Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 4 (2803–2818)

INVARIANCE IN HETEROGENEOUS, LARGE-SCALE AND HIGH-DIMENSIONAL DATA

Peter Bühlmann

Abstract

Statistical inference from large-scale data can benefit from sources of heterogeneity. We discuss recent progress of the mathematical formalization and theory for exploiting heterogeneity towards predictive stability and causal inference in high-dimensional models. The topic is directly motivated by a broad range of applications and we will show an illustration from molecular biology with gene knock out experiments.

1 Introduction

In the advent of large data acquisition we expect that heterogeneity occurs within datasets. The data are typically *not* realizations from independent and identically distributed random variables, nor from a stationary process. We either know that the data come from large-scale experimental perturbations, for example in many bio-molecular applications, or one can empirically detect heterogeneity in terms of shifts, non-stationarities or cluster-membership, for example in macroeconomics.

Rather than considering heterogeneity as a nuisance, one can exploit and use it to obtain more insights and better predictions – in folklore: "Make heterogeneity your friend rather than your enemy". This line of thinking in the context of large-scale data seems "new" Peters, Bühlmann, and Meinshausen [2016]: the foundations though are much older and go back to Trygve Haavelmo in 1943 Haavelmo [1943], who received the Nobel Prize in economics in 1989 "for his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures". Haavelmo has advocated an invariance property across changing (heterogeneous) structures, technically in terms of structural equation models, and this is nowadays adopted in the framework of causality Pearl [2000, cf.].

MSC2010: primary 62J07; secondary 62F99, 68T99.

Keywords: Causality, Hidden variables, High-dimensional models, Structural equation models, Predictive stability.

We discuss here the "reverse relation" where invariance can be extracted from data, enabling predictive stability and towards inference of causal parameters. The corresponding mathematical formulation and theory involve the combination of identifiability from causal inference and techniques from high-dimensional statistical inference (the latter is due to the dimensions of many modern datasets). The methodology and mathematical problems are in close vicinity of applications: we will illustrate an ambitious example of predicting gene knock out perturbations, a fundamental task in molecular-biology for gaining insights into causal gene interactions and for prioritizing future biological experiments.

2 The setting

We consider regression or classification problems with d-dimensional covariates (features) X and a one-dimensional response variable of interest Y.

As a starting point, consider a linear model for n data points, being realizations from

$$Y_i = X_i^T \beta^0 + \varepsilon_i \ (i = 1, \ldots, n),$$

where the covariates X_i, the response Y_i and the errors ε_i are random with $\mathbb{E}[\varepsilon_i | X_i] = 0$ and the $d \times 1$ vector β^0 denotes the unknown true regression parameter of interest. (Because this true underlying parameter is of special interest, we denote it with an additional superscript "^{0}"). The most often used assumption is that the random variables X_i and ε_i are independent from each other and both of them independent and identically distributed (i.i.d.) across $i = 1, \ldots, n$, and hence (Y_i, X_i), $i = 1, \ldots, n$, are i.i.d. as well. We often use the short-hand notation for a linear model

$$\mathbf{Y} = \mathbf{X}\beta^0 + ”, \tag{1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $” = (\varepsilon_1, \ldots, \varepsilon_n)^T$ are $n \times 1$ vectors and $\mathbf{X} = (X_1, \ldots, X_n)^T$ is the $n \times d$ design matrix.

Over the last 15 years, a huge amount of literature has been devoted to the problem of estimating the unknown parameter vector β^0 in the high-dimensional sparse case where $d \gg n$: some of the earlier references include Donoho [1993], Donoho and Johnstone [1994], Tibshirani [1996], Chen, Donoho, and Saunders [2001], Greenshtein and Ritov [2004], Bühlmann [2006], Meinshausen and Bühlmann [2006], Bunea, Tsybakov, and Wegkamp [2007], Zou [2006], Zhao and Yu [2006], Candès and Tao [2007], Bickel, Ritov, and Tsybakov [2009], and Koltchinskii [2009a,b], and see also the monographs Bühlmann and van de Geer [2011], Giraud [2014], and Hastie, Tibshirani, and Wainwright [2015]. Furthermore, estimation of the parameter β^0 in the noiseless case is the same as compressed sensing Donoho and Huo [2001], Donoho [2006], Candès, Romberg, and Tao [2006], and Candès and Tao [2006, cf.], and this itself is a huge field by now.

We will focus on the case where an i.i.d. assumption as above for the random variables (Y_i, X_i) $(i = 1, \ldots n)$ does not hold. This seems particularly relevant for large-scale "big" data. In the advent of large data collection, it is often reasonable to assume that the data exhibits "heterogeneity". Loosely speaking, we mean by this that the data come from e.g.: (i) different regimes, for example across time in applications such as economics, finance or neuroscience; (ii) from different perturbations, for example in molecular biology; (iii) from different sub-populations, for example in online advertisement or auction pricing. In an abstract sense, we generalize the linear model from (1) to

$$(\mathbf{Y}^e, \mathbf{X}^e) \sim F^e,\ e \in \mathcal{E}, \tag{2}$$

where $\mathbf{Y}^e$ is an $n^e \times 1$ vector, $\mathbf{X}^e$ an $n^e \times d$ design matrix, e denotes an environment or a sub-population from a space of environments $\mathcal{E}$, and F^e is the distribution depending on environment e. Typically, we assume that the environments $e \in \mathcal{E}$ are known (observed), but see below for an example where they are unknown.

