What Can the Millions of Random Treatments in Nonexperimental Data Reveal About Causes?

Andre F. Ribeiro¹ · Frank Neffke¹ · Ricardo Hausmann¹

Received: 31 August 2021 / Accepted: 7 July 2022
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2022

Abstract
We propose a new method to estimate causal effects from nonexperimental data. Each pair of sample units is first associated with a stochastic ‘treatment’—differences in factors between units—and an effect—a resultant outcome difference. It is then proposed that all pairs can be combined to provide more accurate estimates of causal effects in nonexperimental data, provided a statistical model relating combinatorial properties of treatments to the accuracy and unbiasedness of their effects. The article introduces one such model and a Bayesian approach to combine the $O(n^2)$ pairwise observations typically available in nonexperimental data. This also leads to an interpretation of nonexperimental datasets as incomplete, or noisy, versions of ideal factorial experimental designs. This approach to causal effect estimation has several advantages: (1) it expands the number of observations, converting thousands of individuals into millions of observational treatments; (2) starting with treatments closest to the experimental ideal, it identifies noncausal variables that can be ignored in the future, making estimation easier in each subsequent iteration while departing minimally from experiment-like treatments; (3) it recovers individual causal effects in heterogeneous populations. We evaluate the method in simulations and the National Supported Work (NSW) program, an intensively studied program whose effects are known from randomized field experiments. We demonstrate that the proposed approach recovers causal effects in common NSW samples, as well as in arbitrary subpopulations and an order-of-magnitude larger supersample with the entire national program data, outperforming Statistical, Econometrics and Machine Learning estimators in all cases. As a tool, the approach also allows researchers to represent and visualize possible causes, and heterogeneous subpopulations, in their samples.

Keywords Causal effect estimation · Experimental design · Signal processing · Effect heterogeneity

Introduction
Most questions of interest in the social, behavioral and life sciences—What makes economies grow? What explains criminal behavior? What can prevent or cure a disease?—are ultimately questions about what causes an outcome of interest. Data used to answer such questions typically have no shortage of correlational patterns, but correlations are often poor guides to the causal process that produced them. The central methodological difficulty in the scientific inquiry remains that of estimating the causal effect of a treatment, or independent variable, on an outcome. Compared to the tremendous success of Machine Learning algorithms in prediction and pattern recognition tasks in correlation-rich data, such as in face recognition and textual topic modeling, Machine Learning approaches are still of limited use when estimating causal effects [1, 2]. This has led to a paradoxical situation: in the midst of the big-data revolution, many prominent scientists have declared randomized field experiments - with often just hundreds of participants—as the sole standard for empirical research [3, 4]. Experiments are attractive because once individuals have been randomly divided into treated and nontreated subgroups, it suffices to compare their average outcomes to estimate causal effects. Yet, randomized trials have many drawbacks: they are expensive, slow and sometimes impossible or unethical to carry out and they elucidate only if treatments work,
rather than why they work [3, 5, 6]. Furthermore, a focus on average effects creates problems when different individuals experience different effects. Indeed, Xie [7] (2016, p. 6263) considers this a fundamental conundrum: ‘the ubiquitous presence of individual-level variability [in social phenomena] makes it impossible to study individual-level causal effects. To draw a causal inference, it is necessary to pool information from different members of a population into aggregates’. As we will demonstrate, effect heterogeneity affects the accuracy of current observational methods even in datasets of moderate size. While discourse about causes are often dominated by all-or-nothing hypothesis testing in the Sciences, there is great practical need for tools that can introduce causal insights into the earlier phases of scientific discovery or provide insights from larger data. Here, we study a problem representation and method that facilitates the use of recent Machine Learning and high-dimensional techniques to that end.

Model Summary

In particular, we consider the problem of estimating Average Treatment Effects (ATE) or Individual Treatment Effects (ITE) of a given treatment on an outcome, \( y \in \mathbb{R} \). The problem of estimating the effect of a treatment-of-interest nonexperimentally has been studied extensively, in particular from comparisons between treated and non-treated subjects’ outcomes [8–10]. Although for nonexperimental estimation, these approaches often draw on Experimental Design concepts and have received attention, especially, in Econometrics and Applied Statistics.

The central goal of the present work is to better understand and exploit the heterogeneity of statistical conditions between pairs of individuals in everyday datasets, as they relate to causality. Consider a nonexperimental dataset with \( n \) observations and \( m \) variables, from a variable set \( \mathcal{X}^m \). An observed difference in outcome \( y_{ij} = y_i - y_j \) between any two individuals, \( 0 < i, j \leq n \), can carry both a lot or very little information about a variable, or variable subset, \( v \subset \mathcal{X}^m \). When individuals differ only by a single variable \( v = \{a\} \), strict claims can be made for the effect of \( a \), as the pairing characterizes an ideal counterfactual (given conditions reviewed below). More commonly, however, pairwise conditions are in a spectrum of usefulness for each variable and individual. Imagine the existence of a prior stochastic process that can describe how individual pairwise conditions translate to the validity of pairwise observations, \( y_{ij} \), for each variable subset \( v \). Such model would allow us to use all pairs, and the full range of experiments ‘run by nature’, to help estimate causal effects.

An obvious analogy is Signal Processing and Fusion [11, 12], where the use of millions of (noisy) observations often outperform any individual observation by orders of magnitude—provided a suitable prior statistical model for measurements. We characterize all population pairwise conditions by considering all their possible combinations of differences and intersections. The model then connects such combinatorial conditions to the validity of pairwise estimates. The effect \( f(v) \) of variables \( v \) is first described as \( f(v) \sim \mathcal{N}(y_{ip}, \sigma_y(v)) \), which carries the assumption of Gaussian measurements having a common mean, for each \( v \), but distinct standard deviations \( \sigma_y \) across pairs. In datasets below, there are in the order of 50–200 \( M \) such pairs. David Mckay provides an excellent allegory to the problem (the ‘7 scientists problem’) [11]. Each of seven scientists draw \( n \) datapoints from a distribution, all of which are Gaussian with a mean \( y \) but with unknown standard deviations \( \sigma_y \). The scientists measure the same \( y \) but have widely-different experimental skills. You expect some of them to do accurate work, and some to turn in widely inaccurate answers (i.e., to have enormous \( \sigma_y \)). What is the maximum a posteriori estimation of \( y \) given observations?

