounding is dense (Guo et al., 2020). In this case, we have a lot of information about the latent factors in the data and the confounding proxies $\hat{H}$ capture the effect of the confounders H well. Furthermore, dense confounding assumption ensures that the scree plot, showing the singular values of the design matrix, has a spiked structure, as several latent factors can explain a relatively large proportion of the variance (fig. 2). This helps estimating the number $\hat{q}$ of the confounding proxies used. As a default choice, we use a permutation method that can be shown to work well under certain assumptions (Dobriban, 2017) and which compares the observed value of the variance explained by the principal components with its expected value over many random permutations of the values in each column of gene expression matrix X.

Algorithm and implementation in R. The presented methods are implemented in the R package *dce* which is freely available on Bioconductor. The function *dce::dce* takes as input the structure of a biological pathway, i.e., the adjacency matrix of a DAG, and two $n \times p$ matrices, with n samples and p genes, storing gene expression data for each of the two conditions respectively. As output, the function returns the estimated DCEs, as well as standard errors and two-sided p-values for the DCE at each edge in the pathway. The results can be easily transformed into a dataframe and plotted for further downstream analyses, together with the p-value measuring the overall pathway enrichment.

Generating synthetic data and benchmarking methods. We assess the behavior of *dce* and its competitors in a controlled setting by generating synthetic data with known DCEs (ground truth). We start by generating a random DAG $\mathcal{G}$. Without loss of generality, we assume the nodes of the DAG to be topologically ordered, i.e., node X_i can only be parent of node X_j, if $i < j$. This ensures that the network $\mathcal{G}$ is a DAG. In practice, we sample edges from a binomial distribution with probability $\hat{p}$ for the upper triangle of $\mathcal{G}$. We further sample the coefficients $\gamma_j^{(i)}$ for every edge as in (2) from a uniform distribution $\mathcal{U}\left(-\gamma_{\max}, \gamma_{\max}\right)$. We generate the data for network $\mathcal{G}$ in the following way. For a node X_i, we set the mean expression count

$$\mu_i = v - \vec{1} \cdot \left(\min_i v_i - \iota\right), \tag{9}$$

where each $X_j \sim \operatorname{Pois}\left(\mu_i\right)$ is a vector of counts, corresponding to gene expression values from experiments like RNA-Seq and depends on its parents by

$$v = \sum_{j \in pa(i)} \gamma_j^{(i)} X_j. \tag{10}$$

$\gamma_j^{(i)}$ represents the direct effect of X_j on X_i, $\iota > 0$ is a small shift, and $\vec{1}$ is a vector of ones. Subtracting the minimum ensures positive values of the mean for each data point. Then, a realization of X_i is drawn from the Poisson distribution $\operatorname{Pois}\left(\mu_i\right)$. We introduce negative binomial noise by drawing a realization of each source node in $\mathcal{G}$ from the negative binomial distribution $\operatorname{NB}\left(\mu, \theta\right)$ with a general mean μ and dispersion θ. We use this setup to control the variance across all nodes, which can blow up for descendants with larger means.

After sampling the data D_A for the nodes of network $\mathcal{G}$ under condition A, we resample a certain fraction of edge weights in order to generate new data D_B under condition B. For an edge weight β^A we sample the new edge weight from a uniform distribution

$$\beta^B \sim \mathcal{U}\left(\left(\beta^A - \delta_{\max}, \beta^A - \delta_{\min}\right) \cup \left(\beta^A + \delta_{\min}, \beta^A + \delta_{\max}\right)\right). \tag{11}$$

This ensures that the absolute difference between the two edge weights lies in $[\delta_{\min}, \delta_{\max}]$.

We also simulate latent variables. They are neither included in the data nor the network $\mathcal{G}$, but have (unknown) outgoing edges to all genes in the data set with non-zero effects. Hence, these latent variables have global effects on the data, e.g., emulating batch effects.

We compare *dce* to correlation (*cor*), partial correlation (*pcor*), the method Fast Gaussian Graphical Models (*fggm*) tailored to DCEs (Wang et al., 2016; He et al., 2019), Latent Differential Graphical Models (Tian et al., 2016), the pathway activity tool CARNIVAL (*car*) (Liu et al., 2019), a differential gene expression approach (*dge*), and random guessing. *cor* is provided by the R package `stats` (R Core Team, 2020). For *pcor* we use the general matrix inversion from the R package `MASS` (Venables and Ripley, 2002) to compute the precision matrix. *fggm* is based on partial correlation, but additionally tries to learn the network structure to adjust for confounding effects. We use the R code provided by the authors (He et al., 2019) to run *fggm*. *ldgm* is also based on partial correlation, but directly computes the differential network instead of two networks for the two data sets. We use the Matlab implementation of *ldgm* for the estimate of the latent correlation matrices by a transformation of Kendall's τ. We also add a permutation test to compute significance or assume Gaussian coefficients, and evaluate only differences corresponding to an edge in the graph. The parameter for *ldgm* is set according to the example at https://github.com/ma-compbio/LDGM/blob/master/Stand_alone_example_by_LDGM/LDGM/LDGM.m. For both *fggm* and *ldgm* we transform the gene expression count matrix by $log(\cdot + 1)$. We use the function `runInverseCarnival` from the R Bioconductor package `CARNIVAL` (Gjerga and Trairatphisan, 2021) to compute normalized edge weights $e \in [0, 100]$, which we normalize to p-values by $1 - \frac{e}{100}$. We use differential gene expression from `edgeR` (Robinson and Smyth, 2007) as input to CARNIVAL. The same differential expression result is used for *dge*. We compute the DCE for the edge between two genes x and y as the difference of the log foldchanges of both genes. We compute the corresponding p-value for the same edge with the minimum of the p-values for both genes x and y. We provide *pcor* with the same adjustment set of confounding variables as *dce*. We run all methods on simulated data for various modeling parameters. The default parameters are a network $\mathcal{G}$ of 100 genes, 200 samples for both sample conditions, an absolute magnitude in effect differences between the two conditions of 1, mean of 100 negative binomial distributed counts with a dispersion of 1 for the source genes in the network $\mathcal{G}$ (no parents), a true positive rate of 50% (edges which have different effects between the two conditions), and library size factors for each sample in the interval $[1, 10]$. The library size factor accounts for different sequencing depth among the samples, i.e., for one sample including more reads because more RNA was available even though the gene expression was the same as in samples with less RNA. We account for different library sizes over all samples by computing transcripts per million (tpm).

Overall we simulate a full data set of $10,000$ genes including the genes in the network $\mathcal{G}$ to allow for the realistic estimation of the library size. As a performance measure we use the area under the receiver operating characteristic (ROC-AUC). We count the number of true/false positive and false negative DCEs based on the edges in the ground truth network and the significant p-values for different significance levels. Based on these true/false positives we can compute the ROC curve and its AUC. For both correlation methods we use a permutation test to compute empirical p-values.

LDGM's runtime was too high for more than ten genes to use a permutation test to compute p-values. For more genes we assumed a Gaussian distribution to compute p-values.

Validation using Perturb-seq. Perturb-seq, a CRISPR-Cas9-based gene knockout method, can be used to inhibit the expression of multiple target genes on a single-cell level (Qi et al., 2013; Adamson et al., 2016). The data set we analyze is a CRISPR knockout screen with global gene expression profiles as the read-out. We can use the known knockout information of these experiments as ground-truth information for a performance evaluation of our method. In (Adamson et al., 2016), this approach was used to systematically analyze the response of an integrated endoplasmic reticulum (ER) stress response pathway to the combinatorial knockout of the three transmembrane sensor proteins IRE1α, ATF6, and PERK. Each considered combinatorial knockout (ATF6, ATF6+EIF2AK3, ATF6+EIF2AK3+ERN1, ATF6+ERN1, EIF2AK3, EIF2AK3+ERN1, ERN1) was treated either with a DMSO control, tunicamycin, or thapsigargin.

We download the raw gene expression count data from NCBI GEO (accession: GSE90546). The repository provided us with a mapping of guide and cell barcodes, and gene expression counts for all cells. We used this information to identify gene knock-outs for each cell. to create a gene expression count matrix of the individual cells labeled by their corresponding knockouts.

We download all pathway networks from KEGG and retain those which contain at least one of the three transmembrane sensor proteins. This yields in the pathways *hsa04137*, *hsa04140*, *hsa04141*, *hsa04210*, *hsa04932*, *hsa05010*, *hsa05016*, *hsa05017*, *hsa05160*, *hsa05162*, *hsa05168*.

For each combination of the three treatments, seven (combinatorial) knockouts and 11 pathways, we compute DCEs if the respective knocked-out gene is contained in the respective pathway. In total, this yields 128 conditions for each of which we run our method.

We compare the performance of *dce* to both *cor* (correlation) and *pcor* (partial correlation). For the two correlation methods, we estimate the significance of whether a difference in correlation is different from zero using a permutation test. The performance of each method is evaluated using the area-under-curve (AUC) metric for the receiver-operating-characteristic (ROC) curve. The false and true positive rates for the ROC curve are computed from the p-value per edge as in the synthetic benchmark.

Exploratory analysis with TCGA data. We retrieve gene expression matrices from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). These matrices have samples as columns and genes as rows, and are from the data category Transcriptome Profiling, data type Gene Expression Quantification, experimental strategy RNA-Seq and workflow type HTSeq-Counts. Pathway structures in the form of adjacency matrices are obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al., 1999).

Unlike the Perturb-seq dataset, data obtained from TCGA is observational instead of interventional. We do thus not have any ground truth information and perform an exploratory analysis. For a given cancer type, the associated samples are first grouped into normal and tumor samples. The tumor samples are subsequently stratified according to their stage. The clinical data needed to stratify the samples is readily available on TCGA as metadata for each gene expression matrix. In particular, we download all normal and tumor gene expression samples from TCGA for breast cancer (*TCGA-BRCA*) and selected all stages with a sufficient number of samples (stage I: 202 samples, stage II: 697 samples, stage III: 276 samples; normal: 113 samples). We use the breast cancer pathway (*hsa05224*) from KEGG which contains 147 nodes and 509 edges. We then compute DCEs between the normal condition and each of the three stages of the tumor condition, respectively.

3 Results

In this section, we first show the performance of *dce* and its competitors on simulated data and a CRISPR data set. Then, we use *dce* for an exploratory analysis of breast cancer data from TCGA and show the progression of pathway dysregulation over different cancer stages.

3.1 Simulation study

Pathway databases contain networks of different sizes. We first investigate the influence of network size on the ability of each method to recover ground truth differential causal effects. *dce* achieves the highest accuracy for all three network sizes considered ($50, 100, 150$ genes). Methods which do not account for confounding variables perform similar to random guessing for large networks (fig. 3a). However, *dce* also outperforms *pcor* with an AUC of 0.62 versus 0.56.

Overall, *ldgm*'s performance decreased a lot when using the Gaussian test to compute p-values. A closer looked showed that *ldgm* computed very small effect sizes, which lead to large p-values for dysregulated edges. While *car* seems very accurate in detecting true positives, we assume performance was less than random guessing due to the high false positve rate.

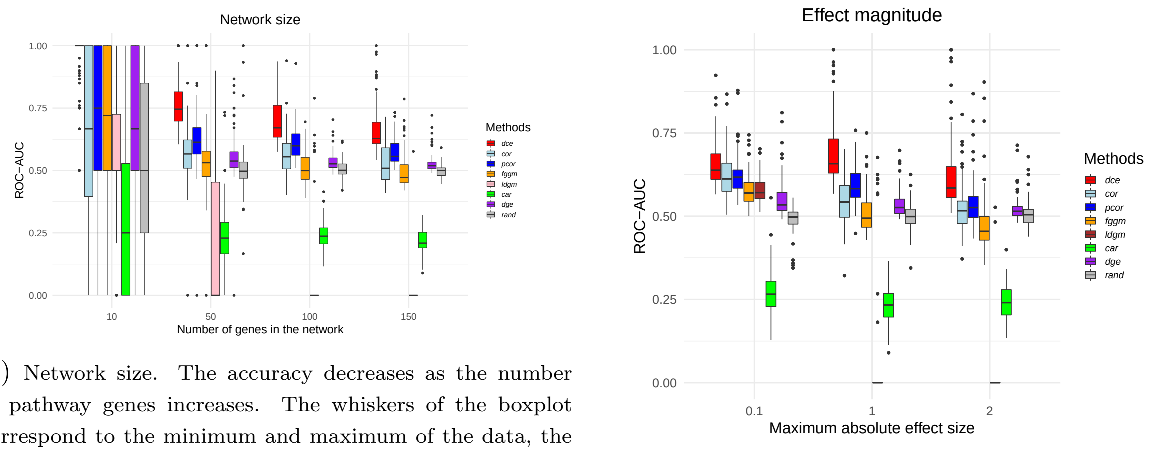
Second, we assess how the magnitude of differential causal effects affects the identification of significant differences. We sample the magnitude from the set $\{0.1, 1, 2\}$. For example, for a magnitude of 1 the edge weights between the network of the wild type samples and the disease samples differ by at most 1. *dce* has difficulty estimating large differences as well as very small differences. However, it still significantly outperforms all other methods, which again show similar performance to random guessing for large effects (fig. 3b).

In additional simulations, *dce* shows increasing accuracy for decreasing dispersion and increasing number of samples (figs. 6 and 7) as is expected due to decreasing noise. We found constant accuracy of *dce* over varying ranges of library size (fig. 8). Different prevalence of positive edges has little effect on the accuracy of *dce* (fig. 9). *dce* with latent variable adjustment performs similarly to *dce* without latent variable integration if we do not simulate any latent variables. But *dce* significantly outperforms *dce* without latent variable integration for five and ten latent variables influencing the data set (fig. 10). This is because without latent confounding adjustment one has a large number of false positives due to the confounding bias (fig. 11).

dce relies heavily on the given network $\mathcal{G}$. Hence, we investigate how well *dce* performs if $\mathcal{G}$ contains false edges or is missing true edges. We find that *dce* is robust to additional false edges in the network, but starts breaking down if true edges are missing in larger fractions (fig. 12).

3.2 Validation experiments using CRISPR knockout data

To benchmark our method using real-life data generated by Perturb-seq (Adamson et al., 2016), we ask whether we can recover the CRISPR knockout from single-cell RNA-seq data using pathways from KEGG which contain the knocked-out genes. Hence we assume that these pathways capture the causal gene interactions governing the response of the cell to the experimental intervention. As seen in the synthetic benchmark, slight deviations of the observed network from the true underlying network have no major impact on the performance of our method (fig. 12). By interpreting a CRISPR knockout as an intervention of the causal pathway, we define the positive class to consist of all edges



(a) Network size. The accuracy decreases as the number of pathway genes increases. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

(b) Effect magnitude. Large causal effects reduce the accuracy of DCE.

Figure 3: Performance benchmark. *dce* is compared to several competitors for varying network size (a) and effect magnitude (b) over 100 synthetic data sets each. *dce* achieves the highest accuracy, which decreases for large networks $\mathcal{G}$ and very large or small differential effects. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

adjacent to a knocked-out gene, and the negative class as all other genes. Consequently, a true positive occurs when an edge adjacent to a CRISPR knocked-out gene is (significantly) associated to a non-zero DCE.

Figure 4a shows an example of this procedure for one of the conditions described above. The CRISPR knockout gene is highlighted in red and a positive DCE of ~1.3 can be observed on the edge connecting ATF6 and DDIT3. This can be seen in more detail in fig. 4b. As this edge is adjacent to the knocked out gene ATF6, it is classified as a true positive for an effect size threshold of $|0.5|$. Following an analogous argument, the edge from EIF2AK3 to EIF2S1 is classified as a false positive.

We find that *dce* is significantly better (Wilcoxon signed-rank test (Wilcoxon, 1992) p-value $\leqslant 10^{-5}$) at recovering the knockout effects with a median ROC-AUC of 0.63 compared to 0.51 for *cor* and 0.53 for *pcor* (fig. 4c). To better understand the variability of the performance measure, we also investigate how performance varies when stratified by treatment and knockout gene (fig. 13). For example, for the knockout gene ATF6 the ROC-AUC of *dce* decreases from 0.89 for treatment 1 to 0.67 for treatment 2. This can be explained by the higher variability of the gene expression counts under treatment 2, as the p-value estimation becomes less stable. This pattern can also be observed for other performance shifts between treatments. We note that *cor* outperforms *dce* for the knockout of ATF6 in treatment 2, as the permutation test is able to better account for the variance of the expression data in this case. In all other cases, *dce* is either better or roughly as good as the competing methods. We conclude that *dce* is able to better recover the dysregulations of single as well as combinatorial knockouts when compared to methods based on correlations.

3.3 Exploratory analysis of TCGA data

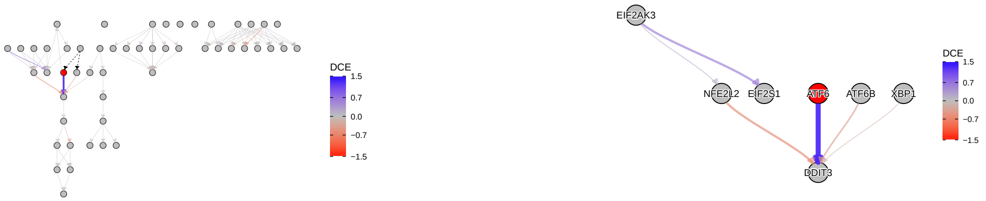
To demonstrate the ability of our method to recover known cancer-related pathway dysregulations as well as to discover new genes of potential biological and clinical relevance, we compute DCEs using breast cancer gene expression data from TCGA on the breast cancer pathway obtained from KEGG. The results for each stage are then visualized on the pathway structure (figs. 5a to 5c). The raw DCE values were transformed to a symmetric logarithm for greater visibility with the following formula

$$\text{symlog}(DCE) = \begin{cases} \text{log10}(DCE) + 1 & \text{if DCE} > 1 \\ -\text{log10}(-DCE) - 1 & \text{if DCE} < -1 \\ DCE & \text{otherwise} \end{cases} . \tag{12}$$

Roughly 40% of all investigated interactions (614 out of 1527) show no difference in causal effects ($|DCE| < 1$ and p-value > 0.05) between normal and stage condition for all stages. We will now discuss the cases with large DCE sizes or significant p-values (fig. 5d). In the following, we will discuss cases with large effect sizes and significant p-values.

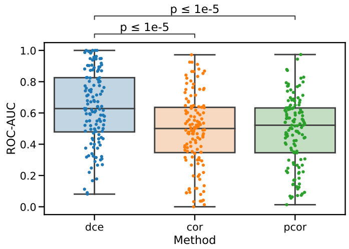
Throughout all stages, interactions between the WNT (Wingless/Int1) and FZD (Frizzled) protein complexes exhibit significant, non-zero DCEs indicating a strong dysregulation of the breast cancer pathway. Most notably, we observe a strongly significant dysregulation of WNT11→FZD1, WNT11→FZD3 and WNT11→FZD7 in stage II (p-value $< 1\text{e}{-20}$), as well as of WNT11→FZD7 in stages I and II. Additionally, the interaction between WNT8A and FZD4 features a strongly positive DCE of $\sim$2000 in all three stages. These observations are expected, because the interactions between the WNT and FZD protein complexes have been implicated in disease formation in general (Dijksterhuis et al., 2015; Chien et al., 2009; Schulte, 2010) and in breast cancer in particular (Yin et al., 2020; Koval and Katanaev, 2018).

Interactions between the FGF (Fibroblast Growth Factor) and FGFR (Fibroblast Growth Factor Receptor) protein complexes show strong negative effect sizes in all three stages (DCE < -100 for most members of these complexes). In particular, the FGF6→FGFR1 link features negative DCEs of -1279, -665, -1961, while the FGF8→FGFR1 link features negative DCEs of -402, -336, -285, in the stages I, II, III respectively. This pair has already been recognized as a promising therapeutic target for breast cancer treatment (Santolla and Maggiolini, 2020).



(a) Protein processing in the endoplasmic reticulum pathway for Homo Sapiens from KEGG (ID *hsa04141*). Each node corresponds to a gene and each edge to an interaction between two genes. Each edge is colored according to the effects size of DCEs computed for the experimental data for knocking out ATF6 and using DMSO control. The width of an edge corresponds to its absolute DCE (the wider an edge, the larger the absolute DCE). Black dashed edges are drawn when one of the two connected nodes has zero coverage (and thus no DCE can be estimated). The gene knocked out in the CRISPR experiment is highlighted in red.

(b) Zoomed-in version of fig. 4a with focus on the genes ATF6, ATF6B, NFE2L2, XBP1, DDIT3, EIF2AK3, EIF2S1. These genes constitute the neighborhood of the knocked-out gene ATF6 and illustrate the edge classification scheme used in the performance evaluation. Assume an effect size threshold of $|0.5|$. The edge ATF6→DDIT3 has a DCE of ~ 1.3 and is adjacent to the knocked-out gene. Consequently, it is classified as a true positive. Both the edge EIF2AK3→EIF2S1 and NFE2L2→DDIT3 have a DCE whose absolute value is larger than 0.5 and are not adjacent to the knocked-out gene. They are thus classified as false positives. All remaining edges are classified as true negatives.



(c) Summary of the performance of the *dce*, *cor* and *pcor* methods in the form of ROC-AUCs for the recovery of the knocked-out genes in all considered pathways. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median. Additionally, each data point is indicated with a dot whose x position has been randomly shifted to improve visibility. The method *dce* shows the best performance with a ROC-AUC of 0.63 (standard deviation (std): 0.23) compared to 0.51 (std: 0.23) for *cor* and 0.53 (std: 0.22) for *pcor*. The significance of the difference between the boxplots has been estimated using the Wilcoxon signed-rank test (Wilcoxon, 1992).

Figure 4: Overview of the CRISPR benchmark.

We also find the interaction between EGFR (Epidermal Growth Factor Receptor) and PIK3CA (Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha) to be significantly (p-values $< 1\mathrm{e}{-14}$) dysregulated with a small negative DCE of approximately -0.2 in stages I and II but not III. EGFR→PIK3CB shows similar behavior for stage II with a DCE of -0.12 and a p-value $< 1\mathrm{e}{-15}$. While the small effect size suggests that there is only a small dysregulation of these interactions, the dysregulation of EGFR together with PIK3CA mutations have been recognized as independent prognostic factors in triple negative breast cancers (Jacot et al., 2015).