Example: Gene knock out perturbations in yeast Meinshausen, Hauser, Mooij, Peters, Versteeg, and Bühlmann [2016].
Among the approximately 6'000 genes in yeast, 1'479 have been knocked out and a phenotypic response is measured. The space $\mathcal{E}$ corresponds to the different gene knock-out perturbations. In the extreme case, every gene knock out corresponds to a single e and the space $\mathcal{E} = \{1, 2, \ldots, 1479\}$. One can also think to pool (some of) the different perturbations and the space $\mathcal{E}$ is then smaller. This example is discussed further in Section 5.4.

Example: Monetary policy in macro economics Pfister, Bühlmann, and Peters [2017].
The data are monthly observations of the Euro – Swiss Franc exchange rate over 18 years and nine macro economic variables such as GDP or inflation rate. The space $\mathcal{E}$ corresponds to unknown time-dependent regimes. Although the environments e (the regimes) in $\mathcal{E}$ are unknown, they are assumed to be present in the observed data.

Consider a space of (mostly unobserved) environments $\mathcal{F}$, typically $\mathcal{F} \supset \mathcal{E}$ being much larger than the observed environments in $\mathcal{E}$. In the examples above, $\mathcal{F}$ would be $\mathcal{E}$ but in addition also including the gene perturbations or the future time-dependent regimes which are not observed in the data. We are interested in the following problems.

1. Prediction for new scenarios or environments in $\mathcal{F}$. When adopting a linear model, consider the following objective:

   $$\beta^* = \text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}[|Y^e - (X^e)^T \beta|^2],$$

   and how to infer β^* from data as in (2) from much fewer observed environments in $\mathcal{E}$. That is, we want to infer the parameter β^* which optimizes the worst case loss within a class of new unobserved scenarios in $\mathcal{F}$. We will discuss in Section 5 (e.g. Theorem 4) some cases for $\mathcal{F}$.

2. Predicting unseen interventions or perturbations. We want to predict Y^e when doing a perturbation on some of the covariates X^e, for $e \in \mathcal{F}$ corresponding to a new perturbation which was not observed in the data. That is, in popular terms, a "what if I do (perturb)" question: the answer to such a question is at the fundamental basis of causal inference Pearl [2000] and Peters, Janzing, and Schölkopf [2017, cf.].

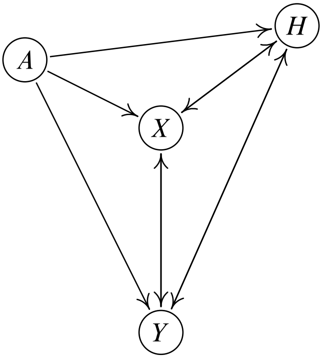
3. Finding stable structures. We also aim to find subsets of covariates $S \subseteq \{1, \ldots, d\}$ leading to (near) invariance in terms of the conditional distribution $\mathcal{L}(Y^e | X^e_S)$ being (nearly) constant across environments $e \in \mathcal{F}$.

All these three points above are closely related. When it comes to inference from finite samples, the setting is usually high-dimensional since in each observed environment e, the sample size n^e is often not so large implying that the covariate dimension $d \gg n^e$. Therefore, the underlying mathematical developments involve high-dimensional statistical theory together with perturbation analysis in structural equation models (see (4)) for analyzing heterogeneity.

3 Modeling heterogeneity and perturbations

We discuss here a general model for heterogeneous data as in (2). There is a d-dimensional covariate X, a q-dimensional hidden (latent) variable H, an r-dimensional "anchor" variable A and a one-dimensional response Y. A discrete space of environments $\mathcal{E} = \{1, \ldots, m\}$ mentioned before can be described with $r = m$ anchor variables, each of them being binary where $A_k = 1$ means that the corresponding environment is $e = k$ $(k = 1, \ldots, m)$.

The involved variables $(Y, X, H, A)^T$ is a $(1 + d + q + r)$-dimensional random vector and each component is corresponding to a node in a directed graph describing the "structure" among the variables. The directed graph is qualitatively as follows:



The (bi-)directed arrows correspond to the structure of a structural equation model (see below). There are structural relations among the components of X, H and A as well but this is not visible in the displayed graph. Bi-directed arrow allow to exhibit directed cycles, i.e., feedback loops. The variables in H are hidden confounders between X and Y which makes it hard to identify effects from components X_j to Y which are not due to the hidden confounding.

A quantitative model on the graph is given by a structural equation model: a linear structural equation model is given below in (4) in its abstract form. To exemplify: the equation for the response, which is of major interest since Y is the target one wants to understand, reads

$$Y \leftarrow \sum_{k \in \mathrm{pa}(Y) \cap \{X\}} \beta_k^0 X_k + \sum_{k \in \mathrm{pa}(Y) \cap \{H\}} \delta_k H_k + \sum_{k \in \mathrm{pa}(Y) \cap \{A\}} \alpha_k A_k + \varepsilon_Y, \tag{3}$$

where $\mathrm{pa}(Y)$ denotes the parental set of a variable Y in the directed graph, $\{X\}$ denotes the nodes corresponding to the random variables from the components of X (and analogously for H, A), and ε_Y is an exogenous stochastic noise term. The directed arrow "←" means that the variable on the left hand side is a "direct function" or "caused" by the variables on the right hand side: it can be replaced by the expression of "equality in distribution". The parameter β^0 is of special interest: in the literature it is called the direct causal effect Pearl [2000, cf.], describing the direct effect from X to Y (see below). In fact, the causal parameter describes what happens when doing a perturbation/intervention on the X-variables, see goal 2. in Section 2, and it is intrinsically related to invariance properties with respect to perturbations Haavelmo [1943] and Peters, Bühlmann, and Meinshausen [2016]: we will take up the latter point in Section 5. We note that an L_2-projection does not lead to β^0: $\mathrm{argmin}_\beta \mathbb{E}[|Y - X^T\beta|^2] \neq \beta^0$; thus, inferring β^0 from data is a more complicated task than using standard regression methodology.