The set of attributes \( x_i \subset \mathcal{X}^m \) that characterize an individual \( i \) is related in some unknown way to his or hers observed outcome \( y_i \). This relationship can be described abstractly by functionals \( g(x), h(x), \ldots \) (such as the complex input-output map of a classifier, regressor, estimator, etc.) Consequent statistical pitfalls, such as omissions, endogeneity, selection biases, are examined as we progress and implement the model. In a typical setting, we have \( y \sim g(x) + \epsilon \) where \( \epsilon \) is some irreducible, additive and typically Gaussian error. Under correct specification, \( \epsilon \) is independent (of \( y \) and \( x \)) and small for the sample. We interpret \( y_{ij} = y_i - y_j \), instead, as a noisy observation of the true causal effect of observed factor differences, \( x_i \Theta x_j \), where \( \Theta \) indicates (asymmetric) set-difference. Noise increases as \( x_i \Theta x_j \) departs from experiment-like conditions. That is, we interpret differences in attributes between a pair of individuals, \( x_i \Theta x_j \), as ‘treatment' differences that cause differences in outcomes \( y_{ij} = y_i - y_j \).

\[
y_{ij} \sim g(x_i \Theta x_j) + \epsilon_{ij}, \\
\epsilon_{ij} \sim \mathcal{N}(0, h(x_i \Theta x_j)), \\
\]

where \( \epsilon_{ij} \) is a Gaussian noise with mean 0 and variance \( \sigma^2 = h(x_i \Theta x_j) \). Equation (1) postulates that \( \epsilon_{ij} \) reflects distortions in observed effects \( y_{ij} \) due to variables that individual \( j \), alone, has. We will say that, when \( \epsilon_{ij} = 0 \), pairwise observations correspond to factorial treatments; pairwise observations with little risk of observed confounding. Or, similarly, that each observational pair, \( x_i \Theta x_j \), in the sample represents a factorial treatment, \( v = x_i \Theta x_j \), with probability \( p_v(x_i, x_j) \approx 1 \).

Similar to typical data fusion problems, we choose, therefore, to model \( \epsilon_{ij} \) and not \( g(x_i \Theta x_j) \). Any estimator we choose to estimate the effect of a treatment, \( x_i \Theta x_j \), can be used to also estimate the expected effect of the pairwise confounding set, \( h = g(x_i \Theta x_j) \). The second is very useful because
it allows us to determine how much \( y_{ij} = y_i - y_j \) is, in turn, useful, as a non-parametric effect statistic. The proposed approach inherits any assumptions the chosen estimators make—such as those related to sample selection or variable omissions—but uses, additionally, sample and population combinatorial properties and all pairs.

This looks at sample differences as counterfactual (noisy) observations with associated statistics. Although often averaged across samples, each pairwise outcome difference \( y_{ij} \) is, fundamentally, an statistic pertaining to the pair, alone, and not the sample. Similarly, how confounded this effect estimate is is a property of the pair, and not the sample. The main identification assumption in this work is \( (x_i \ominus x_j) \perp (y_i - y_j) \mid (x_j \ominus x_i) \), Fig. 1a, left. This is the condition that potential outcomes for the two units are independent of their pairwise treatment, \((x_i \ominus x_j)\), given its commuted factor set, \((x_j \ominus x_i)\). Whenever this condition holds, we are able to say that \((x_i \ominus x_j)\) had effect \( y_{ij} = y_i - y_j \). If \textit{treat} is a fixed treatment indicator, the main difference between this assumption and typical sample treatment assignment conditions \cite{13}, \( \text{treat} \perp Y \mid X \), is that the set of factors \((x_i \ominus x_j)\), enabling pairwise unconfoundedness, will be different for different pairs. When sample units share unobserved common factors, those factors are not confounders...
but simply common causes. This is known as the Common Cause Principle and was fundamentally formulated by Reichenbach [14–16]. Adjusting for $x_i$ is sufficient to make the causal effect of $x_i$ unconfounded and identifiable. This is however not the case if two units differ by some factors. In this more severe case, we say that the pairwise outcome difference $y_{ij} = y_i - y_j$ is not pairwise valid, introducing a measurement error $\hat{h}(x_i \otimes x_j)$ in estimation. That is, consider a pair with $|x_i \otimes x_j| > 0$ (i.e., a valid treatment for $i$). When $|x_i \otimes x_j| = 0$, the treatment can be seen as fully ‘concentrated’ on $i$. Under ignorability, we can confidently take $y_{ij} = y_i - y_j$ as a maximally accurate (heterogeneous) effect of $(x_i \otimes x_j)$ on $i$. When $|x_i \otimes x_j| > 0$, however, the multivariate ‘treatment’ has been spread across different individuals and the scalar difference $y_{ij} = y_i - y_j$ now carries information about $(x_i \otimes x_j) \cup (x_j \otimes x_i)$. With increasing $|x_i \otimes x_j|$, the statistic $y_{ij}$ is increasingly, and simultaneously, uninformative for both sample units. This is illustrated in Fig. 1a, right for 1000 simulations with $n = 600$ and $m = 10$ Binomial variables, where both probabilities and effects are sampled from $U([0, 1])$ and outcomes are sigmoidal. When $(y_{ij} - y_{uv})^2 \approx 0$ for any two sample pairs $(ij$ and $uv$), the effect observed by the first pair is transferable to the second—to mean that effects observed for one pair are valid estimates for the other. Empirical sample pairwise errors, $(y_{ij} - y_{uv})^2$, follow $|x_i \otimes x_j|$ and, consequently, a hypergeometric distribution [19] with these parametric assumptions.

For each sample pair we thus have $y_{ij} = y_i - y_j$ (scalar), $x_i \otimes x_j$, $x_j \otimes x_i$, $x_i \cap x_j$ (factor sets). Regularities on the latter three help characterize the first statistic (and its usefulness). Figure 1b illustrates this sample decomposition. The approach articulates statistical roles for all possible combinatorial conditions appearing in pairings, such as when there are extraneous varying variables, $x_i \otimes x_j$, as well as multivariate treatments, $x_i \otimes x_j$. The formulation of a random treatment model enables the use of all available treatments and treatment types (such as multivariate treatments) in datasets. Estimation can also output, as a consequence, not only effect estimates, but also confidence measures in the form of accumulated (pairwise) confounding. This combinatorial decomposition also leads to a simple geometrical interpretation of observational pairs and a density $p_{\lambda}(x_i, x_j)$ that indicates pairs’ departure from ideal factorial treatments: univariate treatments with no pairwise confounders.