The interaction between DLL3 (Delta Like Canonical Notch Ligand 3) and NOTCH4 (Notch Receptor 4) features a significant DCE of $\sim$140 with p-values $< 1\mathrm{e}{-6}$ in all three stages. The Notch signaling pathway has been shown to play an important role in Pancreatic ductal adenocarcinoma tumor cells, but has not been implicated in breast cancer (Song and Zhang, 2018). Our finding suggests that stromal cells located in the breast may play an important role for disease progression throughout all stages.

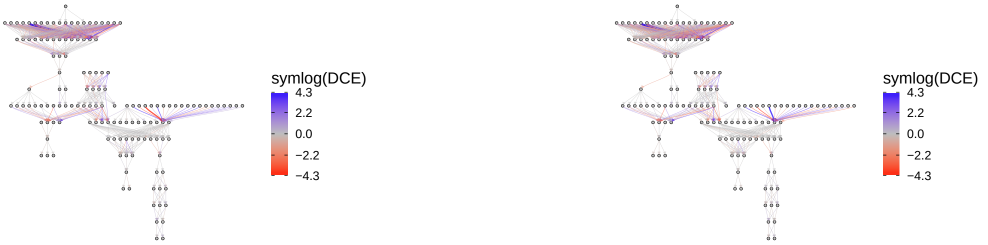
For the interaction between TCF7L2 (Transcription Factor 7 Like 2) and CCND1 (Cyclin D1) we observe a significant negative DCE of -11.9 with a p-value of $< 1\mathrm{e}{-6}$ in stage III. The role of TCF7L2, which participates in the Wnt/β-catenin signaling pathway and is important for cell development and growth regulation, has already been discussed in the context of breast cancer (Connor et al., 2012). However, its interaction with CCND1 has, to the best of our knowledge, not been investigated in the literature. Due to the down-regulation in the diseased condition for stage III, we suggest that an improved understanding of the underlying biological reasons might provide insights into the late-stage behavior of breast cancer.

Overall, we are able to recover both interactions which are known to be dysregulated in breast cancer as well as novel ones. The former indicates that the prioritization of interactions given by *dce* is in accordance with current literature. The latter suggests that *dce* is also able find dysregulated interactions which up to now have only been recognized for other diseases but may play an important role for breast cancer.

4 Discussion

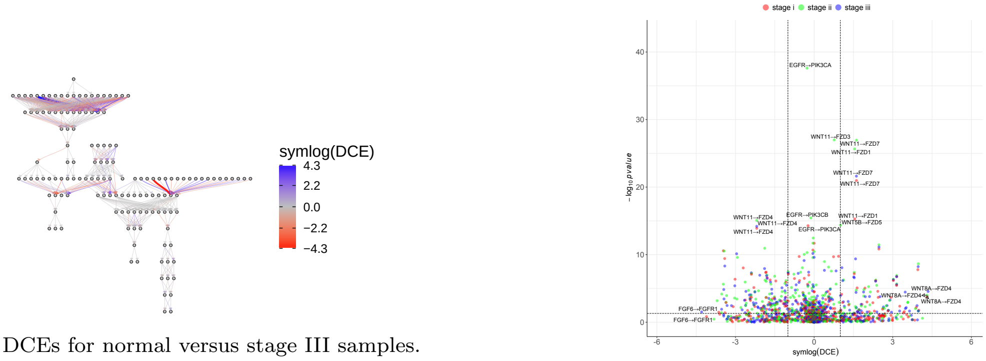
We have presented a new method, *dce*, to compute differential causal effects between two conditions using a regression approach. *dce* enables the edge-specific identification of signaling pathway dysregulations. This piece of information can help to further our understanding of subtle differences on the molecular level in seemingly similar cancer types.

dce assumes a linear relationship among pathway genes. The linear model is solved using



(a) DCEs for normal versus stage I samples.

(b) DCEs for normal versus stage II samples.



(c) DCEs for normal versus stage III samples.

(d) Volcano plot of effect size on the x-axis against the p-value on the y-axis for all interactions over all three stages.

Figure 5: DCEs for *TCGA-BRCA normal* samples versus stage I, stage II, and stage III computed with the *hsa05224* pathway. In (a)-(c), edge thickness and opacity scale with absolute DCE size. More negative DCEs appear red, more positive DCEs appear blue. The color follows a symmetric logarithmic scale for values $|x| \geqslant 1$ and is linear otherwise. (d) shows a volcano plot for the symmetric logarithm of DCE against its associated p-value. DCE thresholds of 1 and -1 as well as a p-value threshold of 0.05 are denoted with grey dashed lines.

network information to account for additional genes confounding the linear relationship between gene pairs. The network information is included via prior knowledge from literature. *dce* also includes latent variables in the model accounting, e.g., for batch effects, which are unknown and not included in the gene network, as confounders.

We have shown in our simulations that *dce* is able to detect changes in causal effects even in the presence of noise and for certain ranges of effect sizes. For a wide array of parameter choices, *dce* outperforms methods using (partial) correlation and *fggm*. Especially in the case of latent confounders we showed that *dce* with the integration of latent variables outperforms *dce* without, except if no latent confounders were used to simulate the data. In this case both methods are equally accurate. Hence, we recommend the integration of latent variables in the model as the default configuration.

In addition to the synthetic benchmark, we have also validated our method on real data derived from Perturb-seq experiments. We have shown that *dce* is able to recover the experimental knockouts with better performance than correlations and partial correlations.

For breast cancer, we have shown that not all parts of the signaling pathway are perturbed and characteristic hotspots exist. Some causal effects between two genes are invariant to stage information, while other causal effects can vary in either magnitude or even sign of their effect size. This indicates that certain areas of such pathways are more relevant than others. This phenomenon has also been observed in other studies ((Song et al., 2014; Feng et al., 2018)). Some parts of a pathway seem to be either more conserved or just not relevant to tumorigenesis. This provides exciting opportunities to identify drugs which target certain parts of a pathway and might explain their efficacy. However, the robustness of our method depends on the availability of enough samples. In many cases, few are available and make our approach infeasible. While *dce* performs still better than random for even 10 samples, it is significantly worse than for higher sample sizes.

In summary, we have proposed a novel application of the concept of differential causal effects which describe the differences in causal effects between two conditions and developed a regression approach to compute those differences. We demonstrate their robustness in a simulation study, and point out interesting results in application to real data, e.g., we show that some dysregulated edges are consistent among breast cancer tumor stages I-III, but that other dysregulations are unique to each stage.

Our simulations show the need for sufficiently large data sets when dealing with large pathways. Additionally *dce* relies on correct network information. While very robust to incorrect edges in the network, *dce*'s performance breaks down significantly when edges are missing from the network. We have also simulated data from DAGs only. However, this assumption is made throughout all analyses. In reality biological pathways include cycles, which could affect the result of *dce*. Similarly, we rely on the assumption that all causal effects are propagated linearly. Other types of causal effect could affect *dce* as well.

Future research should focus on modifying the regression to make working with small data sets more robust, for example, enforcing sparsity by the introduction of L_1- or L_2-norms on the coefficients to avoid outliers produced by artifacts in the data.

We have shown the performance of *dce* on count data from simulations and (single cell) RNA-Seq. However, *dce* is also suited to analyze other types of data like Gaussian data from log-normal microarray intensities.

5 Data availability

The code used to construct the synthetic data sets is available as part of the R software package *dce*. The experimental data used in the Perturb-seq validation is available under the accession GSE90546 from NCBI GEO. The experimental data used in the exploratory breast cancer analysis is available under the accession TCGA-BRCA from The Cancer Genome Atlas. The pathway structures have been obtained from the Kyoto Encyclopedia of Genes and Genome.

6 Code availability

The method *dce* is freely available as an R package on Bioconductor as well as on `https://github.com/cbg-ethz/dce`. The GitHub repository also contains the Snakemake (Mölder et al., 2021) workflows needed to reproduce all results presented here.

7 Acknowledgements

Part of this work has been funded by SystemsX.ch, the Swiss Initiative in Systems Biology, under Grant No. RTD 2013/152 (TargetInfectX - Multi- Pronged Perturbation of Pathogen Infection in Human Cells), evaluated by the Swiss National Science Foundation, and by ERC Synergy Grant 609883 (to NB). The research of DC and PB has been supported by the European Research Council under the Grant Agreement No. 786461 (CausalStats - ERC-2017-ADG).

8 Author information

Contributions KPJ and MP conceived the project. KPJ and MP developed the statistical model of *dce* and implemented the software package. DC contributed to the statistical methodology as well as software implementation. NB and PB supervised the study. KPJ and MP wrote the initial manuscript draft. All authors edited the manuscript.

Corresponding authors Correspondence to Niko Beerenwinkel.

9 Ethics declarations

Competing interests The authors declare no competing interests.

10 Additional Figures

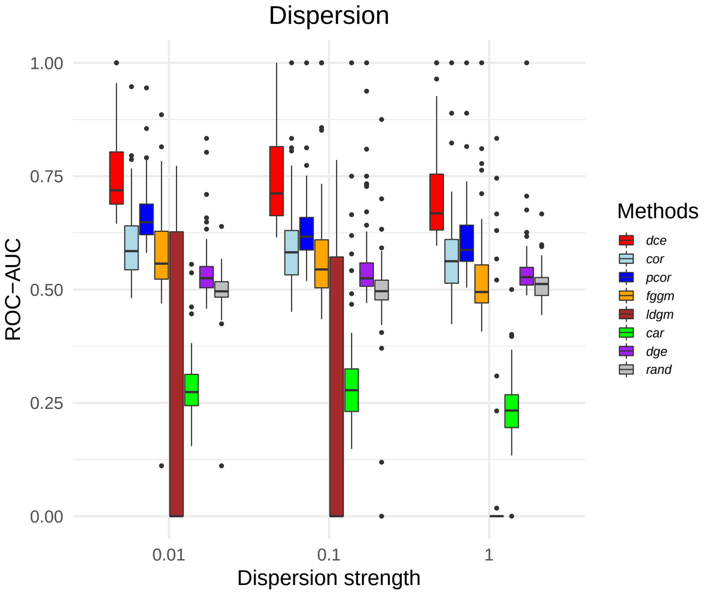


Figure 6: Dispersion. *dce* is compared to its competitors over 100 synthetic data sets with varying dispersion values. Performance decreases for higher dispersion values. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

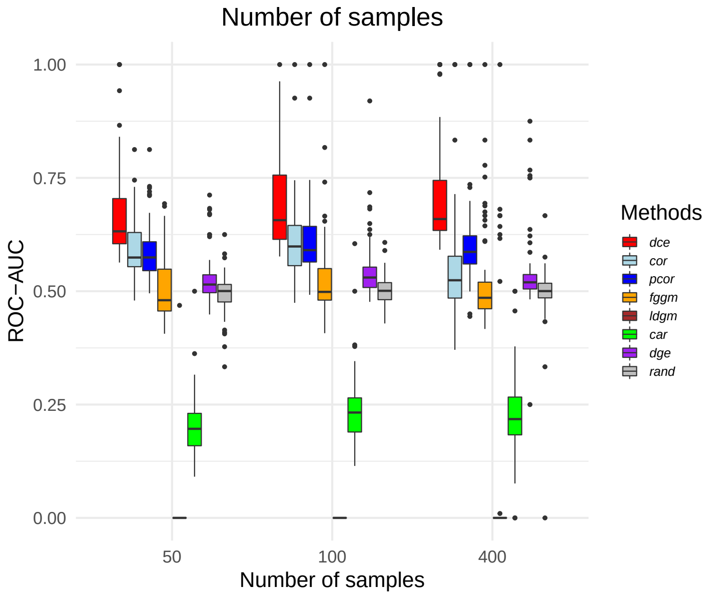


Figure 7: Sample size. *dce* is compared to its competitors over 100 synthetic data sets with varying sample sizes for one condition. The other conditions has a fixed sample size of 200. Performance decreases for lower sample sizes. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

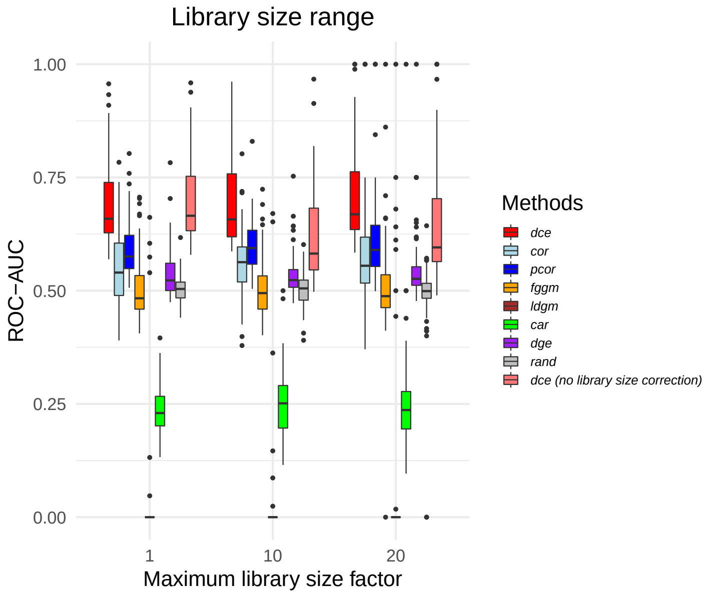


Figure 8: Library size. *dce* is compared to its competitors over 100 synthetic data sets with varying library size factors. Library size has little effect on the accuracy of all methods. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

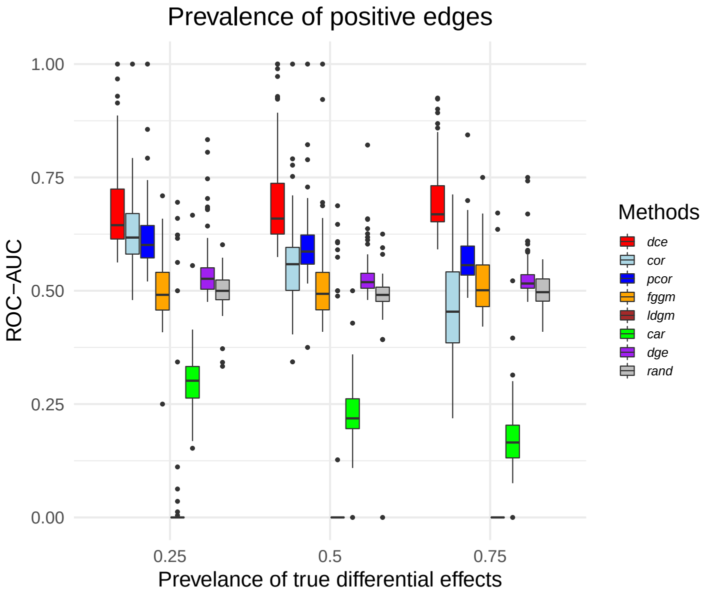


Figure 9: Prevalence. *dce* is compared to its competitors over 100 synthetic data sets with varying prevalence for DCE $\neq 0$. Accuracy decreases for all methods and higher prevalence except for *dce*. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

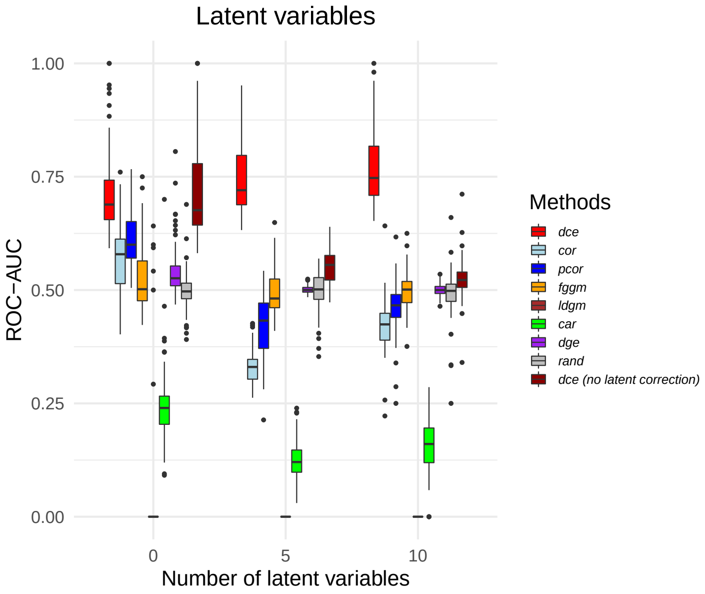


Figure 10: Latent variables. *dce* is compared to its competitors over 100 synthetic data sets with varying numbers of latent variables. *dce*'s accuracy stays robust, if we account for latent variables, but drastically decreases, if we do not. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

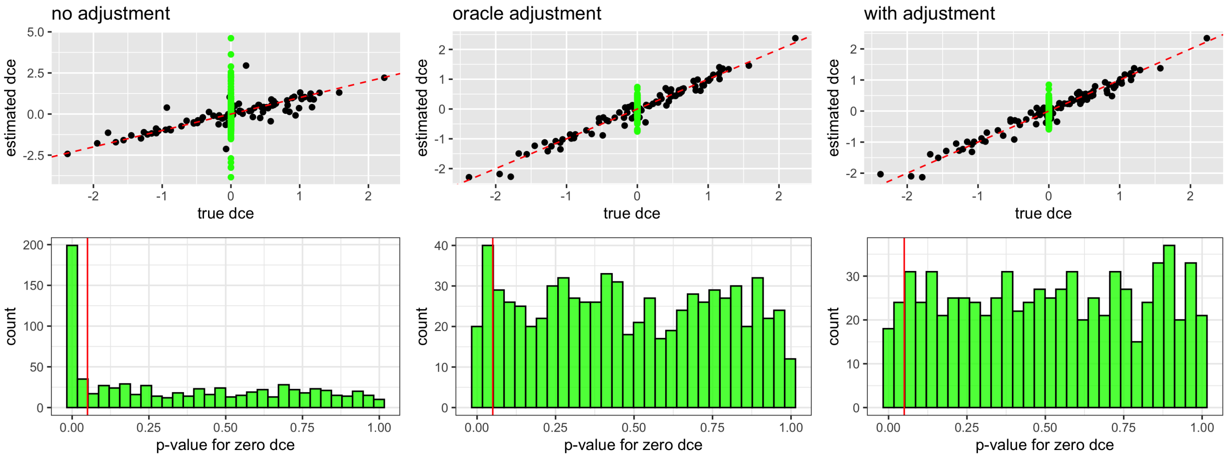


Figure 11: The performance of the *dce* without latent confounding adjustment (left), *dce* using true values of condounders (not known in practice) and *dce* with the latent confounding adjustment. Null DCEs are denoted in green, whereas the non-zero DCEs are denoted in black. This figure uses synthetic data with 300 genes, 300 observations and 3 latent confounders. Red line in the bottom row indicates the 0.05 threshold. The performance with the deconfounding step is close to the performance if we actually observed the latent confounders. Furthermore, it avoids increased number of falsely significant findings due to confounding bias (bottom row).

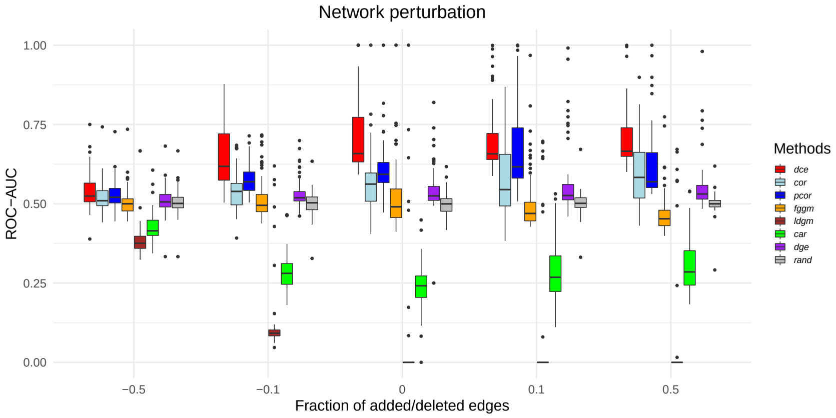


Figure 12: *dce* is compared to its competitors over 100 synthetic data sets with incorrect network information. Performance decreases for networks with missing edges, but stays robust, if additional edges are included. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

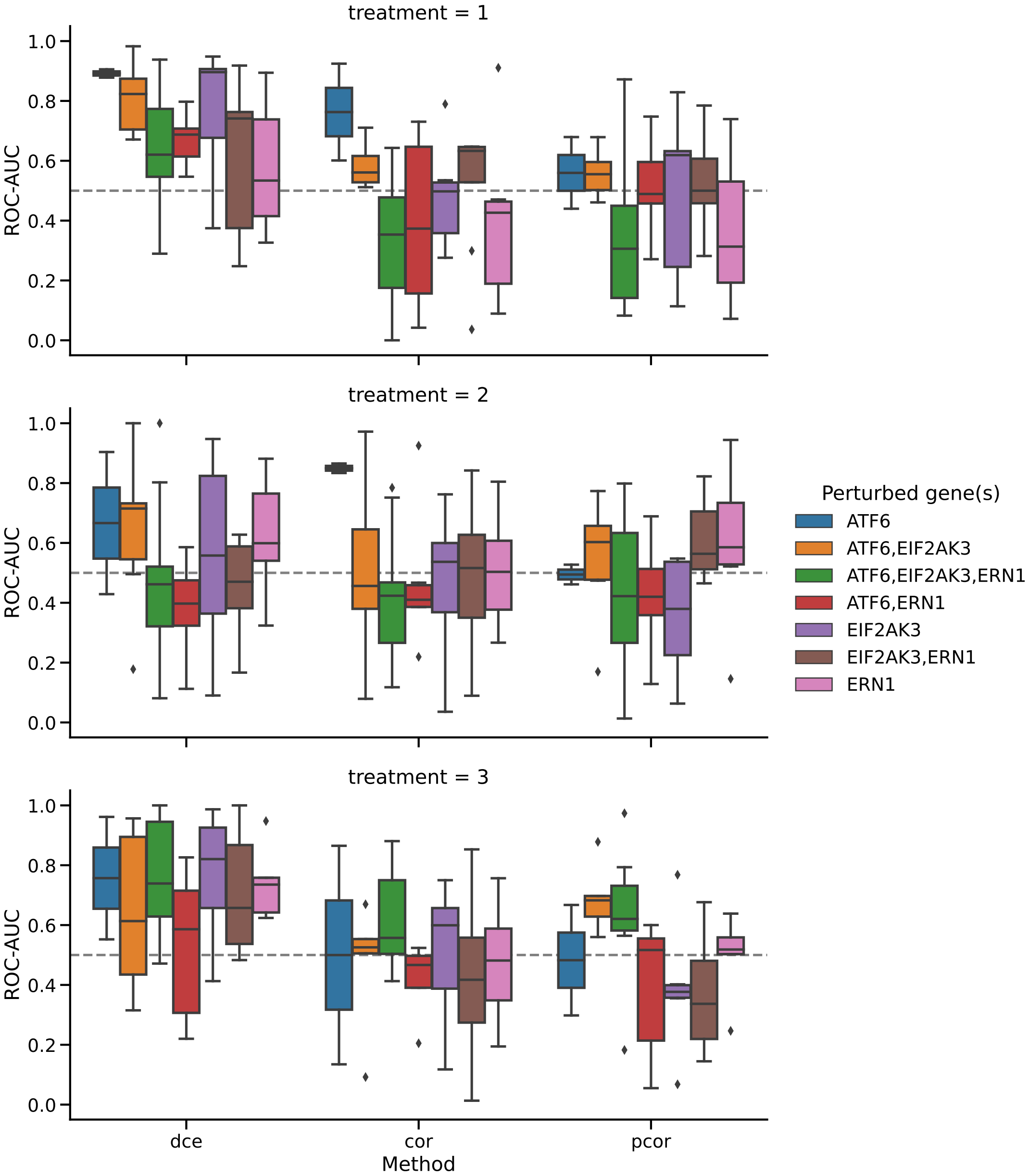


Figure 13: Dependence of the performance for the CRISPR benchmark when stratified by treatment and knockout gene.