Having displayed above the structural equation of the response Y, all other variables have such a structural representation as well, for example

$$\begin{aligned} X_j \leftarrow & \sum_{k \in \mathrm{pa}(X_j) \cap \{X\}} \kappa_{j,k} X_k + \sum_{k \in \mathrm{pa}(X_j) \cap \{H\}} \gamma_{j,k} H_k \\ & + \sum_{k \in \mathrm{pa}(X_j) \cap \{A\}} \alpha_{j,k} A_k + \xi_j Y I(Y \in \mathrm{pa}(X_j)) + \varepsilon_j \end{aligned}$$

We can represent the model in algebraic form as

$$\begin{pmatrix} Y \\ X \\ H \end{pmatrix} = B \begin{pmatrix} Y \\ X \\ H \end{pmatrix} + MA + \varepsilon, \tag{4}$$

where B is a $(1 + d + r) \times (1 + d + r)$ matrix, M a $(1 + d + r) \times r$ matrix and ε a stochastic noise vector of dimension $(1 + d + r) \times 1$. We note that since A is an "anchor" or a source node in the graph, it is exogenous and hence it appears on the right hand side of (4) only.

Of main interest is the equation and dynamics for the response Y: if $(I - B)$ is invertible, see below, we can express Y and all other X, H as a function of A and ε,

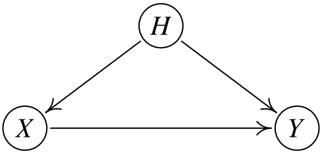
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (I - B)^{-1}(\varepsilon + MA). \tag{5}$$

As mentioned above, the variable A can describe heterogeneity, and the formulation in (5) is a useful representation for the perturbation effect of shift interventions, see Section 5.2. The matrix $(I - B)$ is invertible if the underlying structure (encoding zeroes in B) is a directed acyclic graph; for cyclic graphs, one typically assumes an equilibrium solution of the dynamical system when conditioning on ε and A, for example requiring that the cycle-product is strictly less than one (but we do not need such an equilibrium assumption).

3.1 Some special cases.

Some special cases highlight the generality of the model in (4).

Hidden confounding. This model has no "anchor" variable A but some hidden (unobserved) confounders H. The directed graph looks as follows.



The structural equation model is:

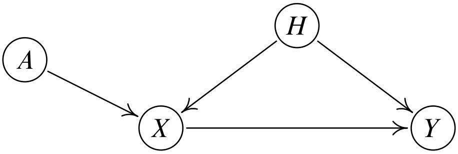
$$
\begin{aligned}
&H = (H_1, \dots, H_q)^T \text{ uncorrelated with covariance matrix } I_q\\
&X \leftarrow \Gamma H + \varepsilon_X,\\
&Y \leftarrow X^T \beta^0 + H^T \delta + \varepsilon_Y
\end{aligned}
\tag{6}
$$

where all the components of $\varepsilon_X, \varepsilon_Y, H$ are jointly independent, Γ is a $d \times q$ matrix and β^0 the $d \times 1$ regression vector. There is an implicit directionality assumption saying that there are no structural directions from Y to some of the components of X (i.e.,, Y is "childless").

We will argue that despite the hidden confounders H, one can estimate the causal coefficient vector β^0 when the setting is high-dimensional (among other conditions).

A prominent example for such a model are genome-wide association studies (GWAS). The response variable Y is often a medical or disease status, the covariates X are genetic biomarkers in terms of single nucleotide polymorphisms (SNPs) and the hidden confounders can come from various sources such as environment or unmeasured genetic profiles. The dimensionality of the SNPs is in the order of $d = O(10^6)$ and a typical sample size is in the range of $O(10^3)$. It is a very high-dimensional setting with the interesting additional information about direction: if there is an association between X and Y, it must point from X to Y; this, because the SNPs are genetic information and the medical status cannot influence the genetics (although there are some exceptions with retroviruses like HIV).

Instrumental variables regression. This is a very popular and well-studied model in economics. The variables in A are called instruments and the directed graph looks as follows.



In contrast to the general situation considered above, there are no directed arrows from either X, Y or H to A, and there are no bi-directed arrows.

The corresponding structural equation model is as in (4) with the constraint from the directed graph above:

$$\begin{aligned}
&H = (H_1, \dots, H_q)^T \text{ from a distribution } F_H,\\
&A = (A_1, \dots, A_r)^T \text{ from a distribution } F_A,\\
&X \leftarrow \Gamma H + MA + \varepsilon_X,\\
&Y \leftarrow X^T \beta^0 + H^T \delta + \varepsilon_Y. \qquad (7)
\end{aligned}$$

A necessary condition for identifiability of β^0 is to have at least as many instruments as covariates, i.e., $r \geq d$. More precisely, the condition $\text{rank}(\mathbb{E}[(AA^T)]M^T) \geq d$, is necessary and sufficient; and this involves also that the coefficient matrix M is not “too degenerate”.