A popular recent strategy [20–22] is to stipulate prior models assuming a proxy confounding variable (either observed or unobserved) in the data. ‘Latent-Confounder’ approaches require assuming and training complex functional models to infer the latent variable, then adjusting effect estimates accordingly. While sometimes calling confounders ‘unobserved’, these parametric estimators assume partially-observed confounding variables (whose correlations with observed variables can be exploited when training a proxy). In practice, pairs of individuals vary in how they differ, and consequently in what variables can confound estimates in each case. Instead of requiring a single variable to account for this potentially complex and contingent relationship, a prior statistical model for pairwise conditions allows estimates to be considered in a pair-by-pair case—simply penalizing pairs, or observations $y_{ij}$, that are ruled unfit in some fundamental manner. Using the proposed combinatorial model, we show that even simple estimators can outperform current complex estimators (e.g., encompassing specific statistical pitfalls like endogeneity and omitted variables) when replicating the outcome of the studied Randomized Experiments.

Most current causal effect estimation solutions demand researchers to specify ex ante the causes, confounders, and/or a parametric form to outcomes in the problems they are studying. These requirements put, however, ‘the cart before the horse’: most research is undertaken because these are unknown. We consider to what extent combining all available treatments can help estimate effects autonomously and non-parametrically. This article’s main contribution is therefore to demonstrate the practical benefits of modeling and combining large samples of disparate pairwise observations. Beyond practical usefulness, understanding how population combinatorial patterns relate to causality is also of great importance to population-based research, such as the study of genetic variation in the Medical Sciences and demographics in the Social Sciences. We discuss the implications of concepts developed here to the study of genetic population variation, and Genome-wide Association studies, in [23].

**Reproducing Effect Estimates from Large Randomized Experiments**

We assess the proposed method’s performance in simulations and a seminal real-world example, comparing it to current Statistics, Econometrics and Machine Learning estimators. We demonstrate that the proposed approach also remains accurate in heterogeneous and diverse samples. The simulations introduce confounders and heterogeneous subpopulations into synthetic data, demonstrating that observational methods generally become biased or inaccurate, unlike the proposed. As a real-world application, we consider the National Supported Work (NSW) program. Starting with a seminal contribution by Lalonde [24], studies have used this randomized control trial to benchmark non-experimental techniques—including an historical ‘face-off between regression and propensity-score matching’ [25]. Observational methods generally fail to recover the experimental

---

1 The assumption that $\lambda^m$ includes all factors associated both with treatment assignment and outcomes [17, 18].
causal effect estimate, except in a smaller handpicked NSW subsample [10, 26–28]. This literature exemplify a typical scenario across disciplines: estimating causal effects nonexperimentally require several (hard to justify) population and variable selection assumptions (in this case, expert selection of samples and variables with desirable economic characteristics). We demonstrate, however, that the proposed approach can recover the NSW experimental effects not only in Lalonde’s original unsolved challenge (with 740 participants and 6 variables) but also in the full NSW data with over 10000 participants and 1000 variables, without an extensive specification from researchers.

We compare the approach to a range of previous solutions, including those making typical ignorability assumptions, as well as approaches relaxing other assumptions, such as missing variables and endogeneity. We show that the previous simple model outperform these solutions, when trying to reproduce the results of the previous Randomized Experiment. Our initial goal was to address the common case of large datasets with many variables. Perhaps surprisingly, however, this is also the case for samples with very few variables, where Matching, Doubly-Robust and Latent-Confounder approaches should be most relevant.

**Stochastic Factorial Estimation (SFE)**

Randomized Controlled Trials (RCTs), and the observational estimators they have inspired [29], often focus on the causal effect of a single treatment or intervention. However, observed outcomes are often the result of many interacting causes. This limitation of RCTs had already been noted by Fisher in 1926: ‘No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time.’ [30] (1926, p. 503) Instead, he proposed submitting Nature ‘logical and carefully thought out questionnaire[s],’ leading to factorial experimental designs. Factorial designs have since been mostly studied for the design or analysis of experiments [31].

A factorial experiment is a complex experiment consisting of many runs. It is designed to estimate the causal effect of \( m \) factors on an outcome of interest \( y \). When factors are binary, the design contains \( 2^m \) factorial runs, or, possible factor combinations. Figure 1c depicts geometrically a 3-factor design with a cube \( C : \{-1, +1\}^3 \). We call the set of factors in which two runs differ a factorial treatment. A factorial run corresponds to \( C \)'s vertices and treatments to edges. We consider all individuals in a nonexperimental dataset as stochastic factorial runs and the entire dataset as an incomplete random factorial design. Critically, the full set of observed factorial treatments express necessary combinatorial patterns of variable variation and fixation necessary to make claims about each variable’s piecewise effects on \( y \).

We consider this geometric representation for the nonexperimental causal effect estimation problem next, then discuss a Bayesian procedure that combines effect estimates from each pair of individuals, given how strongly they depart from ideal factorial treatments. We call the resulting method Stochastic Factorial Estimation (SFE).

**Geometric Representation**

Consider a Factorial Experiment studying the effects of a set of factors \( \mathcal{X}^m = \{a, b, c, \ldots\} \). The experiment’s runs are the vertices of the cube \( \mathcal{C} = \mathcal{X}^m : \{-1, +1\}^m \). Factorial treatments are pairs of runs that differ on a set of factors (the ‘treatment’), while having all other factors in common. Namely, let \( x_i, x_j \in \mathcal{C} \) be runs and their corresponding treatment be the set of factors run \( i \) has exclusively, \( x_i \ominus x_j \). In addition, we let the size of \( C \)'s edges correspond to the causal effect, \( f(x_i \ominus x_j) \), of their associated treatment. We discuss an extension to the continuous variables case in Appendix 2.