Paper

D

Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression.

Ćevid, D., Michel, L., Näff, J., Bühlmann, P., Meinshausen, N.

Submitted to Journal of Machine Learning Research.

Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression

Domagoj Ćevid*, Loris Michel*, Jeffrey Näff*, Peter Bühlmann*, Nicolai Meinshausen*

*Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland.

Abstract

Random Forests (Breiman, 2001) is a successful and widely used regression and classification algorithm. Part of its appeal and reason for its versatility is its (implicit) construction of a kernel-type weighting function on training data, which can also be used for targets other than the original mean estimation. We propose a novel forest construction for multivariate responses based on their joint conditional distribution, independent of the estimation target and the data model. It uses a new splitting criterion based on the MMD distributional metric, which is suitable for detecting heterogeneity in multivariate distributions. The induced weights define an estimate of the full conditional distribution, which in turn can be used for arbitrary and potentially complicated targets of interest. The method is very versatile and convenient to use, as we illustrate on a wide range of examples. The code is available as `Python` and `R` packages `drf`.

Keywords. causality, distributional regression, fairness, Maximal Mean Discrepancy, Random Forests, two-sample testing

1 Introduction

In practice, one often encounters heterogeneous data, whose distribution is not constant, but depends on certain covariates. For example, data can be collected from several different sources, its distribution might differ across certain subpopulations or it could even change with time, etc. Inferring valid conclusions about a certain target of interest from such data can be very challenging as many different aspects of the distribution could potentially change. As an example, in medical studies, the effectiveness of a certain treatment might not be constant throughout the population but depend on certain patient characteristics such as age, race, gender, or medical history. Another issue could be that different patient groups were not equally likely to receive the same treatment in the observed data.

Obviously, pooling all available data together can result in invalid conclusions. On the other hand, if for a given test point of interest one only considers similar training data points, i.e. a small homogeneous subpopulation, one may end up with too few samples for accurate statistical estimation. In this paper, we propose a method based on the Random Forest algorithm (Breiman, 2001) which in a data-adaptive way determines for any given test point which training data points are relevant for it. This in turn can be used for drawing valid conclusions or for accurately estimating any quantity of interest.

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d) \in \mathbb{R}^d$ be a multivariate random variable representing the data of interest, but whose joint distribution is heterogeneous and depends on some subset of a potentially large number of covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$. Throughout the paper, vector quantities are denoted in bold. We aim to estimate a certain target object $\tau(\mathbf{x})$ that depends on the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}) = \mathbb{P}(\mathbf{Y} \mid X_1{=}x_1, \ldots, X_p{=}x_p)$, where $\mathbf{x} = (x_1, \ldots, x_p)$ is an arbitrary point in $\mathbb{R}^p$. The estimation target $\tau(\mathbf{x})$ can range from simple quantities, such as the conditional expectations $\mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}]$ (Breiman, 2001) or quantiles $Q_\alpha[f(\mathbf{Y}) \mid \mathbf{X}]$ (Meinshausen, 2006) for some function $f : \mathbb{R}^d \to \mathbb{R}$, to some more complicated aspects of the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$, such as conditional copulas or conditional independence measures. Given the observed data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the most straightforward way of estimating $\tau(\mathbf{x})$ nonparametrically would be to consider only the data points in some neighborhood $\mathcal{N}_\mathbf{x}$ around $\mathbf{x}$, e.g. by considering the k nearest neighbors according to some metric. However, such methods typically suffer from the curse of dimensionality even when p is only moderately large: for a reasonably small neighborhood, such that the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} \in \mathcal{N}_\mathbf{x})$ is close to the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$, the number of training data points contained in $\mathcal{N}_\mathbf{x}$ will be very small, thus making the accurate estimation of the target $\tau(\mathbf{x})$ difficult. The same phenomenon occurs with other methods which locally weight the training observations such as kernel methods (Silverman, 1986), local MLE (Fan et al., 1998) or weighted regression (Cleveland, 1979) even for the relatively simple problem of estimating the conditional mean $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}]$ for fairly small p. For that reason, more importance should be given to the training data points $(\mathbf{x}_i, \mathbf{y}_i)$ for which the response distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}_i)$ at point $\mathbf{x}_i$ is similar to the target distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$, even if $\mathbf{x}_i$ is not necessarily close to $\mathbf{x}$ in every component.

In this paper, we propose the Distributional Random Forest (DRF) algorithm which estimates the multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ in a locally adaptive fashion. This is done by repeatedly dividing the data points in the spirit of the Random Forest algorithm (Breiman, 2001): at each step, we split the data points into two groups based on some feature X_j in such a way that the distribution of $\mathbf{Y}$ for which $X_j \leqslant l$, for some level l, differs the most compared to the distribution of $\mathbf{Y}$ when $X_j > l$, according to some distributional metric. One can use any multivariate two-sample test statistic, provided it can detect a wide variety of distributional changes. As the default choice, we propose a criterion based on the Maximal Mean Discrepancy (MMD) statistic (Gretton et al., 2007a) with many interesting properties. This splitting procedure partitions the data such that the distribution of the multivariate response $\mathbf{Y}$ in the resulting leaf nodes is as homogeneous as possible, thus defining neighborhoods of relevant training data points for every $\mathbf{x}$. Repeating this many times with randomization induces a weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ as in Lin and Jeon (2002, 2006), described in detail in Section 2, which quantifies the relevance of each training data point $\mathbf{x}_i$ for a given test point $\mathbf{x}$. The conditional distribution is then estimated by an empirical distribution determined by these weights (Meinshausen, 2006). This construction is data-adaptive as it assigns more weight to the training points $\mathbf{x}_i$ that are closer to the test point $\mathbf{x}$ in the components which are more relevant for the distribution of $\mathbf{Y}$.

Our forest construction does not depend on the estimation target $\tau(\mathbf{x})$, but it rather estimates the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ directly and the induced forest weights can be used to estimate $\tau(\mathbf{x})$ in a second step. This approach has several advantages. First, only one DRF fit is required to obtain estimates of many different targets, which has a big computational advantage. Furthermore, since those estimates are obtained from the same forest fit, they are mutually compatible. For example, if the conditional correlation matrix $\{\mathrm{Cor}(Y_i, Y_j \mid \mathbf{X}=\mathbf{x})\}_{i,j=1}^{d}$ were estimated componentwise, the resulting matrix might not be positive semidefinite, and as another example, the CDF estimates $\hat{\mathbb{P}}(\mathbf{Y} \leqslant \mathbf{y} \mid \mathbf{X}=\mathbf{x})$ might not be monotone in $\mathbf{y}$, see Figure 6. Finally, it can be extremely difficult to tailor forest construction to more complex targets $\tau(\mathbf{x})$. The induced weighting function can thus be used not only for obtaining simple distributional aspects such as, for example, the conditional quantiles, conditional correlations, or joint conditional probability statements, but also to obtain more complex objectives, such as conditional independence tests (Zhang et al., 2012), heterogeneous regression (see also Section 4.4 for more details) (Künzel et al., 2019; Wager and Athey, 2018) or semiparametric estimation by fitting a parametric model for $\mathbf{Y}$, having nonparametrically adjusted for $\mathbf{X}$ (Bickel et al., 1993). Representation of the conditional distribution via the weighting function has a great potential for applications in causality such as causal effect estimation or as a way of implementing do-calculus (Pearl, 2009) for finite samples, as we discuss in Section 4.4.

Therefore, DRF is used in two steps: in the first step, we obtain the weighting function $w_{\mathbf{x}}(\cdot)$ describing the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ in a target- and model-free way, which is then used as an input for the second step. Even if the method used in the second step does not directly support weighting of the training data points, one can easily resample the data set with sampling probabilities equal to $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^n$. The two-step approach is visualized in the following diagram:

$$\begin{tikzcd}
\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}) \arrow[r, "\text{1) get } w_{\mathbf{x}}(\cdot) \text{ with DRF}"] \arrow[d, dashed, "\text{objective}"'] & \hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}) \arrow[d, "\text{2) compute from } w_{\mathbf{x}}(\cdot)"] \\
\tau(\mathbb{P}) & \tau(\hat{\mathbb{P}}) \arrow[l, dashed, "\text{induced estimator}"']
\end{tikzcd}$$

1.1 Related work and our contribution

Several adaptations of the Random Forest algorithm have been proposed for targets beyond the original one of the univariate conditional mean $\mathbb{E}[Y \mid \mathbf{X}=\mathbf{x}]$: for survival analysis (Hothorn et al., 2006), conditional quantiles (Meinshausen, 2006), density estimation (Pospisil and Lee, 2018), CDF estimation (Hothorn and Zeileis, 2021) or heterogeneous treatment effects (Wager and Athey, 2018). Almost all such methods use the weights induced by the forest, as described in Section 2, rather than averaging the estimates obtained per tree. This view of Random Forests as a powerful adaptive nearest neighbor method is well known and dates back to Lin and Jeon (2002, 2006). It was first used for targets beyond the conditional mean in Meinshausen (2006), where the original forest construction with univariate Y was used (Breiman, 2001). However, the univariate response setting considered there severely restricts the number of interesting targets $\tau(\mathbf{x})$ and our DRF can thus be viewed as an important generalization of this approach to the multivariate setting.

In order to be able to perform certain tasks or to achieve a better accuracy, many forest-based methods adapt the forest construction by using a custom splitting criterion tailored to their specific target, instead of relying on the standard CART criterion. In Zeileis et al. (2008) and Hothorn and Zeileis (2021), a parametric model for the response $\mathbf{Y} \mid \mathbf{X}=\mathbf{x} \sim f(\theta(\mathbf{x}), \cdot)$ is assumed and recursive splitting is performed based on a permutation test which uses the user-provided score functions. Similarly, Athey et al. (2019) estimate certain univariate targets for which there exist corresponding score functions defining the local estimating equations. The data is split so that the estimates of the target in resulting child nodes differ the most. This is different, though, to the target-free splitting criterion of DRF, which splits so that the distribution of $\mathbf{Y}$ in child nodes is as different as possible.

Since the splitting step is extensively used in the algorithm, its complexity is crucial for the overall computational efficiency of the method, and one often needs to resort to approximating the splitting criterion (Pospisil and Lee, 2018; Athey et al., 2019) to obtain good computational run time. We propose a splitting criterion based on a fast random approximation of the MMD statistic (Gretton et al., 2012a; Zhao and Meng, 2015), which is commonly used in practice for two-sample testing as it is able to detect any change in the multivariate distribution of $\mathbf{Y}$ with good power (Gretton et al., 2007a). DRF with the MMD splitting criterion also has interesting theoretical properties as shown in Section 3 below.

The multivariate response case has not received much attention in the Random Forest literature. Most of the existing forest-based methods focus on either a univariate response Y or on a certain univariate target $\tau(\mathbf{x})$. One interesting line of work considers density estimation (Pospisil and Lee, 2018) and uses aggregation of the CART criteria for different response transformations. Another approach (Kocev et al., 2007; Segal and Xiao, 2011; Ishwaran and Kogalur, 2014) is based on aggregating standard univariate CART splitting criteria for $Y_1, \ldots, Y_d$ and targets only the conditional mean of the responses, a task which could also be solved by separate regression fits for each Y_i. In order to capture any change in the distribution of the multivariate response $\mathbf{Y}$, one needs to not only consider the marginal distributions for each component Y_i, but also to determine whether their dependence structure changes, see e.g. Figure 8.

There are not many methods that nonparametrically estimate the joint multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ in the statistics or machine learning literature. Other than a few simple classical methods such as k-nearest neighbors and kernel regression, there are methods based on normalizing flows such as Inverse Autoregressive Flow (Kingma et al., 2016) or Masked Autoregressive Flow (Papamakarios et al., 2017) and also conditional variants of several popular generative models such as Conditional Generative Adversarial Networks (Mirza and Osindero, 2014) or Conditional Variational Autoencoder (Sohn et al., 2015). The focus of these methods is more on the settings with large d and small p, such as image or text generation. The comparison of DRF with the competing methods for distributional estimation can be found in Section 4.1.

Our contribution, resulting in the proposal of the Distributional Random Forest (DRF), can be summarized as follows: First, we introduce the idea of forest construction based on sequential multivariate two-sample test statistics. It does not depend on a particular estimation target and is completely nonparametric, which makes its implementation and usage very simple and universal. Not only does it not require additional user input such as the log-likelihoods or score functions, but it can be used even for complicated targets for which there is no obvious forest construction. Furthermore, it has a computational advantage as only a single forest fit is needed for producing estimates of many different targets that are additionally compatible with each other. Second, we propose an MMD-based splitting criterion with good statistical and computational properties, for which we also derive interesting theoretical results in Section 3. It underpins our implementation, which we provide as `R` and `Python` packages `drf`. Finally, we show on a broad range of examples in Section 4 how many different statistical estimation problems, some of which not being easily tractable by existing forest-based methods, can be cast to our framework, thus illustrating the usefulness and versatility of DRF.

2 Method

In this section we describe the details of the Distributional Random Forest (DRF) algorithm. We closely follow the implementations of the `grf` (Athey et al., 2019) and `ranger` (Wright and Ziegler, 2015) `R`-packages. A detailed description of the method and its implementation and the corresponding pseudocode can be found in the Appendix A of Ćevid et al. (2020b).

2.1 Forest Building

The trees are grown recursively in a model-free and target-free way as follows: For every parent node P, we determine how to best split it into two child nodes of the form $C_L = \{X_j \leqslant l\}$ and $C_R = \{X_j > l\}$, where the variable X_j is one of the randomly chosen splitting candidates and l denotes its level based on which we perform the splitting. The split is chosen such that we maximize a certain (multivariate) two-sample test statistic

$$\mathcal{D}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right), \tag{1}$$

which measures the difference of the empirical distributions of the data $\mathbf{Y}$ in the two resulting child nodes C_L and C_R. Therefore, in each step we select the candidate predictor X_j which seems to affect the distribution of $\mathbf{Y}$ the most, as measured by the metric $\mathcal{D}(\cdot,\cdot)$. Intuitively, in this way we ensure that the distribution of the data points in every leaf of the resulting tree is as homogeneous as possible, which helps mitigate the bias caused by pooling the heterogeneous data together. A related idea can be found in GRF (Athey et al., 2019), where one attempts to split the data so that the resulting estimates $\hat{\tau}_L$ and $\hat{\tau}_R$, obtained respectively from data points in C_L and C_R, differ the most:

$$\frac{n_L n_R}{n_P^2}\left(\hat{\tau}_L - \hat{\tau}_R\right)^2, \tag{2}$$

where we write $n_P = |\{i \mid \mathbf{x}_i \in P\}|$ and n_L, n_R are defined analogously.

One could construct the forest using any metric $\mathcal{D}(\cdot,\cdot)$ for empirical distributions. However, in order to have a good accuracy of the overall method, the corresponding two-sample test using $\mathcal{D}(\cdot,\cdot)$ needs to have a good power for detecting any kind of change in distribution, which is a difficult task in general, especially for multivariate data (Bai and Saranadasa, 1996; Székely and Rizzo, 2004). Another very important aspect of the choice of distributional metric $\mathcal{D}(\cdot,\cdot)$ is the computational efficiency; one needs to be able to sequentially compute the values of $\mathcal{D}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\},\{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right)$ for every possible split very fast for the overall algorithm to be computationally feasible, even for moderately large datasets. Below, we propose a splitting criterion based on the MMD two-sample test statistic (Gretton et al., 2007a) which has both good statistical and computational properties.

In contrast to other forest-based methods, we do not use any information about our estimation target τ in order to find the best split of the data, which comes with a certain trade-off. On one hand, it is sensible that tailoring the splitting criterion to the target should improve the estimation accuracy; for example, some predictors might affect the conditional distribution of $\mathbf{Y}$, but not necessarily the estimation target τ and splitting on such predictors unnecessarily reduces the number of training points used for estimating τ. On the other hand, our approach has multiple benefits: it is easier to use as it does not require any user input such as the likelihood or score functions and it can also be used for very complicated targets for which one could not easily adapt the splitting criterion. Furthermore, only one DRF fit is necessary for producing estimates of many different targets, which has both computational advantage and the practical advantage that the resulting estimates are mutually compatible (see e.g. Figure 5).

Interestingly, sometimes it could even be beneficial to split based on a predictor which does not affect the target of estimation, but which affects the conditional distribution. This is illustrated by the following toy example. Suppose that for a bivariate response (Y_1, Y_2) we are interested in estimating the slope of the linear regression of Y_2 on Y_1 conditionally on $p = 30$ predictors $\mathbf{X}$, i.e. our target is $\tau(\mathbf{x}) = \mathrm{Cov}(Y_1, Y_2 \mid \mathbf{X}{=}\mathbf{x})/\mathrm{Var}(Y_1 \mid \mathbf{X}{=}\mathbf{x})$. This is one of the main use cases for GRF and its variant which estimates this target is called Causal Forest (Wager and Athey, 2018; Athey et al., 2019). Let us assume that the data has the following distribution:

$$\mathbb{P}\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \,\middle|\, \mathbf{X}{=}\mathbf{x}\right) \sim N\left(\begin{bmatrix} x_1 \\ x_1 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right) \qquad \mathbf{X} \sim N(\mathbf{0}, I_p), \tag{3}$$

i.e. X_1 affects only the mean of the responses, while the other $p-1$ predictors have no effect. In Figure 1 we illustrate the distribution of the data when $n = 300, p = 30, \sigma = 0.2$, together with the DRF and GRF splitting criteria. The true value of the target is $\tau(\mathbf{x}) = 0$, but when σ is not too big, the slope estimates $\hat{\tau}$ on pooled data will be closer to 1.

Therefore, the difference of $\hat{\tau}_L$ and $\hat{\tau}_R$ between the induced slope estimates for a candidate split, which is used for splitting criterion (2) of GRF, might not be large enough for us to decide to split on X_1, or the resulting split might be too unbalanced. This results in worse forest estimates for this toy example, see Figure 1.

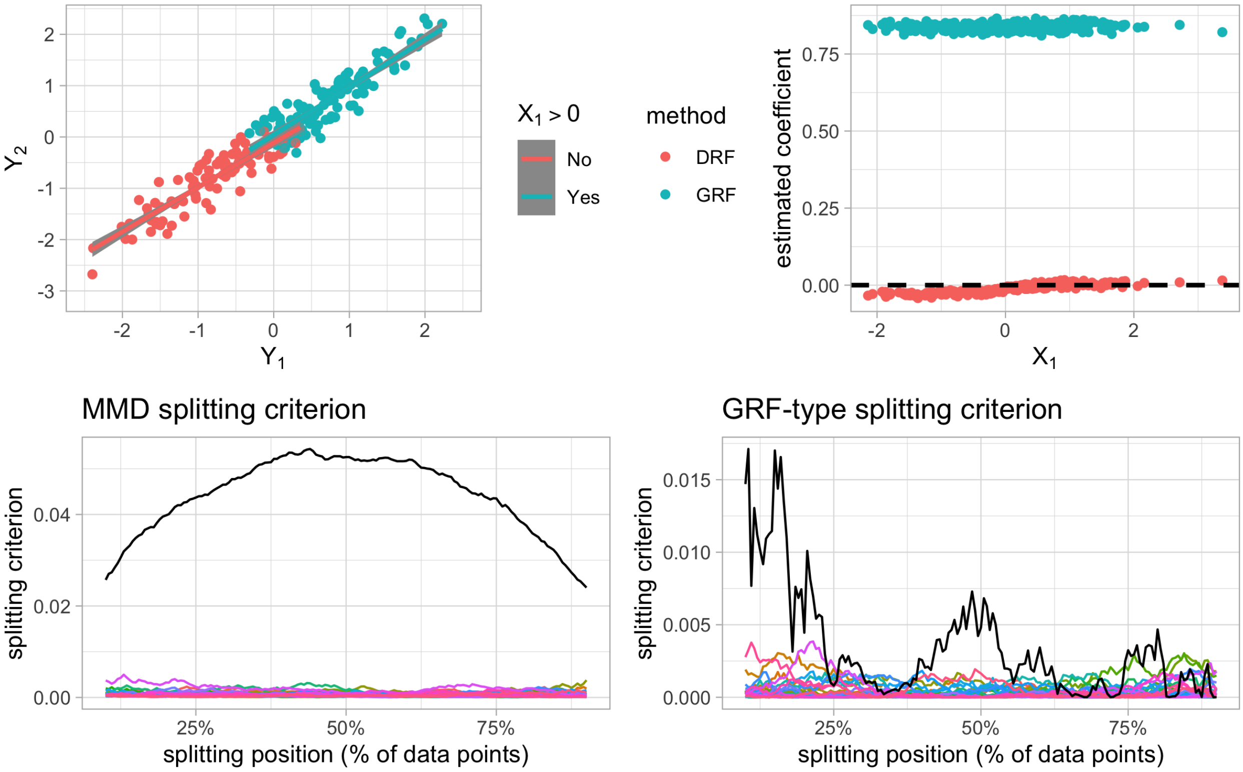


Figure 1: Top left: Illustration of data distribution for the toy example (3) when $n = 300$, $p = 30$. Bottom: The corresponding MMD (12) (left) and GRF (2) splitting criteria (right) at the root node. The curves of different colors correspond to different predictors, with X_1 denoted in black. Top right: Comparison of the estimates of DRF and Causal Forest (Athey et al., 2019) which respectively use those splitting criteria. Test points were randomly generated from the same distribution as the training data. Black dashed line indicates the correct value of the target quantity.