4 Hidden confounding in high-dimensional settings

Consider here the model in (6) with hidden confounder variables. Such hidden confounders can also be thought as generating different regimes or environments in the data. In high-dimensional settings, we assume that the regression parameter β^0 (with a causal interpretation) is sparse.

When using the population least squares principle, we obtain

$$\begin{aligned}
&\beta_{\text{LS}} = \Sigma_X^{-1}\text{Cov}(Y, X) = \beta^0 + b,\\
&b = \Sigma_X^{-1}\Gamma\delta,
\end{aligned}$$

where $\Sigma_X = \text{Cov}(X) = \Gamma\Gamma^T + \text{Cov}(\varepsilon_X)$.

Example: one hidden confounder ($r = 1$) and same noise terms for X: $Cov(\varepsilon_X) = \sigma_\varepsilon^2 I$.

We obtain that the bias equals

$$b = \Gamma \frac{\delta}{\Lambda_{\max}^2(\Gamma^T\Gamma) + \sigma_\varepsilon^2} = \Gamma \frac{\delta}{\|\Gamma\|_2^2 + \sigma_\varepsilon^2},$$

where $\Lambda_{\max}^2(\Gamma\Gamma^T)$ denotes the maximal eigenvalue of $\Gamma\Gamma^T$ (the only non-zero eigenvalue). Suppose that the number of non-zero entries in $\Gamma = m$ and that the non-zero entries in Γ are not too small, i.e., $\|\Gamma\|_2^2 \asymp m \to \infty$ as $d \to \infty$. Then,

$$\|b\|_2^2 = O(m^{-1}) \text{ as } d \to \infty. \qquad (8)$$

Thus, if the hidden variables have an effect which is sufficiently spread out, i.e. m being large, the bias of population least squares (which ignores the hidden confounders) will disappear in high dimensions.

4.1 Sparse plus dense signals. The analysis above can be used as follows. We can write the model by L_2-projection as

$$Y = X^T \beta_{\mathrm{LS}} + \eta, \ \ \mathrm{Cov}(\eta, X) = 0,$$

and thus we can also write

$$Y = X^T(\beta^0 + b) + \eta, \tag{9}$$

with the bias term as above which is typically dense (under fairly natural conditions, see above).

We assume to have observed data $(\mathbf{Y}, \mathbf{X})$ being i.i.d. realizations of (Y_i, X_i) $(i = 1, \ldots n)$ from (6), also involving the unobserved $\mathbf{H}$ from i.i.d. realizations H_i $(i = 1, \ldots, n)$. The sample version of the model in (9) is as follows:

$$\mathbf{Y} = \mathbf{X}(\beta^0 + b) + \eta, \ \ \eta \text{ uncorrelated with } \mathbf{X}.$$

In the high-dimensional regime with $d \gg n$, we have a linear model with a signal being a composition of a sparse β^0 and a dense b. Estimation of the sparse vector β^0 can be done with a combination of ℓ_1- and ℓ_2 regularization, the Lava estimator Chernozhukov, Hansen, Liao, et al. [2017]:

$$\mathrm{argmin}_{\beta,b} \left(\|\mathbf{Y} - \mathbf{X}(\beta + b)\|_2^2/n + \lambda_1 \|\beta\|_1 + \lambda_2 \|b\|_2^2 \right).$$

Note that for $\lambda_2 = \infty$ we obtain the ℓ_1-regularized Lasso estimator, and analogously $\lambda_1 = \infty$ corresponds to the ℓ_2-regularized Ridge procedure (Tikhonov regularization). The solution of this convex optimization problem is given by:

$$\begin{aligned} \hat{\beta} &= \mathrm{argmin}_\beta \left\{ \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta\|_2^2/n + \lambda_1 \|\beta\|_1 \right\}, \\ \hat{b} &= (\mathbf{X}^T\mathbf{X} + n\lambda_2 I)^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\hat{\beta}). \end{aligned} \tag{10}$$

Here, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are given by $K_{\lambda_2}^{1/2}\mathbf{X}$ and $K_{\lambda_2}^{1/2}\mathbf{Y}$ respectively, where

$$K_{\lambda_2} = I - \mathbf{X}(\mathbf{X}^T\mathbf{X} + n\lambda_2 I)^{-1}\mathbf{X}^T.$$

The representation in (10) leads to some consequences and insights. First, the computation can simply be done by an ℓ_1-norm regularization to a transformed problem with response $\tilde{\mathbf{Y}}$ and covariates $\tilde{\mathbf{X}}$. Second, the mathematical properties of $\hat{\beta}$ can be studied from the view point of the theory for ℓ_1-norm regularization (and compressed sensing): there are two obstacles though, namely: (i) the underlying coefficient vector is sparse plus dense, and a bias term will be inherent in sparse approximation; (ii) the transformed design matrix $\tilde{\mathbf{X}}$ has other restricted eigenvalues Bickel, Ritov, and Tsybakov [2009] and van de Geer and Bühlmann [2009] than the original $\mathbf{X}$ and the transformed error $\tilde{\varepsilon} = K_{\lambda_2}^{1/2}$ is correlated and possibly inflates.

The following result holds.