Consider now an observational matrix \( X \) with variables \( X^m \). Figure 1c depicts an example with \( n=8 \) individuals and \( m=3 \) variables, where values have been normalized to the unit interval, \( X : [-1, +1]^m \). The \( O(n^2) \) pairs of individuals are in a myriad of configurations. As a result, different pairs are useful for estimating effects of different variables. Figure 1d illustrates a pair corresponding to a factorial treatment with \( \{a\} \) as treatment. The pair captures the main intuition behind factorial designs: a single variable, \( a \), differs between individuals while all other variables are fixed. Figure 1e depicts another factorial treatment. Here, however, the treatment is multivariate, \( \{a, b\} \). Because the treatment consists of two potential causes, it is impossible to infer their separate effects from the pair alone. However, we can still learn about their combined effect. Figure 1f shows an imperfect factorial treatment. There, the outcome difference is not necessarily due to variables in the treatment and may reflect extraneous variation from other variables. Although not fixed within the pair, these nontreated variables could coincide in expectation across treated and nontreated individuals—i.e., they could be ‘balanced’ in the sample. We can also learn from pairs in this case. Finally, Fig. 1g illustrates a pair without treatment. We disregard such cases in the estimation. We formalize these pairwise conditions further in the next section.

**Treatment Balance and Size Penalties**

In the previous model, Eq. (1), variables \( x_i \ominus x_j \) were assumed capable of confounding pairwise estimates, affecting both the treatment and outcome at the same time. A first way to estimate effects \( f(y) \) is to focus on pairs that
approximate a given factorial treatment \( v \) with near certainty: \( p_j(v) \to 1 \). The defining characteristic for this type of pair is that, when estimating the effect for an individual \( i \), the other individual \( j \) has no observed extraneous factors that could confound the effect \( y_{ij} \), \( x_j \perp x_i \), or \( |x_j \perp x_i| = 0 \). This is the first key condition behind Factorial experimentation. In this first condition, the experimenter keeps all relevant conditions fixed, except for a treatment. This leads to a simple but also conservative model, where the risk of (observed) confounding increases with extraneous variables between pairs. We define penalties describing the extent to which pairs deviate from these ideal treatments. The central penalty reflect Eq. (1): pairs with any pairwise confounders, \( x_j \perp x_i \), are strongly penalized. We then place two priors over \( x_j \perp x_i \) which also allow pairs with pairwise confounders to be used to estimate effects, conditional on their sample balance, \( \phi_{ij}^{\text{bal}} \), and treatment size, \( \phi_{ij}^{\text{size}} \).

Each added assumption thus further scales the number of useful pairs. The first prior condition, \( x_j \perp x_i \perp y \mid x_j \perp x_i \), is illustrated in Fig. 1a, left. With the assumption, pairs where treatment-outcome variables are independent, conditional on non-treated variation, are now also deemed useful. For the full sample and specific treatment case, this condition is, abstractly, ‘the most developed and popular strategy for causal analysis in observational studies’ [32, 33]. It is used by a wide range of methods across disciplines. It is often seen as means to emulate randomized experimentation. In randomized experiments, the randomization enables unbiased estimation of treatment effects across local population groups. For each observed variable, randomization implies, as a simple application of the law of large numbers, that a treatment-subgroup will be ‘balanced’ on average. The assignment of treatments to subjects in observational datasets is typically not random. Most causal effect estimator in use today attempt to reduce treatment assignment bias by increasing a balance score between treatment and control units in use. The idea is to create a subsample of units that received the treatment that is comparable on all observed covariates to units that did not receive the treatment. Most of these estimators requires, however, a further ‘ignorable treatment assignment’ assumption [20, 29]. Namely, it requires that, conditional on the observed variables, there are no unobserved differences between treatment and control groups. The common way to satisfy this assumption is to include in \( \mathcal{X}^{\text{obs}} \) any variable that affect either outcomes or treatments.

In the present approach, the notion of sample balance thus leads to an observational pair’s balance. A balancing score \( b(X) \) is a function of the observed covariates \( X \) such that the conditional distribution of \( X \) given \( b(X) \) is the same for treated and control units—thus reflecting the previous assumption. We assign a penalty \( \phi_{ij}^{\text{bias}} = b(x_j \perp x_i) \) for pairs, which will later be used in optimization. We let \( p(\phi_{ij}^{\text{bal}} = 1) \) denote the probability that the the observational pair \( (ij) \) has a balanced treatment. These are formulated as priors over \( |x_j \perp x_i| \) which means that balance statistics are only used when there are no better pairs (i.e., fully pairwise unconfounded for the same \( v \)). A consequence, demonstrated below, is that balance calculations are mostly useful in small samples. There are three general approaches to derive balance functions. The first and most trivial function is to use \( b(X) = X \), which is the case of exact matching [34].

A second approach, which broadly underlie popular Propensity Score [8, 13] and Latent-Confounder estimators [20–22], is to use dimensionality reduction techniques to define a simpler function \( b(X) = h'(X) \). With this function, it is easier to balance samples, \( \{v \perp y \mid X \rightarrow v\} \approx \{v \perp y \mid h'(X \rightarrow v)\} \) where \( v = x_j \perp x_i \). In Propensity scores or Latent-Confounder approaches, treated and untreated individuals with the same \( h'(x_j \perp x_i) \) are expected to have similar distributions across any observed baseline covariates. Conditioning on the propensity score can also be seen as blocking the backdoor path between \( v \) and \( y \) [35]. On the other hand, the approach requires not only ignorability, but also more complex models, both statistically and computationally. With ignorability and these parametric assumptions, it has been shown analytically [13] that Propensity scores are the ‘coarsest’ balancing function taking the multidimensional \( X \) into one dimension. It uses a logistic regression to calculate the probability of a unit being assigned to a particular treatment, given \( X \). There are serious questions as to whether the previous goal can typically be achieved in practice, and whether these methods’ assumptions, in fact, hold [26, 33, 36, 37].