2.2 Weighting Function

Having constructed our forest, just as the standard Random Forest (Breiman, 2001) can be viewed as the weighted nearest neighbor method (Lin and Jeon, 2002), we can use the induced weighting function to estimate the conditional distribution at any given test point $\mathbf{x}$ and thus any other quantity of interest $\tau(\mathbf{x})$. This approach is commonly used in various forest-based methods for obtaining predictions Hothorn and Zeileis (2021); Pospisil and Lee (2018); Athey et al. (2019).

Suppose that we have built N trees $\mathcal{T}_1, \ldots, \mathcal{T}_N$. Let $\mathcal{L}_k(\mathbf{x})$ be the set of the training data points which end up in the same leaf as $\mathbf{x}$ in the tree $\mathcal{T}_k$. The weighting function

$w_{\mathbf{x}}(\mathbf{x}_i)$ is defined as the average of the corresponding weighting functions per tree (Lin and Jeon, 2006):

$$w_{\mathbf{x}}(\mathbf{x}_i) = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbb{1}\left(\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})\right)}{|\mathcal{L}_k(\mathbf{x})|}. \tag{4}$$

The weights are positive add up to 1: $\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) = 1$. In the case of equally sized leaf nodes, the assigned weight to a training point $\mathbf{x}_i$ is proportional to the number of trees where the test point $\mathbf{x}$ and $\mathbf{x}_i$ end up in the same leaf node. This shows that forest-based methods can in general be viewed as adaptive nearest neighbor methods. The sets $\mathcal{L}_k(\mathbf{x})$ of DRF will contain data points $(\mathbf{x}_i, \mathbf{y}_i)$ such that $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}_i)$ is close to $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$, thus removing bias due to heterogeneity of $\mathbf{Y}$ caused by $\mathbf{X}$. On the other hand, since the trees are constructed randomly and are thus fairly independent (Breiman, 2001), the leaf sets $\mathcal{L}_k(\mathbf{x})$ will be different enough so that the induced weights $w_{\mathbf{x}}(\mathbf{x}_i)$ are not concentrated on a small set of data points, which would lead to high estimation variance. Such good bias-variance tradeoff properties of forest-based methods are also implied by their asymptotic properties (Biau, 2012; Wager, 2014), even though this is a still active area of research and not much can be shown rigorously.

One can estimate the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$ from the weighting function by using the corresponding empirical distribution:

$$\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \cdot \delta_{\mathbf{y}_i}, \tag{5}$$

where $\delta_{\mathbf{y}_i}$ is the point mass at $\mathbf{y}_i$.

The weighting function $w_{\mathbf{x}}(\mathbf{x}_i)$ can directly be used for any target $\tau(\mathbf{x})$ in a second step and not just for estimating the conditional distribution. For example, the estimated conditional joint CDF is given by

$$\hat{F}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}(\mathbf{t}) = \hat{\mathbb{P}}(Y_1 \leqslant t_1, \ldots, Y_d \leqslant t_d \mid \mathbf{X}=\mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \mathbb{1}(\cap_{j=1}^{d} \{(\mathbf{y}_i)_j \leqslant t_j\}). \tag{6}$$

It is important to point out that using the induced weighting function for locally weighted estimation is different than the approach of averaging the noisy estimates obtained per tree (Wager and Athey, 2018), originally used in standard Random Forests (Breiman, 2001). Even though the two approaches are equivalent for conditional mean estimation, the former approach is often much more efficient for more complex targets (Athey et al., 2019), since the number of data points in a single leaf is very small, leading to large variance of the estimates.

For the univariate response, the idea of using the induced weights for estimating targets different than the original target of conditional mean considered in Breiman (2001) dates back to Quantile Regression Forests (QRF) (Meinshausen, 2006), where a lot of emphasis is put on the quantile estimation, as the number of interesting targets is quite limited in the univariate setting.

In the multivariate case, on the other hand, many interesting quantities such as, for example, conditional quantiles, conditional correlations or various conditional probability statements can easily be directly estimated from the weights.

By using the weights as an input for some other method, we can accomplish some more complicated objectives, such as conditional independence testing, causal effect estimation, semiparametric learning, time series prediction or tail-index estimation in extreme value analysis. As an example, suppose that our data $\mathbf{Y}$ come from a certain parametric model, where the parameter θ is not constant, but depends on $\mathbf{X}$ instead, i.e. $\mathbf{Y} \mid \mathbf{X} = \mathbf{x} \sim f(\theta(\mathbf{x}), \cdot)$, see also Zeileis et al. (2008). One can then estimate the parameter $\theta(\mathbf{x})$ by using weighted maximum likelihood estimation:

$$\hat{\theta}(\mathbf{x}) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) \log f(\theta, \mathbf{y}_i).$$

Another example is heterogeneous regression, where we are interested in the regression fit of an outcome $Y \in \mathbb{R}$ on certain predicting variables $\mathbf{W} \in \mathbb{R}^s$ conditionally on some event $\{\mathbf{X} = \mathbf{x}\}$. This can be achieved by weighted regression of Y on $\mathbf{W}$, where the weight $w_{\mathbf{x}}(\mathbf{x}_i)$ assigned to each data point $(\mathbf{w}_i, y_i)$ is obtained from DRF with the multivariate response $(Y, \mathbf{W}) \in \mathbb{R}^{s+1}$ and predictors $\mathbf{X} \in \mathbb{R}^p$, for an illustration see Section 4.4.

The weighting function of DRF is illustrated on the air quality data in Figure 2. Five years (2015−2019) of air pollution measurements were obtained from the US Environmental Protection Agency (EPA) website. Six main air pollutants (nitrogen dioxide (NO_2), carbon monoxide (CO), sulphur dioxide (SO_2), ozone (O_3) and coarse and fine particulate matter (PM10 and PM2.5)) that form the air quality index (AQI) were measured at many different measuring sites in the US for which we know the longitude, latitude, elevation, location setting (rural, urban, suburban) and how the land is used within a 1/4 mile radius. Suppose we would want to know the distribution of the pollutant measurements at some new, unobserved, measurement site. The top row illustrates for a given test site, whose characteristics are indicated in the plot title, how much weight in total is assigned to the measurements from a specific training site. We see that the important sites share many characteristics with the test site and that DRF determines the relevance of each characteristic in a data-adaptive way. The bottom row shows the corresponding estimates of the joint conditional distribution of the pollutants (we choose 2 of them for visualization purposes) and one can clearly see how the estimated pollution levels are larger for the suburban site than for the rural site. The induced weights can be used, for example, for estimating the joint density (whose contours can be seen in the plot) or for estimating the probability that the AQI is below a certain value by summing the weights in the

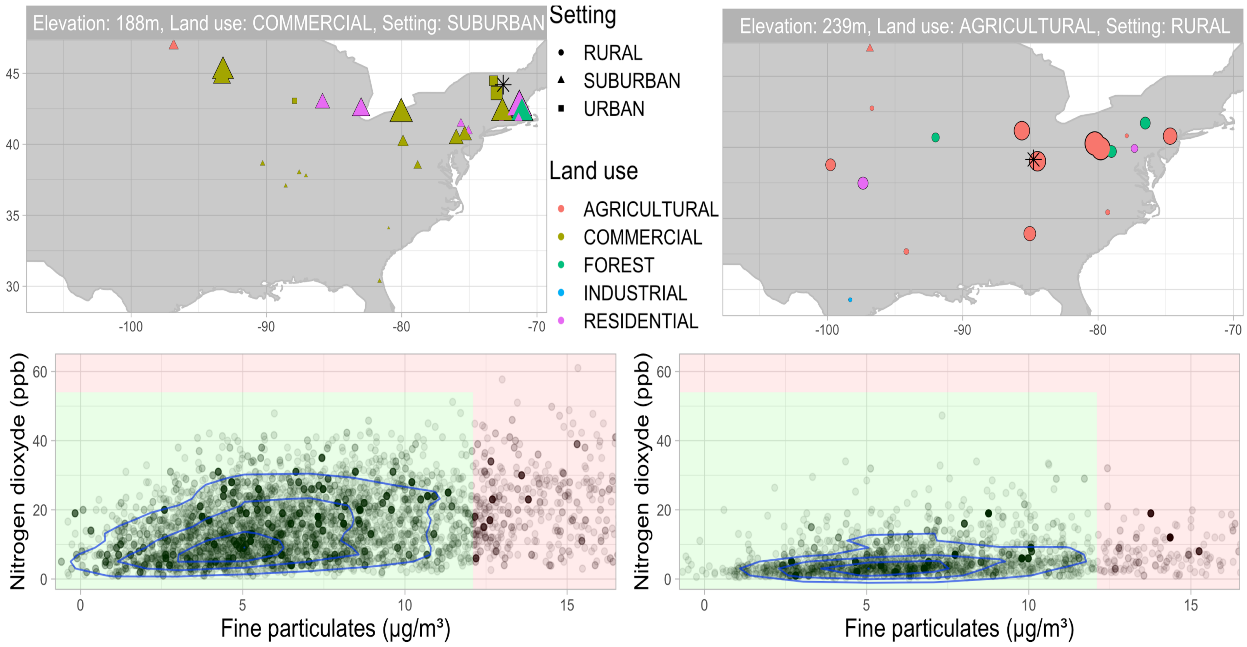


Figure 2: Top: the characteristics of the important training sites, for a fixed test site whose position is indicated by a black star and whose characteristics are indicated in the title. The total weight assigned corresponds to the symbol size. Bottom: estimated joint conditional distribution of two pollutants NO_2 and PM2.5, where the weights correspond to the transparency of the data points. Green area corresponds to 'Good' air quality category (AQI $\leqslant 50$).

corresponding region of space.

2.3 Distributional Metric

In order to determine the best split of a parent node P, i.e. such that the distributions of the responses $\mathbf{Y}$ in the resulting child nodes C_L and C_R differ the most, one needs a good distributional metric $\mathcal{D}(\cdot, \cdot)$ (see Equation (1)) which can detect change in distribution of the response $\mathbf{Y}$ when additionally conditioning on an event $\{X_j > l\}$. Testing equality of distributions from the corresponding samples is an old problem in statistics, called two-sample testing problem. For univariate data, many good tests exist such as Wilcoxon rank test (Wilcoxon, 1946), Welch's t-test (Welch, 1947), Wasserstein two-sample testing (Ramdas et al., 2017), Kolmogorov-Smirnov test (Massey Jr, 1951) and many others, but obtaining an efficient test for multivariate distributions has proven to be quite challenging due to the curse of dimensionality (Friedman and Rafsky, 1979; Baringhaus and Franz, 2004).

Additional requirement for the choice of distributional metric $\mathcal{D}(\cdot, \cdot)$ used for data splitting is that it needs to be computationally very efficient as splitting is used extensively in the algorithm. If we construct N trees from n data points and in each node we consider mtry candidate variables for splitting, the complexity of the standard Random Forest algorithm (Breiman, 2001) in the univariate case is $\mathcal{O}(N \times \text{mtry} \times n \log n)$ provided our splits are balanced. It uses the CART splitting criterion, given by:

$$\frac{1}{n_P}\left(\sum_{\mathbf{x}_i \in C_L}(y_i - \overline{y}_L)^2 + \sum_{\mathbf{x}_i \in C_R}(y_i - \overline{y}_R)^2\right), \tag{7}$$

where $\overline{y}_L = \frac{1}{n_L}\sum_{\mathbf{x}_i \in C_L} y_i$ and $\overline{y}_R$ is defined analogously. This criterion has an advantage that not only it can be computed in $\mathcal{O}(n_P)$ complexity, but this can be done for all possible splits $\{X_j \leqslant l\}$ as cutoff level l varies, since updating the splitting criterion when moving a single training data point from one child node to the other requires only $\mathcal{O}(1)$ computational steps (most easily seen by rewriting the CART criterion as in (13)).

If the time complexity of evaluating the DRF splitting criterion (1) for a single splitting candidate X_j and all cutoffs l of interest (usually taken to range over all possible values) is at least n^c for some $c > 1$, say $\mathcal{O}(f(n_P))$ for some function $f : \mathbb{R} \to \mathbb{R}$, then by solving the recursive relation we obtain that the overall complexity of the method is given by $\mathcal{O}(N \times \text{mtry} \times f(n))$ (Akra and Bazzi, 1998), which can be unfeasible even for moderately large n if f grows too fast.

The problem of sequential two-sample testing is also central to the field of change-point detection (Wolfe and Schechtman, 1984; Brodsky and Darkhovsky, 2013), with the slight difference that in the change-point problems the distribution is assumed to change abruptly at certain points in time, whereas for our forest construction we only are interested in finding the best split of the form $\{X_j \leqslant l\}$ and the conditional distribution $\mathbb{P}(\mathbf{Y} \mid \{\mathbf{X} \in P\} \cap \{X_j \leqslant l\})$ usually changes gradually with l. The testing power and the computational feasibility of the method play a big role in change-point detection as well. However, the state-of-the-art change-point detection algorithms (Li et al., 2019; Matteson and James, 2014) are often too slow for our purpose as sequential testing is done $\mathcal{O}(N \times \text{mtry} \times n)$ times for forest construction, much more frequently than in change-point problems.

2.3.1 MMD splitting criterion

Even though DRF could in theory be constructed with any distributional metric $\mathcal{D}(\cdot,\cdot)$, as a default choice we propose splitting criterion based on the Maximum Mean Discrepancy (MMD) statistic (Gretton et al., 2007a). Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the RKHS of real-valued functions on $\mathbb{R}^d$ induced by some positive-definite kernel k, and let $\varphi : \mathbb{R}^d \to \mathcal{H}$ be the corresponding feature map satisfying that $k(\mathbf{u}, \mathbf{v}) = \langle\varphi(\mathbf{u}), \varphi(\mathbf{v})\rangle_{\mathcal{H}}$.

The MMD statistic $\mathcal{D}_{\text{MMD}(k)}(U, V)$ for kernel k and two samples $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_{|U|}\}$

and $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_{|V|}\}$ is given by:

$$\mathcal{D}_{\mathrm{MMD}(k)}\left(U, V\right) = \frac{1}{|U|^2} \sum_{i,j=1}^{|U|} k(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{|V|^2} \sum_{i,j=1}^{|V|} k(\mathbf{v}_i, \mathbf{v}_j) - \frac{2}{|U||V|} \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} k(\mathbf{u}_i, \mathbf{v}_j). \quad (8)$$

MMD compares the similarities, described by the kernel k, within each sample with the similarities across samples and is commonly used in practice for two-sample testing. It is based on the idea that one can assign to each distribution $\mathcal{P}$ its embedding $\mu(\mathcal{P})$ into the RKHS $\mathcal{H}$, which is the unique element of $\mathcal{H}$ given by

$$\mu(\mathcal{P}) = \mathbb{E}_{\mathbf{Y} \sim \mathcal{P}}[\varphi(\mathbf{Y})]. \quad (9)$$

The MMD two-sample statistic (8) can then can then equivalently be written as the squared distance between the embeddings of the empirical distributions with respect to the RKHS norm $\|\cdot\|_{\mathcal{H}}$:

$$\mathcal{D}_{\mathrm{MMD}(k)}\left(U, V\right) = \left\| \mu\left(\frac{1}{|U|} \sum_{i=1}^{|U|} \delta_{\mathbf{u}_i}\right) - \mu\left(\frac{1}{|V|} \sum_{i=1}^{|V|} \delta_{\mathbf{v}_i}\right) \right\|_{\mathcal{H}}^2, \quad (10)$$

recalling that $\delta_{\mathbf{y}}$ is the point mass at $\mathbf{y}$.

As the sample sizes $|U|$ and $|V|$ grow, the MMD statistic (10) converges to its population version, which is the squared RKHS distance between the corresponding embeddings of the data-generating distributions of U and V. Since the embedding map μ is injective for characteristic kernel k, we see that MMD is able to detect any difference in the distribution. Even though the power of the MMD two sample test also deteriorates as the data dimensionality grows, since the testing problem becomes intrinsically harder (Reddi et al., 2014), it still has good empirical power compared to other multivariate two-sample tests for a wide range of k (Gretton et al., 2012a).

However, the $\mathcal{O}((|U| + |V|)^2)$ complexity for computing $\mathcal{D}_{\mathrm{MMD}(k)}(U, V)$ from (8) is too large for many applications. For that reason, several fast approximations of MMD have been suggested in the literature (Gretton et al., 2012a; Zaremba et al., 2013). As already mentioned, the complexity of the distributional metric $\mathcal{D}(\cdot, \cdot)$ used for DRF is crucial for the overall method to be computationally efficient, since the splitting step is used extensively in the forest construction. We therefore propose splitting based on an MMD statistic computed with an approximate kernel $\tilde{k}$, which is also a fast random approximation of the original MMD (Zhao and Meng, 2015).

Bochner's theorem (see e.g. Wendland (2004, Theorem 6.6)) gives us that any bounded shift-invariant kernel can be written as

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}), \quad (11)$$

i.e. as a Fourier transform of some measure ν. Therefore, by randomly sampling the frequency vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B$ from normalized ν, we can approximate our kernel k by another kernel $\tilde{k}$ (up to a scaling factor) as follows:

$$k(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T(\mathbf{u}-\mathbf{v})} d\nu(\boldsymbol{\omega}) \approx \frac{1}{B} \sum_{b=1}^{B} e^{i\boldsymbol{\omega}_b^T(\mathbf{u}-\mathbf{v})} = \tilde{k}(\mathbf{u}, \mathbf{v}),$$

where we define $\tilde{k}(\mathbf{u}, \mathbf{v}) = \langle \boldsymbol{\varphi}(\mathbf{u}), \boldsymbol{\varphi}(\mathbf{v}) \rangle_{\mathbb{C}^B}$ as the kernel function with the feature map given by

$$\boldsymbol{\varphi}(\mathbf{u}) = \frac{1}{\sqrt{B}} \left(\tilde{\varphi}_{\boldsymbol{\omega}_1}(\mathbf{u}), \ldots, \tilde{\varphi}_{\boldsymbol{\omega}_B}(\mathbf{u}) \right)^T = \frac{1}{\sqrt{B}} \left(e^{i\boldsymbol{\omega}_1^T\mathbf{u}}, \ldots, e^{i\boldsymbol{\omega}_B^T\mathbf{u}} \right)^T,$$

which is a random vector consisting of the Fourier features $\tilde{\varphi}_{\boldsymbol{\omega}}(\mathbf{u}) = e^{i\boldsymbol{\omega}^T\mathbf{u}} \in \mathbb{C}$ (Rahimi and Recht, 2008). Such kernel approximations are frequently used in practice for computational efficiency (Rahimi and Recht, 2009; Le et al., 2013). As a default choice of k we take the Gaussian kernel with bandwidth σ, since in this case we have a convenient expression for the measure ν and we sample $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_B \sim N_d(\mathbf{0}, \sigma^{-2} I_d)$. The bandwidth σ is chosen as the median pairwise distance between all training responses $\{\mathbf{y}_i\}_{i=1}^n$, commonly referred to as the 'median heuristic' (Gretton et al., 2012b).

From the representation of MMD via the distribution embeddings (10) we can obtain that MMD two-sample test statistic $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ using the approximate kernel $\tilde{k}$ is given by

$$\mathcal{D}_{\mathrm{MMD}(\tilde{k})} \left(\{\mathbf{u}_i\}_{i=1}^{|U|}, \{\mathbf{v}_i\}_{i=1}^{|V|} \right) = \frac{1}{B} \sum_{b=1}^{B} \left| \frac{1}{|U|} \sum_{i=1}^{|U|} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{u}_i) - \frac{1}{|V|} \sum_{i=1}^{|V|} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{v}_i) \right|^2.$$

Interestingly, $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ is not only an MMD statistic on its own, but can also be viewed as a random approximation of the original MMD statistic $\mathcal{D}_{\mathrm{MMD}(k)}$ (8) using kernel k; by using the kernel representation (11), it can be written (the derivation can be found in the Appendix B of Ćevid et al. (2020b)) as

$$\mathcal{D}_{\mathrm{MMD}(k)} \left(\{\mathbf{u}_i\}_{i=1}^{|U|}, \{\mathbf{v}_i\}_{i=1}^{|V|} \right) = \int_{\mathbb{R}^d} \left| \frac{1}{|U|} \sum_{i=1}^{|U|} \tilde{\varphi}_{\boldsymbol{\omega}}(\mathbf{u}_i) - \frac{1}{|V|} \sum_{i=1}^{|V|} \tilde{\varphi}_{\boldsymbol{\omega}}(\mathbf{v}_i) \right|^2 d\nu(\boldsymbol{\omega}).$$

Finally, our DRF splitting criterion $\mathcal{D}(\cdot, \cdot)$ (1) is then taken to be the (scaled) MMD statistic $\frac{n_L n_R}{n_P^2} \mathcal{D}_{\mathrm{MMD}(\tilde{k})} \left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\} \right)$ with the approximate random kernel $\tilde{k}$ used instead of k, which can thus be conveniently written as:

$$\frac{1}{B} \sum_{b=1}^{B} \frac{n_L n_R}{n_P^2} \left| \frac{1}{n_L} \sum_{\mathbf{x}_i \in C_L} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{y}_i) - \frac{1}{n_R} \sum_{\mathbf{x}_i \in C_R} \tilde{\varphi}_{\boldsymbol{\omega}_b}(\mathbf{y}_i) \right|^2, \tag{12}$$

where we recall that $n_P = |\{i \mid \mathbf{x}_i \in P\}|$ and n_L, n_R are defined analogously. The additional scaling factor $\frac{n_L n_R}{n_P^2}$ in (12) occurs naturally and compensates the increased variance of the test statistic for unbalanced splits; it also appears in the GRF (2) and CART (see representation (13)) splitting criteria.