Theorem 1. *Ćevid, Bühlmann, and Meinshausen [2018] Assume Gaussian variables $H, \varepsilon_X, \varepsilon_Y$ in (6) with mean zero. In addition, assume condition (A) described below. Denote by $s_0 = |\mathrm{supp}(\beta^0)|$ and assume that $s_0\sqrt{\log(d)/n} = o(1)$ $(d \gg n \to \infty)$. Then, for the Lava estimator in (10), there exist suitable values λ_1 and λ_2 such that with probability tending to one as $d \gg n \to \infty$:*

$$\|\hat{\beta} - \beta^0\|_1 \leq C \frac{\sigma s_0}{\Lambda_{\min}^2(\Sigma)} \sqrt{\log(d)/n}, \tag{11}$$

where $0 < C < \infty$ is a constant, $\sigma^2 = \mathbb{E}|\eta|^2$ in (9) and $\Lambda_{\min}^2(\Sigma_X)$ denotes the minimal eigenvalue of $\Sigma_X = Cov(X) = \Gamma\Gamma^T + Cov(\varepsilon_X)$.

The assumption (A) below ensures that the bias term is asymptotically negligible. There is a broader range of scenarios implying negligible bias, and a simple example is as follows.

(A) In model (6), the entries of the $d \times q$ matrix Γ are i.i.d. from an absolutely continuous distribution w.r.t. Lebesgue measure, $q < \infty$ is a fixed number, and ε_X has i.i.d. components.

The parameter λ_2 can be chosen according to a spectral clustering property and λ_1 then remains as the only tuning parameter as for the ℓ_1-norm regularization scheme with the Lasso. The result in Theorem 1 is based on an analysis in the transformed model with $\tilde{Y}$ and $\tilde{X}$ in (10): we can trade-off between the behavior of the singular values of $\tilde{X}$ and an inflation of the transformed error $K_{\lambda_2}^{1/2}\varepsilon$. It involves recent results about the behavior of singular or spectral values in large random matrices.

The estimation strategy with the Lava estimator in (10) and its justification in Theorem 1 for the hidden confounder model has major implications in practice, including also the broad area of genome-wide association studies where accounting for sub-population structure is important.

5 Predictive stability, invariance and causal regularization

We considered in Section 4 a problem where the direction of an association is known to point from X to Y. Even when having confounding structure, one can then essentially identify the causal regression effect in high-dimensional problems.

For cases where the direction of an association is not known, one needs more to identify the direction of an association and eventually the causal regression parameter β^0 in (3). Heterogeneity and perturbations help towards identifiability of directions and causality. Instrumental variable models as in (7), originating from economics Geary [1949], are now popular in other fields as well. And indeed, if $r = \dim(A) \geq d = \dim(X)$, they typically lead to identifiability of the directed associations and causal parameter β^0 from X to Y. The key idea relies on the fact that A and ε_Y are independent in the model (6): hence we should choose a regression parameter β such that

$$\|\mathbb{E}[A(Y - X^T\beta)]\|_q = 0 \text{ for some } q \geq 1. \tag{12}$$

For $\beta = \beta^0$, the equality in (12) holds, and if A is sufficiently rich, necessarily requiring that $r \geq d$, β^0 is the only solution satisfying (12).

But assuming a structure as in the instrumental variable model in (7) is often more an uncertain bet than a realistic assumption. The model in (4) or (5) relaxes this restriction substantially. In addition, we do not assume to have $r \geq d$ (or more precisely that the number of children of A is larger or equal to d). “Everything” is possible, including cycles, except that the “anchor” A is exogenous (meaning that it is a source node in the graph). In such a general setting, the causal parameter β^0 in (3) is typically not identifiable.

5.1 Causal regularization. It seems natural in general to have a soft version of the constraint in (12). A regression method can be equipped with an additional regularization term:

$$\hat{\beta} = \hat{\beta}_{\gamma;q} = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \gamma \|\mathbf{A}^T(\mathbf{Y} - \mathbf{X}\beta)/n\|_q^2 + \lambda\|\beta\|_1 \right), \tag{13}$$

where $\|\beta\|_1$ is a regularizer for high-dimensionality; typical choices are $q = 2$ or $q = \infty$. Here $\mathbf{A}$ is the $n \times r$ matrix of the observed variables of A. For $\gamma = 0$, we get the usual penalized regression estimator while for $\gamma = \infty$ we enforce the finite-sample version of the restriction in (12). The latter restriction might not be possible to be fulfilled for any β and thus, $\gamma = \infty$ might not be appropriate. The question then becomes: what are the properties of such an estimator in (13) in general? We address this in the next section.

5.2 Predictive stability and invariance under shift interventions: the population case. We consider the problem of predictive stability and invariance of residuals under a class of shift interventions. For example: in the instrumental variable model in (7), the residuals $Y - X\beta^0 = H\delta + \varepsilon_Y$ are invariant under any interventions/perturbations on X which leave the structure and parameters in the structural equation model in (4) unchanged. That is: the causal parameter leads to predictions whose errors remain invariant under arbitrary perturbation scenarios on X.