A third approach is to use explicit, or non-parametric, balance functions. There are additional analytic advantages to this approach [33]—such as, concurrently, decreasing model dependence. Due to the focus on large numbers of pairwise observations, we favor this alternative. That is, for a large number of pairs, it is undesirable to run a large number of regressions. The choice reflects the present alternative Signal Processing perspective, as opposed to the more typical, based on Model Inference. Deriving this non-parametric measure involves assessing whether the distributions of factors are similar between the treated \((v)\) and control groups \((\overline{v})\) for each pair in the sample. Many statistics for balance assessment have been described in the literature. They include: Standardized mean differences, Variance Ratios, Cumulative Density Functions (CDFs), Prognostic scores and more specialized high-dimensional statistics. The standardized mean difference (SMD) is the standardized difference in the means of each factor between groups. It is standardized so that it is on the same scale for all factors. The standardization factor is typically the pooled standard
deviation across both groups. Several recommended thresholds have been published in the literature, with 0.1 and 0.05 indicating prognostically important factors. Several empirical studies have examined the appropriateness of using SMDs in balance assessment, including [36, 38]; in general, they are useful as it’s been shown that the mean or maximum absolute SMD is often highly correlated with the degree of bias in treatment effects. The Variance Ratio is the ratio of the variance of the factor in one group to that in the other. Variance ratios close to 1 indicate good balance because they imply the variances of the samples are similar [39]. Statistics related to the difference in the empirical CDFs for each factor between groups allow assessment of imbalance across the entire factor distribution of that factor rather than just its mean or variance. The maximum CDF difference, also known as the Kolmogorov-Smirnov statistic, is sometimes recommended as a useful supplement to SMDs for assessing balance [37] and is often used as a check in propensity score methods [40, 41]. The prognostic score is an estimate of balance [37] and is often used as a check in propensity score methods [40, 41]. The prognostic score is an estimate of balance [37] and is often used as a check in propensity score methods [40, 41].

The implicit goal of causal effect estimation is to devise effect estimates with high external validity. It is worth considering the impact of multivariate treatments on external validity. Multivariate treatment effects estimate the simultaneous effects of all variables in the treatment, \( x_i \otimes x_j \). Effects need not generalize to the \( x_{1i} \otimes x_{1j} \) different instantiations of the treated variables. Under multivariate treatment conditions, it is impossible to attribute effects to any single cause. As a consequence, the cardinality of a treatment, \( |x_i \otimes x_j| \), is inversely related to the external validity of the derived effect estimates. The notion leads to the pairs’ treatment size, \( \phi^+ = |x_i \otimes x_j| \). We let \( \phi^+ = 1 \) denote the probability that the the observational pair \((ij)\) has a univariate treatment, indicating the propensity for higher external validity.

### Optimization

We consider that \( p_j(x_i, x_j) \to 1 \) when conditions \( \phi^+ = 1 \) and \( \phi^b = 0 \) are fulfilled by stipulating Bayesian priors for: sizes of factorial treatments, \( x_i \otimes x_j \), and balance of non-factorial variations, \( x_i \otimes x_j \). This Bayesian formulation leads to an objective function \( I(x) \) over individual positions \( x \), that we minimize. The overall procedure transforms pairwise distances \( x_i \otimes x_j \) in \( X \) to reflect observed outcome differences, \( y_{ij} \), according to the smallest and most balanced factorial treatments. Problems of integrating noisy diagnostic measurements are sometimes called ‘Inverse problems.’ Tikhonov regularization approaches, such as Ridge and LASSO regressions, are popular solutions to inverse problems. They perform both variable selection and regularization in a common framework to explain noisy observations. Regularized inverse problems can be seen as special cases of Bayesian inference [47, 48].

Equation (1) implies the following likelihood for expected effects\(^2\)

\[
\prod_{i,j} \mathcal{N}(y_{ij} \mid p_j(x_i, x_j)f(x_i \otimes x_j), h(x_i \otimes x_j)),
\]

for every pair \(0 < i, j \leq n\). To consider also learning from pairs with balanced treatments, we introduce a Gaussian prior \( \mathcal{N}(x_i \otimes x_j \mid 0, \phi^b) \) for the probability \( p_j(x_i, x_j) \), where \( \phi^b \) is a strictly positive scalar for each pair. If not a factorial treatment, the probability that the pair represents the treatment \( x_i \otimes x_j \) depends on the likelihood that non-common factors in the pair are balanced in the rest of the sample. Adding the prior to Eq. (3) leads to

\(\text{2 The typical likelihood notation } \mathcal{N}(y \mid \mu, \sigma^2), \text{ for the likelihood of an observation } y, \text{ given a mean } \mu \text{ and variance } \sigma^2, \text{ is used.} \)
\[
\prod_{i,j} \mathcal{N}(y_{ij} \mid p_{ij}(x_i, x_j)f(x_i \otimes x_j), \phi^c_{ij}) \times \mathcal{N}(x_r \otimes x_j \mid 0, \phi^b_{ij}), \tag{4}
\]

where \(\phi^c_{ij} = h(x_i \otimes x_j)\) was taken as the pair’s variance, Eq. (1).

Considering a single position \(x_i\) and its treatment \(v = x_i \otimes x_j\), our goal is to transform \(x_i\) such that \(|x_i - x_j|^2 = p_{ij}(x_i, x_j)f(x_i \otimes x_j)\) for all \(0 < j \leq n\). Taking logarithms and dropping constants from Eq. (4) leads to the objective

\[
\Gamma(x_i) = \min_{x_i} \sum_{j=1}^{n} (1 + \phi^c_{ij})(x_i, x_j) + |y_{ij}|^2 + \phi^b_{ij}(x_i, x_j)^2,
\]

\[
\hat{\phi}^c_{ij} = \frac{|x_i + x_j|}{m},
\]

\[
\hat{\phi}^b_{ij} = \left\lfloor \frac{1}{n} \sum_{k=0}^{n} (x_i, x_k + x_j) \right\rfloor.
\]

where \(\langle \ldots \rangle\) is the dot-product, \(\phi^c_{ij}\) and \(\phi^b_{ij}\) are treatment size and balance estimates. When both \(\phi_{ij}\) terms are zero and \(\langle x_i, x_j \rangle\) is negative, the pair corresponds to a factorial treatment, Fig. 1d–g, and the residual \(\langle x_i, x_j \rangle + |y_{ij}|\) is minimized. Term \(\phi^c_{ij}\) penalizes unbalanced non-factorial treatments, Fig. 1f, and the consequent risk of confounded estimates (given the discussed assumptions). Term \(\phi^b_{ij}\) penalizes multivariate treatments, Fig. 1e, and the consequent risk of low external validity. The estimators illustrated in this case, \(\hat{\phi}^c_{ij}\) and \(\hat{\phi}^b_{ij}\), are simple treatment size and balance estimators derived directly from the previous geometrical representation. This multivariate objective over sample units can be minimized with Stochastic Gradient Descent. Appendix 1 contains further discussions and pseudo-code listings for the approach.