The main advantage of the splitting criterion based on $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ is that by using the representation (1) it can be easily computed for every possible splitting level l in $\mathcal{O}(Bn_P)$ complexity, whereas the MMD statistic $\mathcal{D}_{\mathrm{MMD}(k)}$ using kernel k would require $\mathcal{O}(n_P^2)$ computational steps, which makes the overall complexity of the algorithm $\mathcal{O}\left(B \times N \times \mathrm{mtry} \times n \log n\right)$ instead of much slower $\mathcal{O}\left(N \times \mathrm{mtry} \times n^2\right)$.

We do not use the same approximate random kernel $\tilde{k}$ for different splits; for every parent node P we resample the frequency vectors $\{\omega_b\}_{b=1}^B$ defining the corresponding feature map $\tilde{\varphi}$. Using different $\tilde{k}$ at each node might help to better detect different distributional changes. Furthermore, having different random kernels for each node agrees well with the randomness of the Random Forests and helps making the trees more independent. Since the MMD statistic $\mathcal{D}_{\mathrm{MMD}(\tilde{k})}$ used for our splitting criterion is not only an approximation of $\mathcal{D}_{\mathrm{MMD}(k)}$, but is itself an MMD statistic, it inherits good power for detecting any difference in distribution of $\mathbf{Y}$ in the child nodes for moderately large data dimensionality d, even when B is reasonably small. One could even consider changing the number of random Fourier features B at different levels of the tree, as n_P varies, but for simplicity we take it to be fixed.

There is some similarity of our MMD-based splitting criterion (12) with the standard variance reduction CART splitting criterion (7) when $d = 1$, which can be rewritten as:

$$\frac{n_L n_R}{n_P^2}\left(\frac{1}{n_L}\sum_{\mathbf{x}_i \in C_L} y_i - \frac{1}{n_R}\sum_{\mathbf{x}_i \in C_R} y_i\right)^2. \tag{13}$$

The derivation can be found in Appendix B of Ćevid et al. (2020b). From this representation, we see that the CART splitting criterion (7) is also equivalent to the GRF splitting criterion (2) when our target is the univariate conditional mean $\tau(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X}{=}\mathbf{x}]$ which is estimated for C_L and C_R by the sample means $\hat{\tau}_L = \overline{y}_L$ and $\hat{\tau}_R = \overline{y}_R$. Therefore, as it compares the means of the univariate response Y in the child nodes, the CART criterion can only detect changes in the response mean well, which is sufficient for prediction of Y from $\mathbf{X}$, but might not be suitable for more complex targets. Similarly, for multivariate applications, aggregating the marginal CART criteria (Kocev et al., 2007; Segal and Xiao, 2011) across different components Y_i of the response can only detect changes in the means of their marginal distributions. However, it is possible in the multivariate case that the pairwise correlations or the variances of the responses change, while the marginal means stay (almost) constant. For an illustration on simulated data, see Figure 7. Additionally, aggregating the splitting criteria over d components of the response $\mathbf{Y}$ can reduce the signal size if only the distribution of a few components change. Our MMD-based splitting criterion (12) is able to avoid such difficulties as it implicitly inspects all aspects of the multivariate response distribution.

If we take a trivial kernel $k_{\mathrm{id}}(y_i, y_j) = y_i y_j$ with the identity feature map $\varphi_{\mathrm{id}}(y) = y$, the corresponding distributional embedding (9) is given by $\mu(\mathcal{P}) = \mathbb{E}_{Y\sim\mathcal{P}}Y$ and thus the corresponding splitting criterion based on $\mathcal{D}_{\mathrm{MMD}(k_{\mathrm{id}})}$ (10) corresponds exactly to the CART splitting criterion (7), which can be seen from its equivalent representation (13). Interestingly, Theorem 3 in Section 3 below shows that the MMD splitting criterion with kernel k can also be viewed as the abstract CART criterion in the RKHS $\mathcal{H}$ corresponding to k (Fan et al., 2010). Moreover, it is also shown that DRF with the MMD splitting criterion can thus be viewed asymptotically as a greedy minimization of the squared RKHS distance between the corresponding embeddings of our estimate $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ and the truth $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ averaged over $\mathbf{x}$, thus justifying the proposed method. In Section 3, we exploit this relationship to derive interesting theoretical properties of DRF with the MMD splitting criterion.

3 Theoretical Results

In this section we exploit the properties of kernel mean embedding in order to relate DRF with the MMD splitting criterion to an abstract version of the standard Random Forest with the CART splitting criterion (Breiman, 2001) when the response is taking values in the corresponding RKHS. We further exploit this relationship to adapt the existing theoretical results from the Random Forest literature to show that our estimate (5) of the conditional distribution of the response is consistent with respect to the MMD metric for probability measures and with a good rate. Finally, we show that this implies consistency of the induced DRF estimates for a range interesting targets $\tau(\mathbf{x})$, such as conditional CDFs or quantiles. The proofs of all results can be found in the Appendix B in Ćevid et al. (2020b).

Recalling the notation from above, let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the Reproducing kernel Hilbert space induced by the positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and let $\varphi : \mathbb{R}^d \to \mathcal{H}$ be its corresponding feature map. The kernel embedding function $\mu : \mathcal{M}_b(\mathbb{R}^d) \to \mathcal{H}$ maps any bounded signed Borel measure $\mathcal{P}$ on $\mathbb{R}^d$ to an element $\mu(\mathcal{P}) \in \mathcal{H}$ defined by

$$\mu(\mathcal{P}) = \int_{\mathbb{R}^d} \varphi(\mathbf{y}) d\mathcal{P}(\mathbf{y}),$$

see (9). Boundedness of k ensures that μ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$, while continuity of k ensures that $\mathcal{H}$ is separable Hsing and Eubank (2015).

By considering the kernel embedding $\mu(\cdot)$ and using its linearity, we can write the embedding of the distributional estimate $\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}))$ of DRF (5) as the average of the embeddings of the empirical distributions of $\mathbf{Y}$ in the leaves containing $\mathbf{x}$ over all trees:

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})) = \frac{1}{N}\sum_{k=1}^{N} \mu\left(\frac{1}{|\mathcal{L}_k(\mathbf{x})|}\sum_{\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})} \delta_{\mathbf{y}_i}\right) = \frac{1}{N}\sum_{k=1}^{N} \frac{1}{|\mathcal{L}_k(\mathbf{x})|}\sum_{\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x})} \mu(\delta_{\mathbf{y}_i}). \tag{14}$$

This is analogous to the prediction of the response for the standard Random Forest, but where we average the embeddings $\mu(\delta_{\mathbf{y}_i}) = \varphi(\mathbf{y}_i)$ instead of the response values themselves. Therefore, by using the kernel embedding, we can shift the perspective to the RKHS $\mathcal{H}$ and view DRF as the analogue of the original Random Forest for estimation of $\mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) = \mathbb{E}[\varphi(\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}]$ in some abstract Hilbert space.

With this viewpoint, we can relate the MMD splitting criterion to the original CART criterion (7), which measures the mean squared prediction error for splitting a certain parent node P into children C_L and C_R. On one hand, from Equation (13) we see that the CART criterion measures the squared distance between the response averages $\frac{1}{n_L}\sum_{\mathbf{x}_i \in C_L} y_i$ and $\frac{1}{n_R}\sum_{\mathbf{x}_i \in C_R} y_i$ in the child nodes, but on the other hand, Equation (10) shows that the MMD splitting criterion measures the RKHS distance between the embeddings of the empirical response distributions in C_L and C_R. This is summarized in the following theorem, which not only shows that the MMD splitting criterion can be viewed as the abstract CART criterion in the RKHS $\mathcal{H}$ (Fan et al., 2010), but also that DRF with the MMD splitting criterion can be viewed as greedy minimization of the average squared distance between the estimated and true conditional distributions, as measured by the RKHS norm between the corresponding embeddings to $\mathcal{H}$:

Theorem 3. *For any split of a parent node P into child nodes C_L and C_R, let $\hat{\mathbb{P}}_{split}(\mathbf{x}) = \sum_{j \in \{L,R\}} \mathbb{1}(\mathbf{x} \in C_j)\frac{1}{n_j}\sum_{\mathbf{x}_i \in C_j} \delta_{\boldsymbol{y}_i}$ denote the resulting estimate of the distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$ when $\mathbf{x} \in P$. Then the MMD splitting criterion is equivalent to the abstract version of the CART criterion (7) on $\mathcal{H}$:*

$$\arg\max_{split} \frac{n_L n_R}{n_P^2}\mathcal{D}_{MMD(k)}\left(\{\mathbf{y}_i \mid \mathbf{x}_i \in C_L\}, \{\mathbf{y}_i \mid \mathbf{x}_i \in C_R\}\right) = \arg\min_{split} \frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}_{split}(\mathbf{x}_i))\right\|_{\mathcal{H}}^2.$$

Moreover, for any node P and any fixed distributional estimator $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$, we have:

$$\frac{1}{n_P}\sum_{\mathbf{x}_i \in P}\left\|\mu(\delta_{\mathbf{y}_i}) - \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}_i))\right\|_{\mathcal{H}}^2 = V_P + \mathbb{E}\left[\|\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X})) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P\right] + \mathcal{O}_p(n^{-1/2}),$$

where $V_P = \mathbb{E}\left[\|\mu(\delta_{\mathbf{Y}}) - \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}))\|_{\mathcal{H}}^2 \mid \mathbf{X} \in P\right]$ is a deterministic term not depending on the estimates $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$.

In conclusion, DRF with the MMD splitting criterion can be viewed as the standard Random Forest with the CART splitting criterion, but with the response $\mu(\delta_{\mathbf{Y}})$ taking values in an abstract RKHS $\mathcal{H}$ instead of $\mathbb{R}$. Therefore, one could in principle derive properties of DRF by adapting any theoretical result for standard Random Forests from the literature. However, a lot of care is needed for making the results rigorous in this abstract setup, as many useful properties of $\mathbb{R}$ need not hold for infinite-dimensional $\mathcal{H}$. The remaining part of this section is inspired by the results from Wager and Athey (2018).

We suppose that the forest construction satisfies the following properties, which significantly facilitate the theoretical considerations of the method and ensure that our forest estimator is well behaved, as stated in Wager and Athey (2018):

- **(P1)** (*Data sampling*) The bootstrap sampling with replacement, usually used in forest-based methods, is replaced by a subsampling step, where for each tree we choose a random subset of size s_n out of n training data points. We consider s_n going to infinity with n, with the rate specified below.
- **(P2)** (*Honesty*) The data used for constructing each tree is split into two parts; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response.
- **(P3)** (*α-regularity*) Each split leaves at least a fraction $\alpha \leqslant 0.2$ of the available training sample on each side. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ observations, for some fixed tuning parameter $\kappa \in \mathbb{N}$.
- **(P4)** (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.
- **(P5)** (*Random-split*) At every split point, the probability that the split occurs along the feature X_j is bounded below by π/p, for some $\pi > 0$ and for all $j = 1, \dots, p$.

The validity of the above properties are easily ensured by the forest construction. For more details, see Appendix A of Ćevid et al. (2020b).

From Equation (14), the prediction of DRF for a given test point $\mathbf{x}$ can be viewed as an element of $\mathcal{H}$. If we denote the i-th training observation by $\mathbf{Z}_i = (\mathbf{x}_i, \mu(\delta_{\mathbf{y}_i})) \in \mathbb{R}^p \times \mathcal{H}$, then by (14) we estimate the embedding of the true conditional distribution $\mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}))$ by the average of the corresponding estimates per tree:

$$\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})) = \frac{1}{N} \sum_{j=1}^{N} T(\mathbf{x}; \varepsilon_j, \mathcal{Z}_j),$$

where $\mathcal{Z}_k$ is a random subset of $\{\mathbf{Z}_i\}_{i=1}^n$ of size s_n chosen for constructing the j-th tree $\mathcal{T}_j$ and ε_j is a random variable capturing all randomness in growing $\mathcal{T}_j$, such as the choice of the splitting candidates. $T(\mathbf{x}; \varepsilon, \mathcal{Z})$ denotes the output of a single tree: i.e. the average of the terms $\mu(\delta_{\mathbf{Y}_i})$ over all data points $\mathbf{Z}_i$ contained in the leaf $\mathcal{L}(\mathbf{x})$ of the tree constructed from ε and $\mathcal{Z}$.

Since one can take the number of trees N to be arbitrarily large, we consider an "idealized" version of our estimator, as done in Wager and Athey (2017), which we denote as $\hat{\mu}_n(\mathbf{x})$:

$$\hat{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1 < i_2 < \ldots < i_{s_n}} \mathbb{E}_\varepsilon\, T(\mathbf{x}; \varepsilon; \{\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}\}), \tag{15}$$

where the sum is taken over all $\binom{n}{s_n}$ possible subsets of $\{\mathbf{Z}_i\}_{i=1}^n$. We have that $\mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})) \to \hat{\mu}_n(\mathbf{x})$ as $N \to \infty$, while keeping the other variables constant, and thus we assume for simplicity that those two quantities are the same.

Our main result shows that, under similar assumptions as in Wager and Athey (2017), the embedding of our conditional distribution estimator $\mu_n(\mathbf{x}) = \mu(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}))$ consistently estimates $\mu(\mathbf{x}) := \mu(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}))$ with respect to the RKHS norm with a certain rate:

Theorem 4. *Suppose that our forest construction satisfies properties* ***(P1)**–**(P5)***. *Assume additionally that k is a bounded and continuous kernel and that we have a random design with $\mathbf{X}_1, \ldots, \mathbf{X}_n$ independent and identically distributed on $[0,1]^p$ with a density bounded away from 0 and infinity. If the subsample size s_n is of order n^β for some $0 < \beta < 1$, the mapping*

$$\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[\mu(\delta_\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}] \in \mathcal{H},$$

is Lipschitz and $\sup_{\mathbf{x} \in [0,1]^p} \mathbb{E}[\|\mu(\delta_\mathbf{Y})\|_\mathcal{H}^2 \mid \mathbf{X}{=}\mathbf{x}] < \infty$, we obtain the consistency w.r.t. the RKHS norm:

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_\mathcal{H} = o_p\left(n^{-\gamma}\right), \tag{16}$$

for any $\gamma < \frac{1}{2}\min\left(1 - \beta, \frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p} \cdot \beta\right)$.

Remark. *The rate in* (16) *is analogous to the one from Wager and Athey (2018), who used it further to derive the asymptotic normality of the random forest estimator in $\mathbb{R}$. Indeed, one can show in our case that there exists a sequence of real numbers $\sigma_n \to 0$, such that $(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))/\sigma_n$ as a random element of $\mathcal{H}$ is "asymptotically linear", in the sense that it is indistinguishable asymptotically from an average of independent random elements in $\mathcal{H}$. Unfortunately, this alone is not enough to establish asymptotic normality of $(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))/\sigma_n$ as an element of $\mathcal{H}$, a task left for future research.*

The above result shows that DRF estimate $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$ converges fast to the truth $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$ in the MMD distance, i.e. the RKHS distance between the corresponding embeddings. Even though this is interesting on its own, ultimately we want to relate this result to estimation of certain distributional targets $\tau(\mathbf{x}) = \tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}))$.

For any $f \in \mathcal{H}$, we have that the DRF estimate of the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}]$ equals the dot product $\langle f, \hat{\mu}_n(\mathbf{x})\rangle_{\mathcal{H}}$ in the RKHS:

$$\langle f, \hat{\mu}_n(\mathbf{x})\rangle_{\mathcal{H}} = \left\langle f, \int_{\mathbb{R}^d} \varphi(\mathbf{y}) d\hat{\mathbb{P}}(\mathbf{y} \mid \mathbf{X}{=}\mathbf{x}) \right\rangle_{\mathcal{H}} = \int_{\mathbb{R}^d} f(\mathbf{y})\, d\hat{\mathbb{P}}(\mathbf{y} \mid \mathbf{X}{=}\mathbf{x}) = \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i),$$

where we recall the weighting function $w_{\mathbf{x}}(\cdot)$ induced by the forest (4). Therefore, the consistency result (16) in Theorem 4 directly implies that

$$\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) = \langle f, \hat{\mu}_n(\mathbf{x})\rangle_{\mathcal{H}} \xrightarrow{p} \langle f, \mu(\mathbf{x})\rangle_{\mathcal{H}} = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}] \quad \text{for any } f \in \mathcal{H}, \tag{17}$$

i.e. that the DRF consistently estimates the targets of the form $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}]$, for $f \in \mathcal{H}$. From (16) we also obtain the rate of convergence when $s_n \asymp n^{\beta}$:

$$\left| \sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) - \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}] \right| = o_p\left(n^{-\gamma} \|f\|_{\mathcal{H}}\right),$$

for γ as in Theorem 4. When k is continuous, it is well known that all elements of $\mathcal{H}$ are continuous, see e.g. Hsing and Eubank (2015). Under certain assumptions on the kernel and its input space, holding for several popular kernels, (e.g. the Gaussian kernel) (Sriperumbudur, 2016), we can generalize the convergence result (17) to any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$, as the convergence of measures $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x}) \to \mathbb{P}(\mathbf{Y} \mid \mathbf{X}{=}\mathbf{x})$ in the MMD metric will also imply their weak convergence, i.e. k metrizes weak convergence (Sriperumbudur, 2016; Simon-Gabriel and Schölkopf, 2018; Simon-Gabriel et al., 2020):

Corollary 3.1. Assume that one of the following two sets of conditions holds:

(a) The kernel k is bounded, (jointly) continuous and has

$$\int\int k(\mathbf{x}, \mathbf{y}) d\mathcal{P}(\mathbf{x}) d\mathcal{P}(\mathbf{y}) > 0 \quad \forall \mathcal{P} \in \mathcal{M}_b(\mathbb{R}^d)\backslash\{0\}. \tag{18}$$

Moreover, $\mathbf{y} \mapsto k(\mathbf{y}_0, \mathbf{y})$ is vanishing at infinity, for all $\mathbf{y}_0 \in \mathbb{R}^d$.

(b) The kernel k is bounded, shift-invariant, (jointly) continuous and ν in the Bochner representation in (11) is supported on all of $\mathbb{R}^d$. Moreover, $\mathbf{Y}$ takes its values almost surely in a closed and bounded subset of $\mathbb{R}^d$.

Then, under the conditions of Theorem 4, we have for any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that DRF consistently estimates the target $\tau(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}]$ for any $\mathbf{x} \in [0, 1]^p$:

$$\sum_{i=1}^{n} w_{\mathbf{x}}(\mathbf{x}_i) f(\mathbf{y}_i) \xrightarrow{p} \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}{=}\mathbf{x}].$$

Recalling the Portmanteau Lemma on separable metric spaces, see e.g. Dudley (2002, Chapter 11), this has several other interesting consequences, such as the consistency of CDF and quantile estimates; Let $F_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}(\cdot)$ be the conditional CDF of $\mathbf{Y}$ and for any index $1 \leqslant i \leqslant d$, let $F_{Y_i\,|\,\mathbf{X}=\mathbf{x}}(\cdot)$ be the conditional CDF of Y_i and $F^{-1}_{Y_i\,|\,\mathbf{X}=\mathbf{x}}(\cdot)$ its generalized inverse, i.e. the quantile function. Let $\hat{F}_{Y_i\,|\,\mathbf{X}=\mathbf{x}}(\cdot)$ and $\hat{F}^{-1}_{Y_i\,|\,\mathbf{X}=\mathbf{x}}(\cdot)$ be the corresponding DRF estimates via weighting function (6). Then we have the following result:

Corollary 3.2. Under the conditions of Corollary 3.1, we have

$$\hat{F}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}(\mathbf{t}) \xrightarrow{p} F_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}(\mathbf{t})$$
$$\hat{F}^{-1}_{Y_i\,|\,\mathbf{X}=\mathbf{x}}(t) \xrightarrow{p} F^{-1}_{Y_i\,|\,\mathbf{X}=\mathbf{x}}(t)$$

for all points of continuity $\mathbf{t} \in \mathbb{R}^d$ and $t \in \mathbb{R}$ of $F_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}(\cdot)$ and $F^{-1}_{Y_i\,|\,\mathbf{X}=\mathbf{x}}(\cdot)$ respectively.

4 Applications and Numerical Experiments

The goal of this section is to demonstrate the versatility and applicability of DRF for many practical problems. We show that DRF can be used not only as an estimator of the multivariate conditional distribution, but also as a two-step method to easily obtain out-of-the box estimators for various, and potentially complex, targets $\tau(\mathbf{x})$.

Our main focus lies on the more complicated targets which cannot be that straightforwardly approached by conventional methods. However, we also illustrate the usage of DRF for certain applications for which there already exist several well-established methods. Whenever possible in such cases, we compare the performance of DRF with the specialized, task-specific methods to show that, despite its generality, there is at most a very small loss of precision. However, we should point out that for many targets such as, that can not be written in a form of a conditional mean or a conditional quantile, for example, conditional correlation, direct comparison of the accuracy is not possible for real data, since no suitable loss function exists and the ground truth is unknown. Finally, we show that, in addition to directly estimating certain targets, DRF can also be a very useful tool for many different applications, such as causality and fairness.