We want to understand such invariance and predictive stability in the general model (4) or (5) where A are not instruments (they can point to X, H or Y; and feedback cycles are allowed). For this, we restrict ourselves to shift interventions. We consider shifts v, a $(1+d+r)$-dimensional vector being deterministic or random, which can act on any of the variables Y, X, H (but not on A): the shifted random variables (Y^v, X^v, H^v, A) are given as the solution of

$$\begin{pmatrix} Y^v \\ X^v \\ H^v \end{pmatrix} = B \begin{pmatrix} Y^v \\ X^v \\ H^v \end{pmatrix} + v + MA + \varepsilon. \tag{14}$$

In the random case we assume that v is independent of ε and A. A shift intervention v acts as a shift v_k on the component $(Y, X, H)_k$ (for all k where $v_k \neq 0$) and such shifts v_k are propagated through the SEM, changing the distribution of other components $(Y, X, H)_j$ for $j \neq k$. In the alternative form analogous to (5) we can write

$$\begin{pmatrix} Y^v \\ X^v \\ H^v \end{pmatrix} = (I - B)^{-1}(v + MA + \varepsilon). \tag{15}$$

We now consider a class of shifts of the form

$$C_\gamma^q = \{v;\ v = M\delta \text{ for some } \delta \text{ with } \|\delta\|_q \leq \gamma\}.$$

Thus, C_γ^q includes shifts in the span of M which have at most a certain strength, measured by the ℓ_q-norm of the coefficient vector δ in the representation of the shift. We then see that the term $v + MA$ in (14) or (15) becomes: $v + MA = M(\delta + A)$ which intrinsically links the shift v to a perturbation of A of the form $A + \delta$.

The following fundamental result connects some worst case risk to causal regularization.

Theorem 2. *Rothenhäusler, Bühlmann, Meinshausen, and Peters [2018] Consider the model in* (4) *or* (5) *with* $(I - B)$ *being invertible. Then, for any* $p, q \geq 1$ *with* $p^{-1} + q^{-1} = 1$, *and for any* $b \in \mathbb{R}^d$*:*

$$\max_{v \in C_\gamma^q} \mathbb{E}[|Y^v - X^v b|^2] = \mathbb{E}[|Y - Xb|^2] + \gamma \|\mathbb{E}[A(Y - Xb)]\|_p^2.$$

Theorem 2 has important consequences on predictive stability. First of all, since the result holds for any $b \in \mathbb{R}^d$, we can consider the argmin on both sides of the equality:

$$\begin{aligned} b_{\gamma;q} &= \operatorname{argmin}_{\beta \in \mathbb{R}^d} \max_{v \in C_\gamma^q} \mathbb{E}[|Y^v - X^v b|^2] \\ &= \operatorname{argmin}_{\beta \in \mathbb{R}^d} \left(\mathbb{E}[|Y - X\beta|^2] + \gamma \|\mathbb{E}[A(Y - X\beta)]\|_p^2\right). \end{aligned} \tag{16}$$

Thus, the optimizer for the worst case risk within the class C_γ^q equals exactly the one from the regularized criterion for the non-shifted variables. The interpretation is as follows: the shifted variables (Y^v, X^v) can be associated to test data where some shift perturbations have occurred, while the non-shifted variables (Y, X) are the ones from the non-perturbed training data. Therefore, we obtain predictive stability and protection against some worst case shifts on new test data.

The corner-point of the regularization on the right-hand side of (16) is with $\gamma = \infty$: b_∞ is the parameter b in

$$I = \{b;\ \mathbb{E}[A(Y - X^T b)] = 0\}$$

which minimizes the squared error risk. All the elements of I lead to an interesting invariance.

Theorem 3. *Rothenhäusler, Bühlmann, Meinshausen, and Peters [ibid.] Consider the model in* (4) *or* (5) *with* $(I - B)$ *being invertible. Then:*

$$b \in I \iff Y - Xb \overset{d}{=} Y^v - X^v b \textit{ for all } v \textit{ in } \mathrm{span}(M).$$

The result says that enforcing the constraint $\mathbb{E}[A(Y - X^T b)] = 0$ leads to invariance of the error terms, and protection against any shifts in $\mathrm{span}(M)$ is guaranteed.

5.3 Properties of high-dimensional anchor regression for finite samples.

We consider here the finite-sample estimator in (13). We make the following assumptions for the high-dimensional setting.

(A1) A, X, Y are jointly Gaussian, and the minimal eigenvalue of $\mathrm{Cov}(X)$ satisfies $\Lambda^2_{\min}(\mathrm{Cov}(X))$
$L > 0$;

(A2) $S_0(\gamma; q) = \mathrm{supp}(b_{\gamma;q})$ has cardinality $s_0(\gamma; q) = o(\sqrt{\log(d)/n})$ $(d \gg n \to \infty)$;

The Gaussian assumption in (A1) is only for technical simplicity and extensions to sub-Gaussian distributions are possible.

Theorem 4. *Rothenhäusler, Bühlmann, Meinshausen, and Peters [ibid.] Assume the model in* (4) *or* (5) *and that (A1)-(A2) hold. Consider the estimator in* (13). *Then, with probability tending to one as* $d \gg n \to \infty$, *we have that for any* $\gamma \geq 0$:

$$\|\hat{\beta}_{\gamma;q} - b_{\gamma;q}\|_1 \leq C \lambda s_0(\gamma; q)\ (q \in \{2, \infty\}),$$

$$\textit{for } q = 2{:}\lambda \asymp \sqrt{r \max(\log(r), \log(d))/n},$$

$$\textit{for } q = \infty{:}\lambda \asymp \sqrt{\log(r) log(d)/n}.$$

where $0 < C < \infty$ is a constant. Furthermore, for the risk $R(v, \beta) = \mathbb{E}[|Y^v - X^v \beta|^2]$ with shift v, we have that

$$\max_{v \in C_\gamma^q} R(v, \hat{\beta}_{\gamma;q}) \leq \max_{v \in C_\gamma^q} R(v, b_{\gamma;q}) + C g(\lambda) s_0(\gamma; q),$$

where $g(\lambda) = \lambda^2$ for $q = 2$ and $g(\lambda) = \lambda$ for $q = \infty$, with λ as specified above for $q \in \{2, \infty\}$. Note that $\max_{v \in C_\gamma^q} R(v, b_{\gamma;q}) = \min_\beta \max_{v \in C_\gamma^q} R(v, b_{\gamma;q})$.