In the output space, \(T_i(X)\), the Average Treatment Effect (ATE) of any factor \(a\) can be calculated simply as the difference in coordinates \(a\) between the mean position of all individuals with factor \(a\), \(\hat{X}^+a\), and those without, \(\hat{X}^-a\).

\[
ATE(a) = \sqrt{|\hat{X}^+a - \hat{X}^-a|}(a), \tag{6}
\]

for any \(a\) and where \(\hat{X} \in T_i(X)\). That is, in \(T_i(X)\) coordinate differences correspond to treatment effects and distances to outcome differences (the squared-sum of treatment effects). Since each factor \(a\) with non-zero effects divides a population in two subpopulations, \((+a, -a)\), the method’s output also gives researchers means to represent and visualize relevant subpopulations in their samples.

### Results

#### Simulation

We first test SFE in synthetic datasets. The design is similar to previous studies [26, 49]. In the first simulations, the treatment effect is homogeneous. In a second scenario effects are heterogeneous in 3 subpopulations. Each of these 3 subpopulations are, also, randomly underrepresented in treatment assignment. We simulated 1000 data sets of size 1000, comprised of a continuous outcome \(Y\), a binary indicator of treatment \(treat\), and three baseline factors \(X = \{a, b, c\}\). Specifically, \(X\) are normal random variables and the \(treat\) is binomial with a probability of success dependent on \(X\), and the outcome \(Y\) is random normal with mean dependent on \(X\) and \(treat\). Specifically, \(treat\) was generated as a Binomial random variable with success probability \(P(treat \mid X) \sim 0.5 - U((0,0.3))\). This describes a scenario where subpopulations deviate uniformly from the perfectly balanced scenario, \(P(treat \mid X) \sim 0.5\). Each subpopulation is therefore unbalanced randomly and independently.

We employ ATE estimators that adjust for covariates based on the treatment mechanism, the outcome mechanism, or both (i.e., doubly-robust methods), from several popular methods, which includes all methods in [10] plus 5 methods making use of Machine Learning (superlearner, genetic, latent, g-computation, sfe), Distance Metric Learning (genetic, latent, sfe) and epidemiological (g-computation) techniques.\(^3\) Methods are: propensity-score matching (ps), propensity-score with inverse probability of treatment weighting (IPTW) (ps-iptw), mahalanobis covariate matching (mahab) with 1-3 neighbor matchings, deconfounder (deconfounder) [21], optimal matching (optimal), ordinary least-squares (ols), genetic balance optimization (genetic) [41], SuperLearner ensembles (superlearner) [10, 50], high-dimensional g-computation (g-computation) [51], latent causal variables deep-learning (latent) [20] and stochastic factorial estimation (sfe), using Eq. (6). Ordinary least-squares is a maximum likelihood linear regression with all available variables (i.e., a single stratum analysis of variance). Formally, the propensity score is \(p_i = P(treat_i = 1 \mid x_i)\), i.e. the probability that unit 0 < \(i \leq n\) will be treated according to his or hers characteristics \(x_i\) at the time of the treatment assignment. It is often estimated using a logistic regression. Inverse probability weighting (IPTW) then makes it possible to reduce confounding by correcting the contribution of each subject \(i\) by a weight \(w_i\). For ATE, Xu et al. [52] defined \(w_i = [treat_i \times P(treat_i = 1)/p_i] +[(1 - treat_i) \times P(treat_i = 0)/(1 - p_i)]\). Stabilized weighting has been shown to produce a suitable estimate of the

\(^3\) See [10] for further algorithmic details.
variance even when there are subjects with extremely large weights [53]. The multivariable logistic regression \( \logit(P(Y = 1 \mid X, \text{treat}) = \gamma \text{treat} + \beta X) \) is often called the Q-model [51]. Once it is fitted, one can compute for all sample units \( \hat{P}(y_i = 1 \mid \text{do(treat} = 1), x_i) \) and \( \hat{P}(y_i = 1 \mid \text{do(treat} = 0), x_i) \), i.e. the expected, counterfactual outcome probabilities as if they were treated or not [51]. For ATE, one can obtain, as result, \( \pi_1 = n^{-1} \sum \hat{P}(y_i = 1 \mid \text{do(treat} = 1), x_i) \) and \( \pi_0 = n^{-1} \sum \hat{P}(y_i = 0 \mid \text{do(treat} = 0), x_i) \). ATE corresponds to the difference between these means, \( \hat{f}(\text{treat}) = \pi_1 - \pi_0 \).

We obtained \( \text{Var}[\hat{f}(\text{treat})] \) by simulating the parameters of the multivariable logistic regression assuming a multinormal distribution [54]. Both the deconfounder [21] and latent deep-learning [20] infer a latent variable as substitute for unobserved confounders, to perform causal inference. They make different different model assumptions (the latter focusing on the multiple-cause scenario), and use different techniques (Bayesian predictive model checking and deep learning). The specific deconfounder model employed is described in Sect. 3.1 of [21] (linear Bayesian factor model fit with Variational Bayes, logistic outcomes, and Normal priors).

Super-learner is a general two-stage substitution estimator. In the first stage, it models outcomes as a function of treatment assignment and factors. The second stage is a bias reduction step that iteratively updates the parameter estimates using models of treatment assignment given factors. This updating step also makes the estimator double-robust, asymptotically normal, and asymptotically efficient. Super-learner is typically implemented with semi-parametric Ensemble methods. It begins by fitting the Q-model to estimate the two expected hypothetical probabilities \( \pi_1 \) and \( \pi_0 \). The additional “targeting” step involves estimation of the treatment assignment mechanism, i.e., the propensity score \( P(\text{treat} = 1 \mid x) \), which is then used to update the initial estimates. In the presence of residual confounding, the propensity score provides additional information to improve the initial estimates. Finally, the updated estimates of \( \pi_1 \) and \( \pi_0 \) are used to generate estimates for \( \hat{f}(\text{treat}) = \pi_1 - \pi_0 \). We used the efficient influence curve to obtain standard errors.