Detailed descriptions of all data sets and the corresponding analyses, together with additional simulations can be found in the Appendix C of Ćevid et al. (2020b).

4.1 Estimation of Conditional Multivariate Distributions

In order to provide good estimates for any target $\tau(\mathbf{x}) = \tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x}))$, our method needs to estimate the conditional multivariate distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$ well. Therefore, we first investigate here the accuracy of the DRF estimate (5) of the full conditional distribution and compare its performance with the performance of several existing methods.

There are not many algorithms in the literature that nonparametrically estimate the multivariate conditional distribution $\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})$. In addition to a few simple methods such as the k-nearest neighbors or the kernel regression, which locally weight the training points, we also consider advanced machine learning methods such as the Conditional Generative Adversarial Network (CGAN) (Mirza and Osindero, 2014; Aggarwal et al., 2019), Conditional Variational Autoencoder (CVAE) (Sohn et al., 2015) and Masked Autoregressive Flow (Papamakarios et al., 2017). It is worth mentioning that the focus in the machine learning literature has been more on applications where d is very large (e.g. pixels of an image) and p is very small (such as image labels). Even though some methods do not provide the estimated conditional distribution in a form as simple as DRF, one is still able to sample from the estimated distribution and thus perform any subsequent analysis and make fair comparisons between the methods.

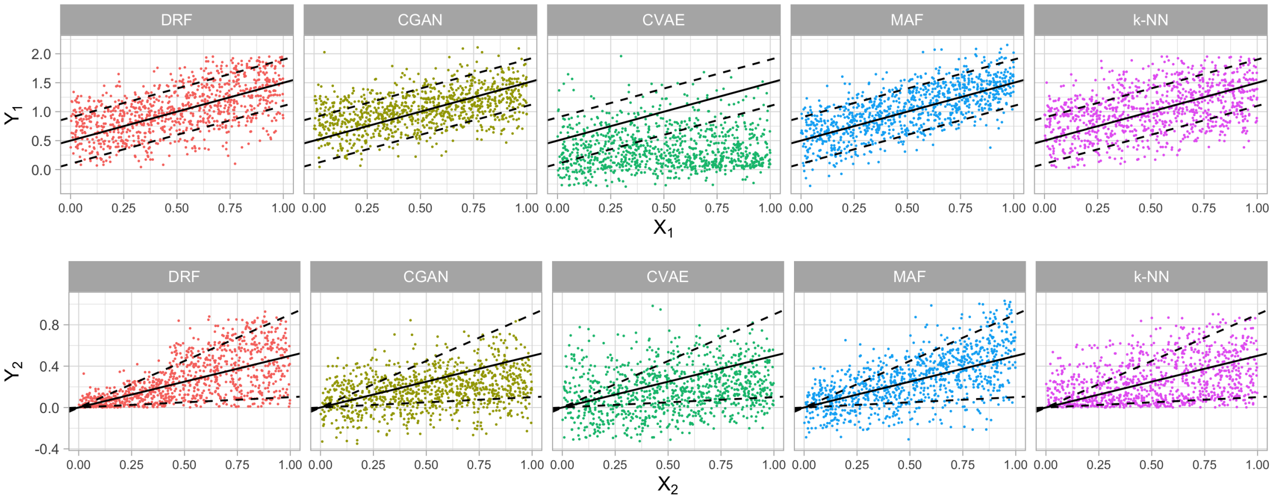


Figure 3: The illustration of the estimated joint conditional distribution obtained by different methods for the toy example (19). For 1000 randomly generated test points $\mathbf{X}_{\text{test}} \sim U(0,1)^p$ the top row shows the estimated distribution of the response component Y_1, whereas the bottom row shows the estimated distribution of Y_2. The 0.1 and 0.9 quantiles of the true conditional distribution are indicated by a dashed black line, whereas the conditional mean is shown as a black solid line.

We first illustrate the estimated distributions of the above method on a toy example where $n = 1000, p = 10, d = 2$ and

$$Y_1 \perp\!\!\!\perp Y_2 \mid \mathbf{X}{=}\mathbf{x}, \quad Y_1 \mid \mathbf{X}{=}\mathbf{x} \sim U(x_1, x_1+1), \quad Y_2 \mid \mathbf{X}{=}\mathbf{x} \sim U(0, x_2), \quad \mathbf{X} \sim U(0,1)^p. \tag{19}$$

That is, in the above example X_1 affects the mean of Y_1, whereas X_2 affects the both mean and variance of Y_2, and $X_3, \ldots, X_p$ have no impact. The results can be seen in Figure 3 for the above methods. We see that, unlike some other methods, DRF is able to balance the importance of the predictors X_1 and X_2 and thus to estimate the distributions of Y_1 and Y_2 well.

One can do a more extensive comparison on real data sets. We use the benchmark data sets from the multi-target regression literature (Tsoumakas et al., 2011) together with some additional ones created from the real data sets described throughout this paper. The performance of DRF is compared with the performance of other existing methods for nonparametric estimation of multivariate distributions by using the Negative Log Predictive Density (NLPD) loss, which evaluates the logarithm of the induced multivariate density estimate (Quinonero-Candela et al., 2005). As the number of test points grows to infinity, NLPD loss becomes equivalent to the average KL divergence between the estimated and the true conditional distribution and is thus able to capture how well one estimates the whole distribution instead of only its mean.

In addition to the methods mentioned above, we also include the methods that are intended only for mean prediction, by assuming that the distribution of the response around its mean is homogeneous, i.e. that the conditional distribution $\mathbb{P}\left(\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}] \mid \mathbf{X}=\mathbf{x}\right)$ does not depend on $\mathbf{x}$. This is fitted by regressing each component of $\mathbf{Y}$ separately on $\mathbf{X}$ and using the pooled residuals. We consider the standard nonparametric regression methods such as Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), and Deep Neural Networks (Goodfellow et al., 2016).

The results are shown in Table 1. We see that DRF performs well for a wide range of sample size and problem dimensionality, especially in problems where p is large and d is moderately big. It does so without the need for any tuning or involved numerical optimization. More detailed analysis and descriptions of each competing method and the loss function can be found in the Appendix C of Ćevid et al. (2020b).

4.2 Estimation of Statistical Functionals

Because DRF represents the estimated conditional distribution $\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}) = \sum_i w_{\mathbf{x}}(\mathbf{x}_i) \cdot \delta_{\mathbf{y}_i}$ in a convenient form by using weights $w_{\mathbf{x}}(\mathbf{x}_i)$, a plug-in estimator $\tau(\hat{\mathbb{P}}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}))$ of many common real valued statistical functionals $\tau(\mathbb{P}(\mathbf{Y} \mid \mathbf{X} = \mathbf{x})) \in \mathbb{R}$ can be easily constructed from $w_{\mathbf{x}}(\cdot)$.

We first investigate the performance for the classical problem of univariate quantile estimation on simulated data. We consider the following three data generating mechanisms with $p = 40, n = 2000$ and $\mathbf{X}_i \overset{i.i.d.}{\sim} U(-1,1)^p$:

- Scenario 1: $Y \sim N(0.8 \cdot \mathbb{1}(X_1 > 0), 1)$ (mean shift based on X_1)
- Scenario 2: $Y \sim N(0, (1 + \mathbb{1}(X_1 > 0))^2)$ (variance shift based on X_1)

	jura	slump	wq	enb	atp1d	atp7d	scpf	sf1	sf2	copula	wage	births1	births2	air
n	359	103	1K	768	337	296	143	323	1K	5K	10K	10K	10K	10K
p	15	7	16	8	370	370	8	21	22	10	73	23	24	15
d	3	3	14	2	6	6	3	3	6	2	2	2	4	6
DRF	**3.9**	**4.0**	**22.5**	2.1	7.3	**7.0**	**2.0**	**-24.2**	**-24.3**	**2.8**	**2.8**	2.5	**4.2**	**8.5**
CGAN	10.8	5.3	27.3	3.5	10.4	363	4.8	9.8	21.1	5.8	360	**2.4**	>1K	11.8
CVAE	4.8	37.8	36.8	2.6	>1K	>1K	108.8	8.6	>1K	2.9	>1K	>1K	49.7	9.6
MAF	4.6	4.5	23.9	3.0	8.0	8.1	2.6	4.7	3.8	2.9	3.0	2.5	>1K	8.5
k-NN	4.5	5.0	23.4	2.4	8.8	8.6	4.1	-22.4	-19.7	2.9	**2.8**	2.7	4.4	8.8
kernel	4.1	4.2	23.0	**2.0**	**6.6**	7.1	2.9	-23.0	-20.6	2.8	2.9	2.6	4.3	8.4
RF	7.1	12.1	35.2	5.7	12.7	13.3	16.7	3.9	2.2	5.8	6.1	5.0	8.3	13.9
XGBoost	11.4	38.3	25.9	3.0	>1K	>1K	>1K	0.3	1.6	3.5	2.9	>1K	>1K	12.8
DNN	4.0	4.2	23.3	2.6	8.6	8.7	2.6	2.3	2.2	2.9	3.0	2.6	5.4	8.6

Table 1: NLPD loss computed on out-of-sample observations for the estimated conditional distributions obtained by several different methods (corresponding to rows) for many real data sets (corresponding to columns). The best method is indicated in bold. Detailed description of both the data sets and the competing methods can be found in Appendix C of Ćevid et al. (2020b)

- Scenario 3: $Y \sim \mathbb{1}(X_1 \leqslant 0) \cdot N(1,1) + \mathbb{1}(X_1 > 0) \cdot \mathrm{Exp}(1)$ (distribution shift based on X_1, constant mean and variance)

The first two scenarios correspond exactly to the examples given in Athey et al. (2019).

In Figure 4 we can see the corresponding estimates of the conditional quantiles for DRF, Quantile Regression Forest (QRF) (Meinshausen, 2006), which uses the same forest construction with CART splitting criterion as the original Random Forest (Breiman, 2001) but estimates the quantiles from the induced weighting function, Generalized Random Forests (GRF) (Athey et al., 2019) with a splitting criterion specifically designed for quantile estimation and Transformation Forests (TRF) (Hothorn and Zeileis, 2021). We see that DRF is performing very well even compared to methods that are specifically tailored to quantile estimation. For more detailed analysis and some additional examples, such as the univariate mean regression, we refer the reader to Appendix D in Ćevid et al. (2020b).

The multivariate setting is however more interesting, as one can use DRF to compute much more interesting statistical functionals $\tau(\mathbf{x})$. We illustrate this in Figure 5 for the the air quality data set, described in Section 2.2. The left plot shows one value of the estimated multivariate CDF, specifically the estimated probability of the event that the air quality index (AQI) is at most 50 at a given test site. This corresponds to the "Good"

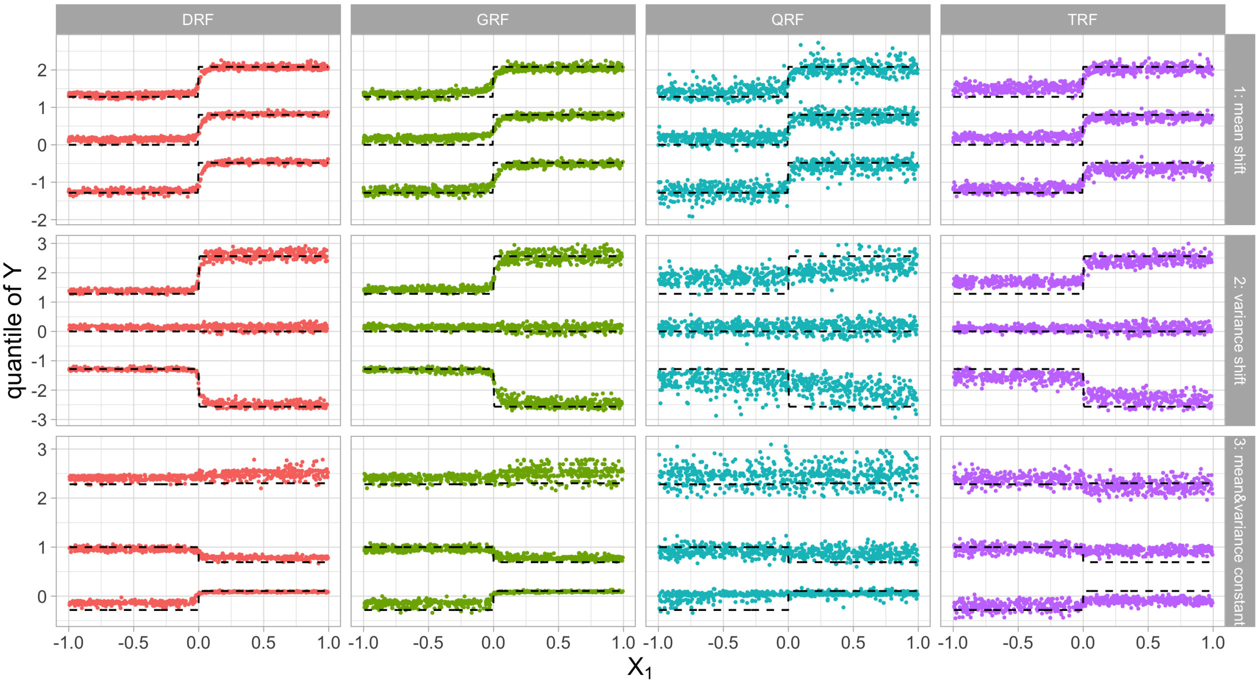


Figure 4: Scatter plot of predictions of the 0.1, 0.5 and 0.9 quantiles against X_1 for randomly generated 500 test data points $\mathbf{X}_{\text{test}} \sim U(-1, 1)^p$. The true values of the quantiles are displayed by black dashed lines. The columns corresponds to different methods DRF (red), GRF (green), QRF (blue), TRF (purple). The rows correspond to different simulation scenarios. The first two are taken from Athey et al. (2019).

category and means that the amount of every air pollutant is below a certain threshold determined by the EPA. Such probability estimates can be easily obtained by summing the weights of the training points belonging to the event of interest.

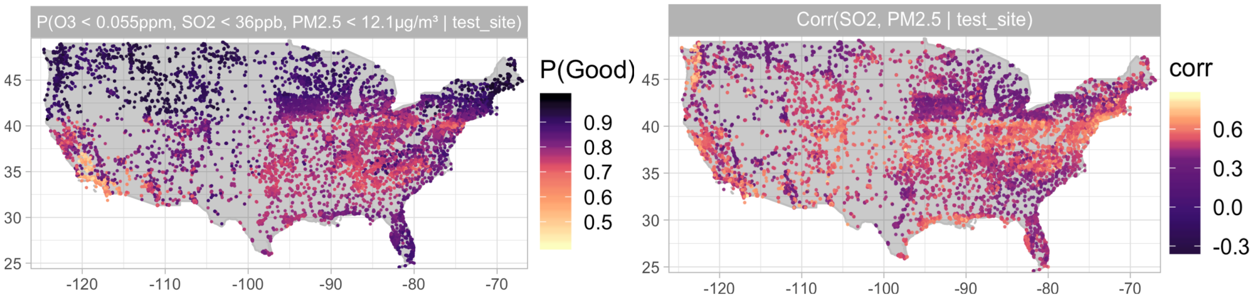

P(O3 < 0.055ppm, SO2 < 36ppb, PM2.5 < 12.1µg/m³ | test_site)
P(Good)
0.9
0.8
0.7
0.6
0.5
Corr(SO2, PM2.5 | test_site)
corr
0.6
0.3
0.0
-0.3
45
40
35
30
25
-120
-110
-100
-90
-80
-70

Figure 5: Estimates of the probability $\mathbb{P}(\text{AQI} \leqslant 50 \mid \text{test site})$ (left) and the conditional correlation (right) derived from the DRF estimate of the multivariate conditional distribution.

In order to investigate the accuracy of the conditional CDF obtained by DRF, we compare the estimated probabilities with estimates of the standard univariate classification forest (Breiman, 2001) with the response $\mathbb{1}(\text{AQI} \leqslant 50)$. In the left plot of Figure 6, we can see that the DRF estimates of the $\mathbb{P}(\text{AQI} \leqslant 50 \mid \mathbf{X} = \mathbf{x})$ (also visualized in Figure 5) are quite similar to the estimates of the classification forest predicting the outcome

$\mathbb{1}(\text{AQI} \leqslant 50)$. Furthermore, the cross-entropy loss evaluated on the held-out measurements equals 0.4671 and 0.4663 respectively, showing almost no loss of precision. In general, estimating the simple functionals from the weights provided by DRF comes usually at a small to no loss compared to the classical methods specifically designed for this task.

In addition to the classical functionals $\tau(\mathbf{x})$ in the form of an expectation $\mathbb{E}(f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x})$ or a quantile $Q_\alpha(f(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x})$ for some function $f : \mathbb{R}^d \to \mathbb{R}$, which can also be computed by solving the corresponding one-dimensional problems, additional interesting statistical functionals with intrinsically multivariate nature that are not that simple to estimate directly are accessible by DRF, such as, for example, the conditional correlations $\mathrm{Cor}(Y_i, Y_j \mid \mathbf{X}=\mathbf{x})$. As an illustration, the estimated correlation of the sulfur dioxide (SO_2) and fine particulate matter (PM2.5) is shown in the right plot of Figure 5. The plot reveals also that the local correlation in many big cities is slightly larger than in its surroundings, which can be explained by the fact that the industrial production directly affects the levels of both pollutants.

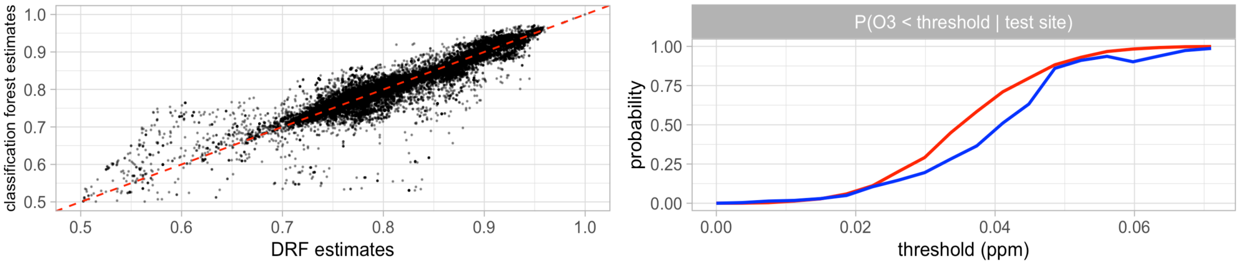


Figure 6: Left: Comparison of the CDF estimates obtained by DRF (displayed also in the left plot of Figure 5) and by the classification forest. Right: Example how the CDF estimated by using the classification forest (blue) need not be monotone, whereas the DRF estimates (red) are well-behaved.

A big advantage of the target-free forest construction of DRF is that all subsequent targets are computed from same the weighting function $w_{\mathbf{x}}$ obtained from a single forest fit. First, this is computationally more efficient, since we do not need for every target of interest to fit the method specifically tailored to it. For example, estimating the CDF with classification forests requires fitting one forest for each function value. Secondly and even more importantly, since all statistical functionals are plug-in estimates computed from the same weighting function, the obtained estimates are mathematically well-behaved and mutually compatible. For example, if we estimate $\mathrm{Cor}(Y_i, Y_j \mid \mathbf{X}=\mathbf{x})$ by separately estimating the terms $\mathrm{Cov}(Y_i, Y_j \mid \mathbf{X}=\mathbf{x})$, $\mathrm{Var}(Y_i \mid \mathbf{X}=\mathbf{x})$, and $\mathrm{Var}(Y_j \mid \mathbf{X}=\mathbf{x})$, one can not in general guarantee the estimate to be in the range $[-1, 1]$, but this is possible with DRF. Alternatively, the correlation or covariance matrices that are estimated entrywise are guaranteed to be positive semi-definite if one uses DRF. As an additional illustration, Figure 6 shows that the estimated (univariate) CDF using the classification forest need not be monotone due to random errors in each predicted value, which can not happen with the DRF estimates.

4.3 Conditional Copulas and Conditional Independence Testing

One can use the weighting function not only to estimate certain functionals, but also to obtain more complex objects, such as, for example, the conditional copulas. The well-known Sklar's theorem (Sklar, 1959) implies that at a point $\mathbf{x} \in \mathbb{R}^p$, the conditional CDF $\mathbb{P}(\mathbf{Y} \leqslant \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_1 \leqslant y_1, \dots, Y_d \leqslant y_d \mid \mathbf{X} = \mathbf{x})$ can be represented by a CDF $C_{\mathbf{x}}$ on $[0,1]^d$, the conditional copula at $\mathbf{x}$, and d conditional marginal CDFs $F_{Y_i \mid \mathbf{X}=\mathbf{x}}(y) = \mathbb{P}(Y_i \leqslant y \mid \mathbf{X} = \mathbf{x})$ for $1 \leqslant i \leqslant d$, as follows:

$$\mathbb{P}(\mathbf{Y} \leqslant \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = C_{\mathbf{x}}\left(F_{Y_1 \mid \mathbf{X}=\mathbf{x}}(y_1), \dots, F_{Y_d \mid \mathbf{X}=\mathbf{x}}(y_d)\right). \tag{20}$$

Copulas capture the dependence of the components Y_i by the joint distribution of the corresponding quantile levels of the marginal distributions: $F_{Y_i \mid \mathbf{X}=\mathbf{x}}(Y_i) \in [0,1]$. Decomposing the full multivariate distribution to marginal distributions and the copula is a very useful technique used in many fields such as risk analysis or finance (Cherubini et al., 2004). Using DRF enables us to estimate copulas conditionally, either by fitting certain parametric model or nonparametrically, directly from the weights.