For the case with high-dimensional anchors with $r \gg n$ we should choose $q = \infty$. For small values of $r = \dim(A)$, $q = 2$ seems to be a more natural choice in terms of the class C_γ^q for protection or predictive stability, see Theorems 2 and 3.

5.4 Predicting single gene knock out experiments. As described in Sections 5.1–5.3, the methodology and corresponding theory is tailored for predictive stability and prediction of new unseen perturbations. A score for measuring the effect-strength of a perturbation at covariate X_j for the response Y is given by the parameter $(b_{\gamma;q})_j$ in (16). Note that for $\gamma = \infty$ and in identifiable scenarios, $(b_{\gamma=\infty})_j = \beta_j^0$ equals the direct causal effect in (3).

We summarize some findings for predicting single gene knock out experiments in yeast (Saccharomyces cerevisiae) Meinshausen, Hauser, Mooij, Peters, Versteeg, and Bühlmann [2016]. The observed data is for the expressions of 6170 genes in yeast, and there are $n_{\text{observ}} = 160$ observational data points (from the system in steady state, without any interventions, from wild-type yeast) and $n_{\text{interv}} = 1479$ interventional data points, each of them corresponding to a single gene knock out experiment where a single strain has been deleted. The response Y is the expression of (say) gene k, and the covariates correspond to the expressions of all the genes without gene k, thus being of dimension $d = 6170 - 1 = 6169$. Consider this encoding into response and covariates for all $k = 1, 2, \ldots, 6170$, that is, the expression of each gene is once the response: and thus, we can predict the effect-strength of a perturbation at each gene to another one. The model in (4) or (5) is used with two environments corresponding to $r = 2$ binary components in A: A_1 encodes the observational data environments with 160 samples, A_2 is encoding all interventional data by (crudely) pooling all of them into a single environment with 1479 samples.

Holding out a random third (repeatedly three times) of the 1479 interventional samples enables us to validate the predictions. We aim to predict a response Y under an intervention at one of the covariates X_j in the hold-out data: the parameters for the prediction are trained (estimated) based on the 160 observational and two thirds of the interventional samples. Thanks to the hold-out data, we can then validate the performance of the prediction. We consider binarized outcomes: if the prediction for Y is large (in absolute value), we denote it as a predicted "positive" and otherwise as a predicted "negative". Similarly

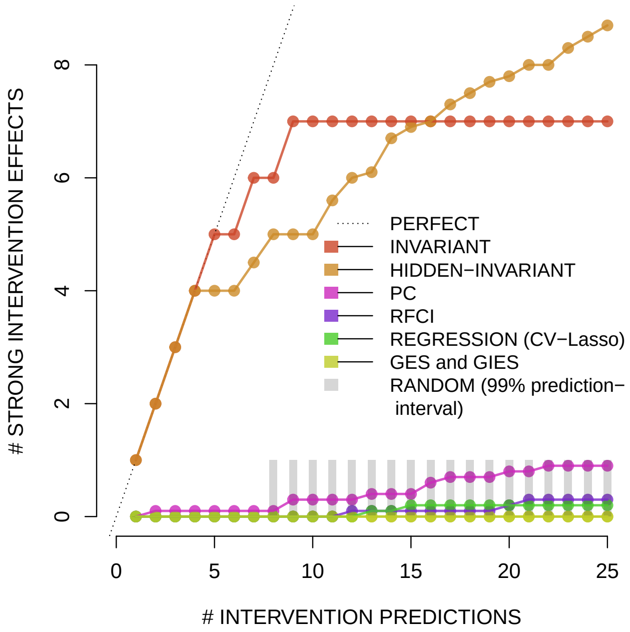


Figure 1: Prediction of single gene perturbations. x-axis: number of false positives; y-axis: number of true positives. Red (no hidden variables) and orange (including hidden variables) lines are algorithms exploiting (near) invariance similarly as described in Theorem 3. Other colored lines correspond to some competitor methods, and the gray bars indicate random guessing. The figure is taken from Meinshausen, Hauser, Mooij, Peters, Versteeg, and Bühlmann [2016].

for the true value in the hold-out observational data: if an intervention at a covariate has a strong effect on Y, we denote it as "true", otherwise as false. One can then validate methods and algorithms in terms of their capacity to predict "true positives" (predicted "positive" and actually being "true") in relation to "false positives" (predicted "positive" but actually being "false"). Figure 1 summarizes the results. The problem of correctly predicting unseen gene interventions is very ambitious: we predict only a few strong intervention effects, but the highest scoring predictions (the first 5 or 4, respectively) are all correct (i.e., "true").

6 Conclusions

Large-scale data with heterogeneity from different environments or perturbations provides novel opportunities for predictive stability and causal inference. The word "causal" is ambitious and perhaps a bit philosophical: in a nutshell, its meaning is to predict the outcome of an unseen (in the data) perturbation, a policy or treatment. This fundamental prediction problem is very different from the standard one where we want to predict an outcome from roughly the same population from which we have collected data. We are now just at the beginning of new developments of methodology, algorithms and fundamental mathematical understanding for statistical inference from heterogeneous large-scale data.