A recent tutorial provides a step-by-step guided implementation of Super-learner [51, 55].

Figure 2a shows the Data Generating Process (DGP) in graphical-model notation for the first set of simulations and Fig. 2b ATE bias (difference from ground-truth) and their Mean Squared Errors (MSE) from distinct methods. The ground-truth counterfactual outcomes \( y^{\text{treat}} \) and \( y^{\text{control}} \) for each individual and \( \text{ITE}(\text{treat}) = y^{\text{treat}} - y^{\text{control}} \) are known. In this setting, all methods recover this ground-truth (dashed line) with little bias. SFE has, however, the smallest MSE, even below the ground-truth’s MSE, illustrating the advantage of using many pairs.

In the next simulations, we assume that subpopulations respond differently to the treatment. Subpopulation \( b \) observes double the expected effect, whereas \( c \) is immune. The resulting heterogeneity introduces significant biases in most estimates, Fig. 2c. Whereas most estimators become increasingly inaccurate as populations become more diverse, SFE continues to provide accurate and unbiased ATE estimates.

Figure 2e shows another set of simulations, now with 10 Binomial subpopulations (in addition to the treatment/control group), effects are additive and outcomes sigmoidal. Subpopulation probabilities and effects are both sampled from a Uniform distributions over \([0, 1]\). Fig. 2d shows the DGP. The horizontal axis shows increasing number of observations, \( n \). The panel on the right shows ATE biases for SFE with distinct balance penalties, \( \phi^b_m \). It demonstrates that, in the proposed system, the geometric estimator in Eq. (2) outperforms IPTW-weighted propensity scores and SMD balance estimates. Notice that the first requires \( O(n^2) \) logistic regressions while the geometric estimator requires the same order of matrix multiplications. The figure also shows the case when \( \phi^b_m \) is not used, illustrating that the bulk of gains from the model comes from Eq. (1). The figure demonstrates that while a central factor in the present work is factorial completeness, the assumption, in fact, impact other methods more severely. Making such assumptions explicit is, we believe, one of the proposed representation’s strengths. This additionally illustrates that (1) in larger samples, ‘causal estimation without inference’ (from difference observations alone without prior model assumptions) is a possible approach, (2) for smaller samples, inference and sample balancing is, reversely, useful, and that (3) the proposed approach combines the two cases.

Figure 3a illustrates SFE’s factorial representation. It shows a 3D subspace of the estimated space \( T_i(X) \). Dots show individuals’ positions and their outcomes (colors). Spatial differences in \( T_i(X) \) reflect differences in outcomes \( y \). The \( \text{ATE} \) corresponds to differences in the \text{treat} coordinate between treated and nontreated subpopulations, Eq. (6). Results demonstrate that SFE achieves lower MSE in homogeneous synthetic samples and can recover unbiased individual effects under heterogeneity.

**National Supported Work (NSW) Program**

We now consider a real-world application: the NSW employment program (details in the Appendix). Where eligible applicants were randomized into treatment and control groups. In his seminal article [24], Lalonde selected a subsample of the NSW participants and replaced its nontreated subgroup with samples from national surveys, leading to 6 distinct datasets. By doing so, he
'unbalanced' the NSW data (i.e., subpopulations' treatment propensities)—previously balanced by the NSW’s experimental design. Lalonde then showed that observational methods failed to recover the experimental effect, a finding corroborated by later authors [26, 28]. Subsequent research [27, 28, 56] showed, however, that in a more restricted sample (henceforth the ‘DW’ subsample) covariate matching and other methods recover the experimental effect.
effect. This small sample continues to be used to this day [10].

Figure 4a shows ATE estimates calculated by different methods (columns), using Lalonde’s variables and sample restrictions, as well as the experimental (dashed line). Each dot is an estimate in one Lalonde sample. As in Lalonde’s study, methods struggle to recover the NSW effect. In contrast, SFE estimates are consistently close to the experimental effect—within US$37, well inside the experimental 95% confidence interval.

ATEs neglect that the NSW effect may differ across subpopulations. Figure 4b shows experimental effects for several subpopulations. Far from homogeneous, the program’s effect was particularly large for older, married and relatively educated workers. Figure 4c shows that, unlike other methods, SFE estimates these heterogeneous effects with negligible bias in all subsamples.

Unknown or Unconsidered Causes

Both Lalonde and Dehejia and Whaba restricted the sample population and model variables. Figure 5a shows results in 4 nested subsamples: DW, Lalonde, all males in the NSW and, finally, the full NSW dataset. This figure uses variables Lalonde picked based on his expert judgment. All methods perform well in the DW sample. However, their performance degrades as sample restrictions are relaxed. Figures 5a and 4c suggest that SFE, in contrast, can estimate effects in heterogeneous populations, relieving the need for population selection. Can SFE also help determine which variables to include as causes? Figure 5b show estimates with variables selected by SFE from the 1232 NSW variables (selection procedure details in the Appendix 2). Using this alternative set of 5 variables improves SFE performance. More surprisingly, it also improves other estimators. Using these variables, all observational methods approach the ground-truth (matching methods, in particular), even as sample restrictions are removed. This suggests that the autonomously identified DGP approximates the true DGP more closely than the one derived from expert judgment.

But why did all observational methods perform well in the DW sample? To explain this, Fig. 6a plots individuals’ coordinates on the two variables with the highest ATEs in a NSW-National Survey matched dataset: work-ethics and college ranking. It illustrates how wages increase exponentially with college ranking, while, at the same time, this relation varies with individuals’ work-ethic. The figure shows in red the matched DW subsample. This suggests that the reason why observational methods recover effects in the DW sample with apparent ease is that this subsample consist of individuals with high effect homogeneity.