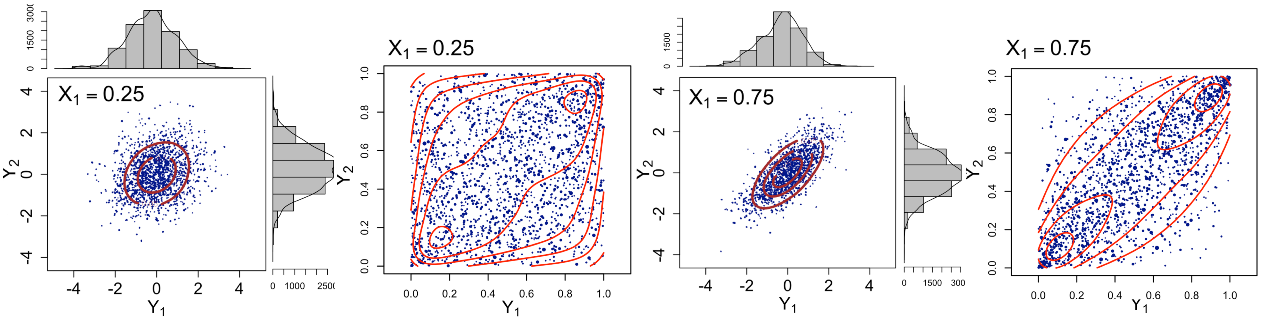


Figure 7: Estimated conditional joint distribution of (Y_1, Y_2) and conditional copulas obtained by DRF at different test points $\mathbf{x}$, where x_1 equals 0.25 and 0.75 respectively. The red lines are the contours of the true multivariate density function.

To illustrate this, consider an example where the 5-dimensional $\mathbf{Y}$ is generated from the equicorrelated Gaussian copula $\mathbf{Y} = (Y_1, \dots, Y_5) \mid \mathbf{X} = \mathbf{x} \sim C^{\text{Gauss}}_{\rho(\mathbf{x})}$ conditionally on the covariates $\mathbf{X}$ with distribution $\mathbf{X}_i \overset{i.i.d.}{\sim} U(0,1)^p$, where $p = 30$ and $n = 5000$. All Y_i have a $N(0,1)$ distribution marginally, but their conditional correlation for $i \neq j$ is given by $\mathrm{Cor}(Y_i, Y_j) = \rho(\mathbf{x}) = x_1$. Figure 7 shows that DRF estimates the full conditional distribution at different test points $\mathbf{x}$ quite accurately and thus we can obtain a good nonparametric estimate of the conditional copula as follows. First, for each component

Y_i, we compute the corresponding marginal CDF estimate $\hat{F}_{Y_i \,|\, \mathbf{X}=\mathbf{x}}(\cdot)$ from the weights. Second, we map each response $\mathbf{y}_i \to \mathbf{u}_i := \left(\hat{F}_{Y_1 \,|\, \mathbf{X}=\mathbf{x}}\left((\mathbf{y}_i)_1\right), \ldots, \hat{F}_{Y_d \,|\, \mathbf{X}=\mathbf{x}}\left((\mathbf{y}_i)_d\right)\right)$. The copula estimate is finally obtained from the weighted distribution $\sum_{i=1}^n w_{\mathbf{x}}(\mathbf{x}_i)\delta_{\mathbf{u}_i}$, from which we sample the points in Figure 7 in order to visualize the copula.

If we want to instead estimate the copula parametrically, we need to find the choice of parameters for a given model family which best matches the estimated conditional distribution, e.g. by weighted maximum likelihood estimation (MLE). For the above example, the correlation parameter of the Gaussian copula can be estimated by computing the weighted correlation with weights $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^n$. The left plot in Figure 8 shows the resulting estimates of the conditional correlation $\text{Cor}\left(Y_1, Y_2 \mid \mathbf{X} = \mathbf{x}\right)$ obtained from DRF_{MMD}, which uses the MMD splitting criterion (12) described in Section 2.3.1, and DRF_{CART}, which aggregates the marginal CART criteria (Kocev et al., 2007; Segal and Xiao, 2011). We see that DRF_{MMD} is able to detect the distributional heterogeneity and provide good estimates of the conditional correlation. On the other hand, DRF_{CART} cannot detect the change in distribution of $\mathbf{Y}$ caused by X_1 that well. The distributional heterogeneity can not only occur in marginal distribution of the responses (a case extensively studied in the literature), but also in their interdependence structure described by the conditional copula $C_{\mathbf{x}}$, as one can see from decomposition (20). Since DRF_{MMD} relies on a distributional metric for its splitting criterion, it is capable of detecting any change in distribution (Gretton et al., 2007a), whereas aggregating marginal CART criteria for $Y_1, \ldots, Y_d$ in DRF_{CART} only captures the changes in the marginal means.

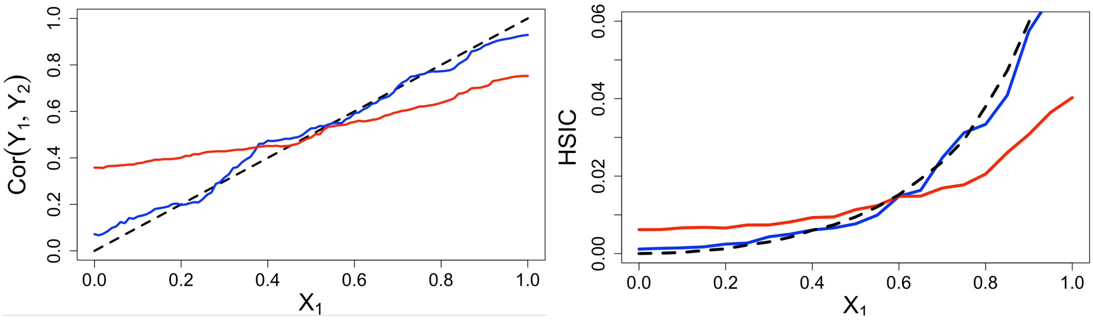


Figure 8: Estimated conditional correlation of Y_1 and Y_2 (left) and estimated conditional dependence quantified by HSIC statistic (right), obtained by DRF_{MMD} (blue) and DRF_{CART} (red) respectively. For every test point, we set $X_j = 0.5, j \neq 1$. Black dashed curve indicates the population values.

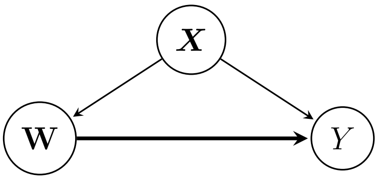
This is further illustrated for a related application of conditional independence testing, where we compute some dependence measure from the obtained weights. For example, we can test the independence $Y_1 \perp\!\!\!\perp Y_2$ conditionally on the event $\mathbf{X} = \mathbf{x}$ by using the Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al., 2007b), which measures the difference between the joint distribution and the product of the marginal distributions.

The right plot of Figure 8 shows that the DRF_{MMD} estimates are quite close to the population value of the HSIC, unlike the ones obtained by DRF_{CART}.

4.4 Heterogeneous Regression and Causal Effect Estimation

In this and the following section, we illustrate that, in addition to direct estimation of certain targets, DRF can also be a useful tool for complex statistical problems and applications, such as causality.

Suppose we would like to investigate the relationship between some (univariate) quantity of interest Y and certain predictors $\mathbf{W}$ from heterogeneous data, where the change in distribution of $(\mathbf{W}, Y)$ can be explained by some other covariates $\mathbf{X}$. Very often in causality applications, $\mathbf{W}$ is a (multivariate) treatment variable, Y is the outcome, which is commonly, but not necessarily, discrete, and $\mathbf{X}$ is a set of observed confounding variables for which we need to adjust if we are interested in the causal effect of $\mathbf{W}$ on Y. This is illustrated by the following causal graph:



The problem of nonparametric confounding adjustment is hard; not only can the marginal distributions of Y and $\mathbf{W}$ be affected by $\mathbf{X}$, thus inducing spurious associations due to confounding, but the way how $\mathbf{W}$ affects Y can itself depend on $\mathbf{X}$, i.e. the treatment effect might be heterogeneous. The total causal effect can be computed by using the adjustment formula (Pearl, 2009):

$$\begin{aligned}\mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w})] &= \int \mathbb{E}[Y \mid do(\mathbf{W}=\mathbf{w}), \mathbf{X}=\mathbf{x}]\, \mathbb{P}(\mathbf{X}=\mathbf{x} \mid do(\mathbf{W}=\mathbf{w}))d\mathbf{x} \\ &= \int \mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]\, \mathbb{P}(\mathbf{X}=\mathbf{x})d\mathbf{x}. \qquad (21)\end{aligned}$$

However, implementing do-calculus for finite samples and potentially non-discrete data might not be straightforward and comes with certain difficulties. The standard approach would be to estimate the conditional mean $\mathbb{E}[Y \mid \mathbf{W}=\mathbf{w}, \mathbf{X}=\mathbf{x}]$ nonparametrically by regressing Y on $(\mathbf{X}, \mathbf{W})$ with some method of choice and to average out the estimates over different $\mathbf{x}$ sampled from the observed distribution of $\mathbf{X}$. Using DRF for this is not necessary, but has an advantage that one can easily estimate the full interventional distribution $\mathbb{P}(Y \mid do(\mathbf{W}=\mathbf{w}))$ and not only the interventional mean.

Another way to compute the causal effect is explained in the following, which allows to add more structure to the problem. We use DRF to first fit the forest with the multivariate response $(\mathbf{W}, Y)$ and the predictors $\mathbf{X}$. In this way, one can for any point of interest $\mathbf{x}$ obtain the joint distribution of $(\mathbf{W}, Y)$ conditionally on the event $\mathbf{X} = \mathbf{x}$ and then the weights $\{w_{\mathbf{x}}(\mathbf{x}_i)\}_{i=1}^n$ can be used as an input for some regression method for regressing Y on $\mathbf{W}$ in the second step. This conditional regression fit might be of separate interest and it can also be used for estimating the causal effect $\mathbb{E}[Y \mid do(\mathbf{W} = \mathbf{w})]$ from (21), by averaging the estimates $\mathbb{E}[Y \mid \mathbf{W} = \mathbf{w}, \mathbf{X} = \mathbf{x}]$ over $\mathbf{x}$, where $\mathbf{x}$ is sampled from the empirical observation of $\mathbf{X}$. In this way one can efficiently exploit and incorporate any prior knowledge of the relationship between $\mathbf{W}$ and Y, such as, for example, monotonicity, smoothness or that it satisfies a certain parametric regression model, without imposing any assumptions on the effect of $\mathbf{X}$ on $(\mathbf{W}, Y)$. Furthermore, one might be able to better extrapolate to the regions of space where $\mathbb{P}(\mathbf{W} = \mathbf{w}, \mathbf{X} = \mathbf{x})$ is small, compared to the standard approach which computes $\mathbb{E}[Y \mid \mathbf{W} = \mathbf{w}, \mathbf{X} = \mathbf{x}]$ directly, by regressing Y on $(\mathbf{W}, \mathbf{X})$. Extrapolation is crucial for causal applications, since for computing $\mathbb{E}[Y \mid do(\mathbf{W} = \mathbf{w})]$ we are interested in what would happen with Y when our treatment variable $\mathbf{W}$ is set to be $\mathbf{w}$, regardless of the value achieved by $\mathbf{X}$. However, it can easily happen that for this specific combination of $\mathbf{X}$ and $\mathbf{W}$ there are very few observed data points, thus making the estimation of the causal effect hard (Pearl, 2009).

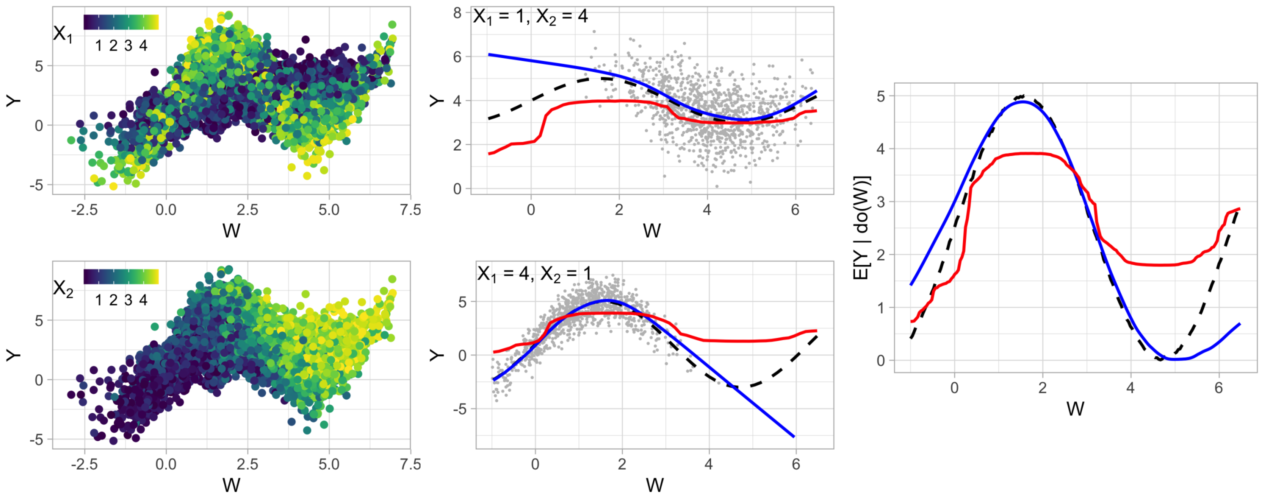


Figure 9: Left: Visualization of heterogeneous synthetic example (22). Middle: Gray points depict joint distribution of (W, Y) conditionally on $\mathbf{X} = \mathbf{x}$, for some choices of $\mathbf{x}$ indicated in the top left corner. Black curve indicates the true conditional mean $\mathbb{E}[Y \mid W = w, \mathbf{X} = \mathbf{x}]$, the blue curve represents the estimate obtained by DRF with response (W, Y) and predictors $\mathbf{X}$ in combination with smoothing splines regression, whereas the red curve represents the estimate obtained by standard Random Forest. Right: The corresponding estimates for both methods of the causal effect $\mathbb{E}[Y \mid do(W = w)]$ computed from (21). The true causal effect is denoted by a black dashed curve.

As an illustration, we consider the following synthetic data example, with continuous outcome Y, continuous univariate treatment W, $n = 5000$ and $p = 20$:

$$\mathbf{X} \sim U(0,5)^p, \quad W \mid \mathbf{X} \sim N(X_2, 1), \quad Y \mid \mathbf{X}, W \sim N(X_2 + X_1 \sin(W), 1). \tag{22}$$

A visualization of the data can be seen on the left side of Figure 9; treatment W affects Y non-linearly, X_2 is a confounding variable that affects the marginal distributions of Y and W and X_1 makes the treatment effect heterogeneous. The middle part of Figure 9 shows the conditional regression fits, i.e. the estimates of $\mathbb{E}[Y \mid W{=}w, \mathbf{X}{=}\mathbf{x}]$ as w varies and $\mathbf{x}$ is fixed. We see that combination of DRF with response (Y, W) and predictors $\mathbf{X}$ with the smoothing splines regression of Y on W (blue curve) is more accurate than the estimates obtained by standard Random Forest (Breiman, 2001) with response Y and predictors $(W, \mathbf{X})$ (red curve). Furthermore, we see that the former approach can extrapolate better to regions with small number of data points, which enables us to better estimate the causal effect $\mathbb{E}[Y \mid do(W{=}w)]$ from (21), by averaging the corresponding estimates of $\mathbb{E}[Y \mid W{=}w, \mathbf{X}{=}\mathbf{x}]$ over observed $\mathbf{x}$, as shown in the right plot of Figure 9.

The conditional regression fit $\mathbb{E}[Y \mid \mathbf{W}{=}\mathbf{w}, \mathbf{X}{=}\mathbf{x}]$ is related to the concept of the conditional average treatment effect (CATE) as it quantifies the effect of $\mathbf{W}$ on Y for the subpopulation for which $\mathbf{X} = \mathbf{x}$. There exist many successful methods in the literature for estimating the causal effects and the (conditional) average treatment effects for a wide range of settings (Abadie and Imbens, 2006; Chernozhukov et al., 2018; Wager and Athey, 2018; Künzel et al., 2019). Due to its versatility, DRF can easily be used when the underlying assumptions of existing methods are violated, when some additional structure is given in the problem or for the general, nonparametric, settings (Imbens, 2004; Ernest et al., 2015; Kennedy et al., 2017). Appendix D of Ćevid et al. (2020b) contains additional comparisons with some existing methods for causal effect estimation.

4.4.1 Births data

We further illustrate the applicability of DRF for causality-related problems on the natality data obtained from the Centers for Disease Control and Prevention (CDC) website, where we have information about all recorded births in the USA in 2018. We investigate the relationship between the pregnancy length and the birthweight, an important indicator of baby's health. Not only is this relationship complex, but it also depends on many different factors, such as parents' race, baby's gender, birth multiplicity (single, twins, triplets...) etc. In the left two plots of Figure 10 one can see the estimated joint distribution of birthweight and pregnancy length conditionally on many different covariates, as indicated in the plot. The black curves denote the subsequent regression fit, based on smoothing splines and described in detail in Appendix C of Ćevid et al. (2020b). In addition to the estimate of the mean, indicated by the solid curve, we also include the estimates of the conditional 0.1 and 0.9 quantiles, indicated by dashed curves, which is very useful in practice for determining whether a baby is large or small for its gestational age. Notice how DRF assigns less importance to the mother's race when the point of interest is a twin (middle plot), as in this case more weight is given to twin births, regardless of the race of the parents.

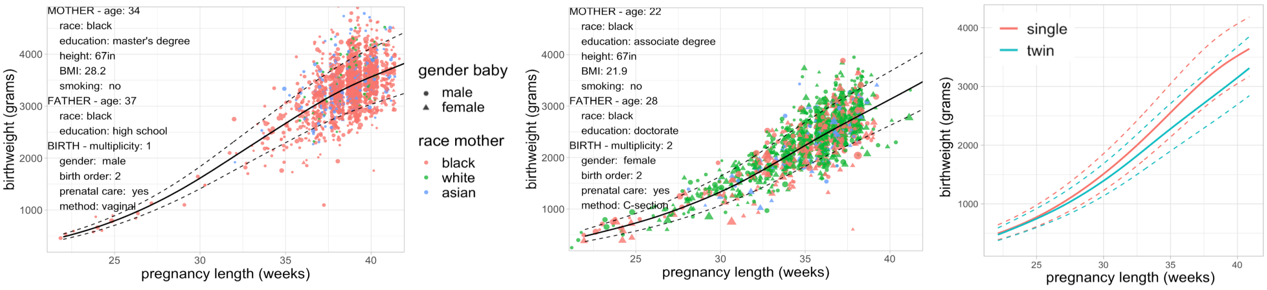
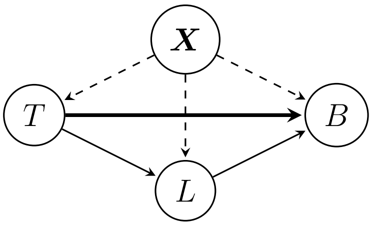


Figure 10: Left and middle: estimated relationship of pregnancy length and birthweight, conditionally on the criteria indicated in the upper left corner. Right: estimated interventional effect of twin birth on the birthweight for a fixed pregnancy length. In all plots the solid curves denote the estimated conditional mean and the dashed denote the estimated 0.1 and 0.9 quantiles.

Suppose now we would like to understand how a twin birth T causally affects the birthweight B, but ignoring the obvious indirect effect due to shorter pregnancy length L. For example, sharing of resources between the babies might have some effect on their birthweight. We additionally need to be careful to adjust for other confounding variables $\mathbf{X}$, such as, for example, the parents' race, which can affect B, T and L. We assume that this is represented by the following causal graph:



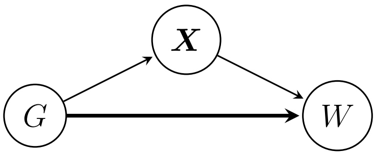
In order to answer the above question, we investigate the causal quantity $\mathbb{P}(B \mid do(T=t, L=l))$. Even though one cannot make such do-intervention in practice, this quantity describes the total causal effect if the birth multiplicity and the length of the pregnancy could be manipulated and thus for a fixed pregnancy length l, we can see the difference in birthweight due to T. We compute this quantity as already stated above, by using DRF with subsequent regression fits (described in detail in Appendix C of Ćevid et al. (2020b)), which has the advantage of better extrapolating to regions with small probability, such as long twin pregnancies (see middle plot of Figure 10). In the right plot of Figure 10 we show the mean and quantiles of the estimated interventional distribution and we see that, as one might expect, a twin birth causes smaller birthweight on average, with the difference increasing with the length of the pregnancy.

4.5 Fairness

Being able to compute different causal quantities with DRF could prove useful in a range of applications, including fairness (Kusner et al., 2017). We investigate the data on approximately 1 million full-time employees from the 2018 American Community Survey by the US Census Bureau from which we have extracted the salary information and all covariates that might be relevant for salaries. In the bottom left plot of Figure 11 one can see the distribution of hourly salary of men and women (on the logarithmic scale). The overall salary was scaled with working hours to account for working part-time and for the fact that certain jobs have different working hours. We can see that men are paid more in general, especially for the very high salaries. The difference between the median hourly salaries, a commonly used statistic in practice, amounts 17% for this data set.

We would like to answer whether the observed gender pay gap in the data is indeed unfair, i.e. only due to the gender, or whether it can at least in part be explained by some other factors, such as age, job type, number of children, geography, race, attained education level and many others. Hypothetically, it could be, for example, that women have a preference for jobs that are paid less, thus causing the gender pay gap.