References

P. Bickel, Y. Ritov, and A. Tsybakov (2009). "Simultaneous analysis of Lasso and Dantzig selector." *Annals of Statistics* 37, pp. 1705–1732 (cit. on pp. 2804, 2812).

P. Bühlmann (2006). "Boosting for high-dimensional linear models". *Annals of Statistics* 34, pp. 559–583 (cit. on p. 2804).

P. Bühlmann and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer (cit. on p. 2804).

F. Bunea, A. Tsybakov, and M.H. Wegkamp (2007). "Sparsity oracle inequalities for the Lasso". *Electronic Journal of Statistics* 1, pp. 169–194 (cit. on p. 2804).

E. Candès and T. Tao (2007). "The Dantzig selector: statistical estimation when p is much larger than n (with discussion)". *Annals of Statistics* 35, pp. 2313–2404 (cit. on p. 2804).

E.J. Candès, J.K. Romberg, and T. Tao (2006). "Stable signal recovery from incomplete and inaccurate measurements". *Communications on Pure and Applied Mathematics* 59, pp. 1207–1223 (cit. on p. 2804).

E.J. Candès and T. Tao (2006). "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory* 52, pp. 5406–5425 (cit. on p. 2804).

D. Ćevid, P. Bühlmann, and N. Meinshausen (2018). *Work in progress* (cit. on p. 2812).

S.S. Chen, D.L. Donoho, and M.A. Saunders (2001). "Atomic decomposition by basis pursuit". *SIAM review* 43, pp. 129–159 (cit. on p. 2804).

V. Chernozhukov, C. Hansen, Y. Liao, et al. (2017). "A lava attack on the recovery of sums of dense and sparse signals". *Annals of Statistics* 45, pp. 39–76 (cit. on p. 2811).

D.L. Donoho (1993). "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data". In: *In Proceedings of Symposia in Applied Mathematics* (cit. on p. 2804).

– (2006). "Compressed sensing". *IEEE Transactions on Information Theory* 52, pp. 1289–1306 (cit. on p. 2804).

D.L. Donoho and X. Huo (2001). "Uncertainty principles and ideal atomic decomposition". *IEEE Transactions on Information Theory* 47, pp. 2845–2862 (cit. on p. 2804).

D.L. Donoho and J.M. Johnstone (1994). "Ideal spatial adaptation by wavelet shrinkage". *Biometrika* 81, pp. 425–455 (cit. on p. 2804).

R.C. Geary (1949). "Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown". *Econometrica: Journal of the Econometric Society*, pp. 30–58 (cit. on p. 2813).

C. Giraud (2014). *Introduction to High-Dimensional Statistics*. CRC Press (cit. on p. 2804).

E. Greenshtein and Y. Ritov (2004). "Persistence in high-dimensional predictor selection and the virtue of over-parametrization". *Bernoulli* 10, pp. 971–988 (cit. on p. 2804).

T. Haavelmo (1943). "The statistical implications of a system of simultaneous equations". *Econometrica*, pp. 1–12 (cit. on pp. 2803, 2807).

T. Hastie, R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press (cit. on p. 2804).

V. Koltchinskii (2009a). "Sparsity in penalized empirical risk minimization". *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 45, pp. 7–57 (cit. on p. 2804).

– (2009b). "The Dantzig selector and sparsity oracle inequalities". *Bernoulli* 15, pp. 799–828 (cit. on p. 2804).

N. Meinshausen and P. Bühlmann (2006). "High-dimensional graphs and variable selection with the Lasso". *Annals of Statistics* 34, pp. 1436–1462 (cit. on p. 2804).

N. Meinshausen, A. Hauser, J.M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann (2016). "Methods for causal inference from gene perturbation experiments and validation". *Proc. National Academy of Sciences USA* 113, pp. 7361–7368 (cit. on pp. 2805, 2816, 2817).

J. Pearl (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press (cit. on pp. 2803, 2806, 2807).

J. Peters, P. Bühlmann, and N. Meinshausen (2016). "Causal inference using invariant prediction: identification and confidence interval (with discussion)". *J. Royal Statistical Society, Series B* 78, pp. 947–1012 (cit. on pp. 2803, 2807).

J. Peters, D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference*. MIT Press (cit. on p. 2806).

N. Pfister, P. Bühlmann, and J. Peters (2017). *Invariant causal prediction for sequential data*. arXiv: 1706.08058 (cit. on p. 2805).

D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters (2018). *Anchor regression: heterogeneous data meets causality*. arXiv: 1801.06229 (cit. on pp. 2814, 2815).

R. Tibshirani (1996). "Regression shrinkage and selection via the Lasso". *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288 (cit. on p. 2804).

S. van de Geer and P. Bühlmann (2009). “On the conditions used to prove oracle results for the Lasso”. *Electronic Journal of Statistics* 3, pp. 1360–1392 (cit. on p. 2812).

P. Zhao and B. Yu (2006). “On model selection consistency of Lasso”. *Journal of Machine Learning Research* 7, pp. 2541–2563 (cit. on p. 2804).

H. Zou (2006). “The adaptive Lasso and its oracle properties”. *Journal of the American Statistical Association* 101, pp. 1418–1429 (cit. on p. 2804).

Received 2017-11-24.

Peter Bühlmann
Seminar for Statistics
Department of Mathematics
ETH Zürich
Switzerland
peter.buehlmann@stat.math.ethz.ch