In order to answer this question, we assume that the data is obtained from the following causal graph, where G denotes the gender, W the hourly wage and all other factors are denoted by $\mathbf{X}$:



i.e. G is a source node and W is a sink node in the graph. In order to determine the direct effect of the gender on wage that is not mediated by other factors, we would like to compute the distribution of the nested counterfactual $W(\text{male}, \mathbf{X}(\text{female}))$, which is interpreted in the data-generating process as the women's wage had they been treated in same way as men by their employers for determining the salary, but without changing their propensities for other characteristics, such as the choice of occupation. Therefore, it can be obtained from the observed distribution as follows:

$$\begin{aligned}\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right) &= \int \mathbb{P}\left(W(G{=}\text{male}, \mathbf{X}{=}\mathbf{x})\right) \mathbb{P}(\mathbf{X}{=}\mathbf{x} \mid G{=}\text{female}) d\mathbf{x} \\ &= \int \mathbb{P}\left(W \mid G{=}\text{male}, \mathbf{X}{=}\mathbf{x}\right) \mathbb{P}(\mathbf{X}{=}\mathbf{x} \mid G{=}\text{female}) d\mathbf{x}, \qquad (23)\end{aligned}$$

Put in the language of the fairness literature, it quantifies the unfairness when all variables $\mathbf{X}$ are assumed to be resolving (Kilbertus et al., 2017), meaning that any difference in salaries directly due to factors $\mathbf{X}$ is not viewed as gender discrimination. For example, one does not consider unfair if people with low education level get lower salaries, even if the gender distribution in this group is not balanced.

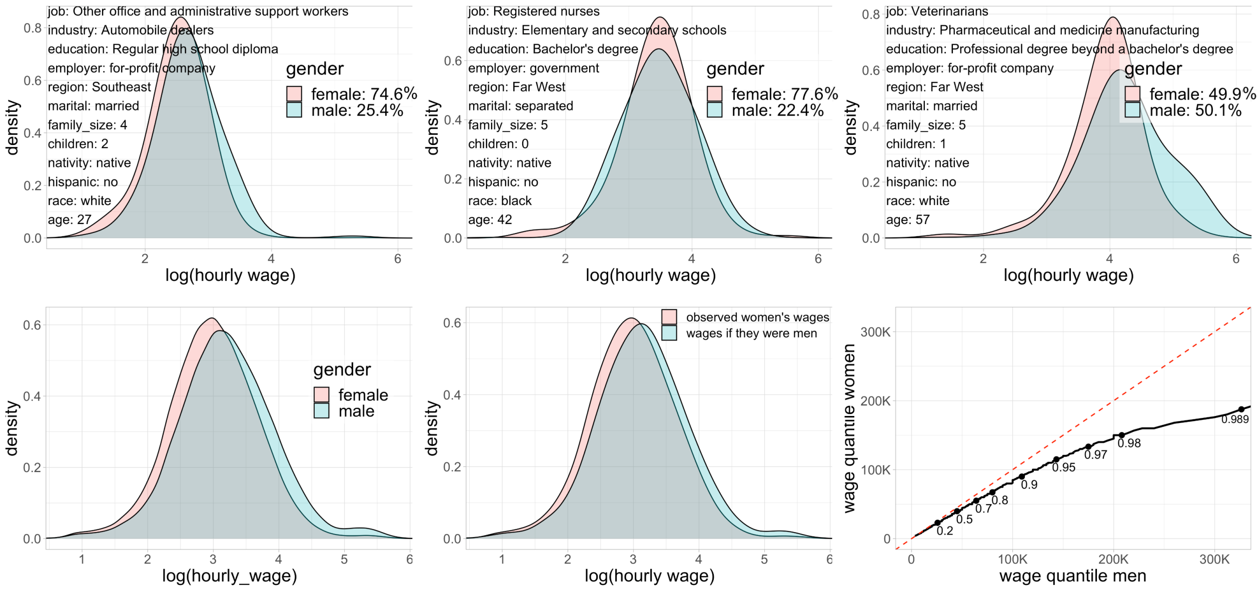


Figure 11: Top row: Estimated joint distribution of wage and gender for some fixed values of other covariates $\mathbf{X}$ indicated in the top left part of each plot. Bottom row: observed overall distribution of salaries (left), estimated counterfactual distribution $\mathbb{P}\left(W(\text{male}, \mathbf{X}(\text{female}))\right)$ of women's salaries (middle) and the quantile comparison of the counterfactual distribution of women's salaries and the observed distribution of men's salaries (right).

There are several ways how one can compute the distribution of $W(\text{male}, \mathbf{X}(\text{female}))$ from (23) with DRF. The most straightforward option is to take W as the response and $(G, \mathbf{X})$ as predictors in order to compute the conditional distribution $\mathbb{P}\left(W \mid G = \text{male}, \mathbf{X} = \mathbf{x}\right)$. However, with this approach it could happen that for predicting $\mathbb{P}\left(W \mid G = \text{male}, \mathbf{X} = \mathbf{x}\right)$ we also assign weight to training data points for which $G = \text{female}$. This happens if in some trees we did not split on variable G, which is likely, for example, if $\mathbb{P}(G = \text{male} \mid \mathbf{X} = \mathbf{x})$ is low. Using salaries of both genders to estimate the distribution of men's salaries might be an issue if our goal is to objectively compare how women and men are paid.

Another approach is to take (W, G) as a multivariate response and $\mathbf{X}$ as the predictors for DRF and thus obtain joint distribution of (W, G) conditionally on the event $\mathbf{X} = \mathbf{x}$. In this way we can also quantify the gender discrimination of a single individual with characteristics $\mathbf{x}$ by comparing his/her salary to the corresponding quantile of the salary distribution of people of the opposite gender with the same characteristics $\mathbf{x}$ (Plečko and Meinshausen, 2020). This is interesting because the distribution of salaries, and thus also the gender discrimination, can be quite different depending on other factors such as the industry sector or job type, as illustrated for a few choices of $\mathbf{x}$ in the top row of Figure 11.

Finally, by averaging the DRF estimates of $\mathbb{P}\left(W \mid \mathbf{X}{=}\mathbf{x},\, G{=}\text{male}\right)$, conveniently represented via the weights, over different $\mathbf{x}$ sampled from the distribution $\mathbb{P}(\mathbf{X} \mid G{=}\text{female})$, we can compute the distribution of the nested counterfactual $W(\text{male},\, \mathbf{X}(\text{female}))$. In the middle panel in the bottom row of Figure 11 a noticeable difference in the means, also called natural direct effect in the causality literature (Pearl, 2009), is still visible between the observed distribution of women's salaries and the hypothetical distribution of their salaries had they been treated as men, despite adjusting for indirect effects of the gender via covariates $\mathbf{X}$. By further matching the quantiles of the counterfactual distribution $\mathbb{P}\left(W(\text{male},\, \mathbf{X}(\text{female}))\right)$ with the corresponding quantiles of the observed distribution of men's salaries in the bottom right panel of Figure 11, we can also see that the adjusted gender pay gap even increases for larger salaries. Median hourly wage for women is still 11% lower than the median wage for the hypothetical population of men with exactly the same characteristics $\mathbf{X}$ as women, indicating that only a minor proportion of the actually observed hourly wage difference of 17% can be explained by other demographic factors.

5 Conclusion

We have shown that DRF is a flexible, general and powerful tool, which exploits the well-known properties of the Random Forest as an adaptive nearest neighbor method via the induced weighting function. Not only does it estimate multivariate conditional distributions well, but it constructs the forest in a model- and target-free way and is thus an easy to use out-of-the-box algorithm for many, potentially complex, learning problems in a wide range of applications, including also causality and fairness, with competitive performance even for problems with existing tailored methods.

Bibliography

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., et al. (2016). A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.

Aggarwal, K., Kirchmeyer, M., Yadav, P., Keerthi, S. S., and Gallinari, P. (2019). Benchmarking regression methods: A comparison with CGAN. *arXiv preprint arXiv:1905.12868*.

Akra, M. and Bazzi, L. (1998). On the solution of linear recurrence equations. *Computational Optimization and Applications*, 10(2):195–210.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, volume 2. Wiley New York.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329.

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bing, X., Bunea, F., Ning, Y., and Wegkamp, M. (2017). Adaptive estimation in structured factor models with applications to overlapping clustering. *arXiv preprint arXiv:1704.06977.*

Bing, X., Bunea, F., Wegkamp, M., and Strimas-Mackey, S. (2019). Essential regression. *arXiv preprint arXiv:1905.12696.*

Boef, A. G., Dekkers, O. M., Vandenbroucke, J. P., and le Cessie, S. (2014). Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. *Journal of clinical Epidemiology*, 67(11):1258–1264.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brodsky, E. and Darkhovsky, B. S. (2013). *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media.

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., and Schneeweiss, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6 0):S114.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, 26(5):2333–2355.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective.* Chapman and Hall/CRC.

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

Ćevid, D., Bühlmann, P., and Meinshausen, N. (2020a). Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21:232.

Ćevid, D., Michel, L., Meinshausen, N., and Bühlmann, P. (2020b). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *arXiv preprint arXiv:2005.14458.*

Chamberlain, G. and Rothschild, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets. Technical report, National Bureau of Economic Research.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 16, pages 785–794, New York, NY, USA. ACM.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Hansen, C., Liao, Y., et al. (2017). A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance.* John Wiley & Sons.

Chien, A. J., Conrad, W. H., and Moon, R. T. (2009). A wnt survival guide: from flies to human disease. *Journal of Investigative Dermatology*, 129(7):1614–1627.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Connor, A. E., Baumgartner, R. N., Baumgartner, K. B., Kerber, R. A., Pinkston, C., John, E. M., Torres-Mejia, G., Hines, L., Giuliano, A., Wolff, R. K., et al. (2012). Associations between tcf7l2 polymorphisms and risk of breast cancer among hispanic and non-hispanic white women: the breast cancer health disparities study. *Breast cancer research and treatment*, 136(2):593–602.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719.

Dijksterhuis, J. P., Baljinnyam, B., Stanger, K., Sercan, H. O., Ji, Y., Andres, O., Rubin, J. S., Hannoush, R. N., and Schulte, G. (2015). Systematic mapping of wnt-fzd protein interactions reveals functional selectivity by distinct wnt-fzd pairs. *Journal of Biological Chemistry*, 290(11):6789–6798.

Dobriban, E. (2017). Permutation methods for factor analysis and pca. *arXiv preprint arXiv:1710.00479*.

Donoho, D. L., Gavish, M., and Johnstone, I. M. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851*.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Dutta, B., Wallqvist, A., and Reifman, J. (2012). Pathnet: a tool for pathway analysis using topological information. *Source code for biology and medicine*, 7(1):10.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 59–82. Berkeley, CA: University of California Press.

Ernest, J., Bühlmann, P., et al. (2015). Marginal integration for nonparametric causal inference. *Electronic Journal of Statistics*, 9(2):3155–3194.

Fan, G., Wang, Z., and Wang, J. (2010). Cw-ssim kernel based random forest for image classification. In *Visual Communications and Image Processing 2010*, volume 7744, page 774425. International Society for Optics and Photonics.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J., Farmen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):591–608.

Fan, J., Ke, Y., and Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of Econometrics*.

Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. *The Annals of Statistics*, 42(3):872–917.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, J., Liao, Y., Wang, W., et al. (2016). Projected principal component analysis in factor models. *The Annals of Statistics*, 44(1):219–254.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W., Liu, B., Lei, Y., Du, S., Vuppalapati, A., Luu, H. H., Haydon, R. C., He, T.-C., and Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & diseases*, 5(2):77–106.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.

Gagnon-Bartsch, J. A., Jacob, L., and Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112.

Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552.

Gautier, E. and Rose, C. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.

Gerard, D. and Stephens, M. (2017). Empirical bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *arXiv preprint arXiv:1709.10066*.

Gerard, D. and Stephens, M. (2020). Empirical bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics*, 21(1):15–32.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Gjerga, E. and Trairatphisan, P. (2021). *CARNIVAL: A CAusal Reasoning tool for Network Identification (from gene expression data) using Integer VALue programming*. R package version 2.2.0.

Gold, D., Lederer, J., and Tao, J. (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1):79–111.

Goldszmidt, M. and Pearl, J. (1992). Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In *Proceeding of the 3rd Conference on Knowledge Representation*, pages 661–672.

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Götze, F. and Tikhomirov, A. (2002). Asymptotic distribution of quadratic forms and applications. *Journal of Theoretical Probability*, 15(2):423–475.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical science*, pages 29–46.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. (2007a). A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007b). A kernel statistical test of independence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 585–592, Red Hook, NY, USA. Curran Associates Inc.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213.

Guertin, J. R., Rahme, E., and LeLorier, J. (2016). Performance of the high-dimensional propensity score in adjusting for unmeasured confounders. *European journal of Clinical Pharmacology*, 72(12):1497–1505.

Guo, Z., Ćevid, D., and Bühlmann, P. (2020). Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. *arXiv preprint arXiv:2004.03758*.

Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815.

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427.

Han, C. (2008). Detecting invalid instruments using l1-gmm. *Economics Letters*, 101(3):285–287.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hawkes, N. (2019). Cancer survival data emphasise importance of early diagnosis. *BMJ*, 364.

He, H., Cao, S., Zhang, J.-g., Shen, H., Wang, Y.-P., and Deng, H.-w. (2019). A statistical test for differential network analysis based on inference of gaussian graphical model. *Scientific Reports*, 9(1):10863.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3):355–373.

Hothorn, T. and Zeileis, A. (2021). Predictive distribution modeling using transformation forests. *Journal of Computational and Graphical Statistics*, 0(0):1–16.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley.

Huang, J. C. and Jojic, N. (2011). Variable selection through correlation sifting. In *International Conference on Research in Computational Molecular Biology*, pages 106–123. Springer.

Huber, P. J. (2011). *Robust statistics*. Springer.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Ishwaran, H. and Kogalur, U. B. (2014). Randomforestsrc: Random forests for survival, regression and classification (rf-src). *R package version*, 1(0).

Jacot, W., Mollevi, C., Fina, F., Lopez-Crapez, E., Martin, P.-M., Colombo, P.-E., Bibeau, F., Romieu, G., and Lamy, P.-J. (2015). High egfr protein expression and exon 9 pik3ca mutations are independent prognostic factors in triple negative breast cancers. *BMC cancer*, 15(1):1–10.

Jankova, J. and van de Geer, S. (2018). Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359.

Janzing, D. and Schölkopf, B. (2018). Detecting non-causal artifacts in multivariate linear regression models. *arXiv preprint arXiv:1803.00810*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Jia, J., Rohe, K., et al. (2015). Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9(1):1150–1172.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(4):1229.

Khakabimamaghani, S., Ding, D., Snow, O., and Ester, M. (2019). Uncovering the subtype-specific temporal order of cancer pathway dysregulation. *PLoS computational biology*, 15(11).

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Kocev, D., Vens, C., Struyf, J., and Džeroski, S. (2007). Ensembles of multi-objective decision trees. In *European conference on machine learning*, pages 624–631. Springer.

Koval, A. and Katanaev, V. L. (2018). Dramatic dysbalancing of the wnt pathway in breast cancers. *Scientific reports*, 8(1):1–10.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

Lam, C., Fan, J., et al. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Le, Q., Sarlós, T., and Smola, A. (2013). Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161.

Li, S., Xie, Y., Dai, H., and Song, L. (2019). Scan B-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544.

Lin, W., Feng, R., and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288.

Lin, Y. and Jeon, Y. (2002). Random forests and adaptive nearest neighbors (technical report no. 1055). *University of Wisconsin.*

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Liu, A., Trairatphisan, P., Gjerga, E., Didangelos, A., Barratt, J., and Saez-Rodriguez, J. (2019). From expression footprints to causal pathways: contextualizing large signaling networks with carnival. *NPJ systems biology and applications*, 5(1):1–10.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585.

Manghnani, K., Drake, A., Wan, N., and Haque, I. (2018). Metcc: Metric learning for confounder control making distance matter in high dimensional biological analysis. *arXiv preprint arXiv:1812.03188.*

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., and Thomas, P. D. (2021). Panther version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive api. *Nucleic Acids Research*, 49(D1):D394–D403.

Miller, K. D., Nogueira, L., Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Alfano, C. M., Jemal, A., Kramer, J. L., and Siegel, R. L. (2019). Cancer treatment and survivorship statistics, 2019. *CA: a cancer journal for clinicians*, 69(5):363–385.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., et al. (2021). Sustainable data analysis with snakemake. *F1000Research*, 10(33):33.

Neykov, M., Ning, Y., Liu, J. S., and Liu, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443.

Nishimura, D. (2001). Biocarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, 2(3):117–120.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., and Nelson, M. R. (2008). Genes mirror geography within europe. *Nature*, 456(7218):98–101.

Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34.

P. Spirtes, C. Glymour, R. S. (2000). Causation, prediction and search. *The MIT Press*, 2nd edition.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Paul, D., Bair, E., Hastie, T., and Tibshirani, R. (2008). "preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, pages 1595–1618.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge University Press, Cambridge, UK.

Pearl, J. (2009). *Causality.* Cambridge university press.

Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44.

Pospisil, T. and Lee, A. B. (2018). Rfcde: Random forests for conditional density estimation. *arXiv preprint arXiv:1804.05753.*

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., and Lim, W. A. (2013). Repurposing crispr as an rna-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183.

Quinonero-Candela, J., Rasmussen, C. E., Sinz, F., Bousquet, O., and Schölkopf, B. (2005). Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Rahimi, A. and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320.

Ramdas, A., Trillos, N. G., and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2).

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259.

Reddi, S. J., Ramdas, A., Póczos, B., Singh, A., and Wasserman, L. (2014). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *arXiv preprint arXiv:1406.2083*.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S., and Sorger, P. K. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology*, 5(1):331.

Santolla, M. F. and Maggiolini, M. (2020). The fgf/fgfr system in breast cancer: Oncogenic features and therapeutic perspectives. *Cancers*, 12(10):3029.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl_1):D674–D679.

Schulte, G. (2010). International union of basic and clinical pharmacology. lxxx. the class frizzled receptors. *Pharmacological reviews*, 62(4):632–667.

Sedgewick, A. J., Shi, I., Donovan, R. M., and Benos, P. V. (2016). Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC bioinformatics*, 17(5):307–318.

Segal, M. and Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):80–87.

Shah, R. D., Frot, B., Thanei, G.-A., and Meinshausen, N. (2020). Right singular vector projection graphs: fast high-dimensional covariance matrix estimation under latent confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82:361–389.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Simon-Gabriel, C.-J., Barp, A., and Mackey, L. (2020). Metrizing weak convergence with maximum mean discrepancies. *arXiv preprint arXiv:2006.09268*.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *The Journal of Machine Learning Research*, 19(1):1708–1736.

Sklar, A. (1959). *Fonctions de Répartition À N Dimensions Et Leurs Marges.* Université Paris 8.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491.

Song, D., Cui, M., Zhao, G., Fan, Z., Nolan, K., Yang, Y., Lee, P., Ye, F., and Zhang, D. Y. (2014). Pathway-based analysis of breast cancer. *Am J Transl Res*, 6(3):302–311.

Song, H. and Zhang, Y. (2018). Regulation of pancreatic stellate cell activation by notch3. *BMC cancer*, 18(1):1–12.

Sriperumbudur, B. (2016). On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839 – 1893.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., and Asiedu, J. K. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Sun, Y., Zhang, N. R., Owen, A. B., et al. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics*, 6(4):1664–1688.

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82.

Tian, D., Gu, Q., and Ma, J. (2016). Identifying gene regulatory network rewiring using latent differential graphical models. *Nucleic acids research*, 44(17):e140–e140.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Troester, M. A. and Swift-Scanlan, T. (2009). Challenges in studying the etiology of breast cancer subtypes. *Breast Cancer Research*, 11(3):104.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011). Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12.

Van de Geer, S. (2016). Estimation and testing under sparsity. *Lecture Notes in Mathematics*, 2159.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyniok, G., editors, *Compressed sensing: theory and applications*, pages 210–268. Cambridge University Press.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127):1546–1558.

Wager, S. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017a). Confounder adjustment in multiple hypothesis testing. *Annals of statistics*, 45(5):1863–1894.

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., Sweet, R. A., Wang, J., and Chen, W. (2016). Fastggm: An efficient algorithm for the inference of gaussian graphical model in biological networks. *PLOS Computational Biology*, 12(2):1–16.

Wang, W., Fan, J., et al. (2017b). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342–1374.

Wang, Y. and Blei, D. M. (2018). The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*.

Wang, Y. and Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596.

Wang, Y., Squires, C., Belyaeva, A., and Uhler, C. (2018). Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems*, pages 3770–3781.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Wendland, H. (2004). *Scattered Data Approximation.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Whirl-Carrillo, M., McDonogh, E., Herbet, J., Gong, L., Sangkuhl, K., Thotn, C., Altman, R., and Klein, E. (2012). Pharmacogenomics knowledge for personlized medicine. clinical pharmacology and therpeutics 92, 4 (2012), 414–417.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838.

Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. *Journal of economic entomology*, 39(2):269–270.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350.

Wolfe, D. A. and Schechtman, E. (1984). Nonparametric statistical procedures for the changepoint problem. *Journal of Statistical Planning and Inference*, 9(3):389–396.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, M. N. and Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133.

Yin, P., Wang, W., Gao, J., Bai, Y., Wang, Z., Na, L., Sun, Y., and Zhao, C. (2020). Fzd2 contributes to breast cancer cell mesenchymal-like stemness and drug resistance. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 28(3):273–284.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286.

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, pages 755–763.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Zhao, J. and Meng, D. (2015). Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*.

Zhu, Y. (2018). Sparse linear models and l1-regularized 2sls with high-dimensional endogenous regressors and instruments. *Journal of Econometrics*, 202(2):196–213.

Curriculum Vitae

Personal data

Name	Domagoj Ćevid
Date of birth	December 3, 1994
Nationality	Croatian

Education

2017–2021	**Seminar für Statistik, ETH Zürich**, Switzerland PhD in Mathematics
2016–2017	**University of Cambridge**, United Kingdom MMath in Mathematics
2013–2016	**University of Cambridge**, United Kingdom BA in Mathematics
2009–2013	**Fifth Grammar School**, Croatia

Professional experience

2017–2021	**ETH Zürich**, Switzerland Research Assistant
2016	**Google Zürich**, Switzerland Software Engineering Intern
2015	**Jane Street Capital**, United Kingdom Trading Intern

Awards and Honors

IMO	Silver Medal (2011), Gold Medal (2012, 2013)
IOI	Silver Medal (2012, 2013)
ACM ICPC	World Finals Participation (2014)