Which DGP was selected by SFE? Table 1 in the Appendix 2 lists the 20 variables with the largest effects against 20 selected by a traditional model-selection algorithm, a regularized (LASSO) regression. The two lists are very different. Among the variables selected by a LASSO are

- alimony money received
- unemployment in past two years
- money received from training
- money from

Fig. 3 a Factorial subspace \(\{b, c, treat\}\) of \(T(X)\) for the simulations in Fig. 2b, pluses (+) show the (average) positions of treated and non-treated subpopulations, the ATE corresponds to coordinate differences between these subpopulations on the treat axis; b subspace \(\{b, treat\}\), treated individuals in green, nontreated in red, diagonal axis indicates the variables’ combined effect, the ATE is an aggregation of individual effects, distances (solid lines) and their projections, \(\delta_i\) and \(\delta_{ij}\), illustrate individual treatment effects for two example pairs: pair \((ij)\) with both individuals in subpopulation \(b\), with \(+b\) and distinct \(treat\), and \((kl)\) with individuals not in \(b\), with \(-b\) and distinct \(treat\).
social security. Several of these reflect consequences, or just components, of an individual’s income. In contrast, the top variables selected by SFE are work-ethics, race (African-American), NSW treatment, recent school attendance and recent employment. All these variables are arguably connected to causes of income differences such as education, work attitudes and discrimination.

These results suggest that SFE may also shed light on the direction of causation. To explore this further, we revisit the earlier simulations, adding variables that are consequences, not causes, of the outcome variable. We progressively add 10 variables \( u_t, 0 < t \leq 10 \), which are increasingly correlated with \( y \), with expected Pearson correlations \( 1 - t^{-1} \). Figure 2e shows the expanded graphical-model and Fig. 2f ATE estimates. For most methods, even a small numbers of consequences significantly biases estimates. In contrast, SFE estimates remain unbiased.

**Conclusion**

We introduced SFE in this article, a computational tool for nonexperimental causal effect estimation which we compared to several estimators from the Statistics and Machine Learning literature. SFE allows researchers to represent nonexperimental data as incomplete factorial designs. We have shown that, as result, it can recover the ground-truth in synthetic data and in Lalonde’s seminal setting—estimating causal effects with less bias and error than alternatives. We have also shown effect estimates at the individual level and in the entire nationwide NSW program, not relying on ex-ante model and population selection criteria, outperforming estimates that used expert specifications. A more abstract goal was to demonstrate that the troves of data on pairwise treatments and confounders in common nonexperimental data can be very useful when estimating causal effects. Many
Fig. 5  a ATE difference from experimental groundtruth (in dollars) using Lalonde’s expert DGP and progressively fewer ex ante assumptions from previous studies \([10, 24, 27, 28]\), leading to the full NSW dataset (rightmost) with 1231 variables, 1923 treated and 8001 non-treated individuals; estimates become progressively harder for most approaches as sample restrictions are loosened with the exception of SFE; b ATE estimates with an automatically devised DGP (Appendix 2 details the procedure), yielding a DGP with variables: worker’s work-ethic, race (African–American), previous school attendance, previous work, alcoholic drink consumption and location (New York city); all methods perform well throughout all samples using the DGP devised with SFE.

Fig. 6  a Subspace of \(T(x)\) consisting of two variables (ranking of attended college and work-ethics) versus outcome \(y\) (wages), variables selected after running SFE on a matched NSW-National survey (Appendix 2 describes the procedure, dataset has 441 variables, 8001 individuals), survey respondents are blue and individuals in the Dehejia and Wahba (DW) sample are red; b illustration of diminishing exponential returns across subpopulations and the area of concentration of the DW sample (small cube); c \((\delta_i - \bar{\delta})^2\) histogram for treated individuals in DW, Lalonde and the full NSW experimental samples, \(\delta_i\) are before-after individual outcome differences, \(0 < i \leq n\), and \(\bar{\delta}\) their mean in each sample; results illustrate how the DW sample contains a highly homogeneous subpopulation.
fields, from Medicine to the Social Sciences, face new realities where historical data is increasingly accessible and new data is constantly accumulating. The tool could enable new uses for such data in scientific investigation.

**Appendix 1: Objective Function**

In this section, we devise Eq. (5) for binary observed variables. We consider an extension for continuous variables in Appendix 2.

**Treatment Likelihood**

Let’s first define requirements for a pair of individuals (ij) to represent an univariate treatment {a} with certainty, a ∈ Xm. A first requirement relates to the treated variable a itself.4 The requirement is that xi(a) · ¬yi(a) = 1, where ¬ and · are the boolean NOT and AND operators and 1 is a m-sized vector with all +1 values. A second requirement relates to other variables, b ≠ a. The requirement is that these variables are either also treated, xi(b) · ¬yi(b) = 1, or common, xi(b) · yi(b) = 1, between i and j.

With these requirements, we will define individuals’ positions, xi, as random observations of factorial runs. The norm of vectors, |xi| and |xj|, relate to the likelihood of treatment and angles, θij, to observe confounding conditions among pairs of individuals. Their dot-product, ⟨xi, xj⟩ = |xi||xj|cos θij, will reflect both factors and become a key element in the optimization.

Before formulating this relationship in detail, reconsider Eq. (1). Let g(x) = max(0, x) which makes nonpositive coordinates in x zero.5 We can decompose an individual pair into vectors for their difference, xi−xj, and sum, xi+xj. Due to the sign convention, the first contains treated coordinates and the second non-treated coordinates. The dot-product relates the sum of the two vectors geometrically when g(x)=h(x). According to Eq. (1), this corresponds to the assumption that the variance is proportional to the expected effect of non-treated variables, i.e., the expected amount of confounding. The relationship leads to a general least-squares solution (considered in further detail below), where we minimize a residual, yij = |g(xi−xj)|, and a penalty, |h(xi+xj)|. Notice that |g(xi−xj)| is also a distance. Figure 7 sketches the (distance) residual and cost for an example pair. Letting x' = g(x), the Law of Cosines leads to

\[
\begin{align*}
(y_{ij} - |x'_i - x'_j|) + |x'_i + x'_j|,
&= (y_{ij} - |x'_i|^2 - |x'_j|^2 + 2(x'_i, x'_j)) \\
&+ |x'_i|^2 + |x'_j|^2 + 2(x'_i, x'_j),
&= y_{ij} + 4(x'_i, x'_j),
\end{align*}
\]

Let’s then define the probability p§(xi, xj) and its relation to the dot-product in further detail. We defined a variable a ∈ Xm as being under a factorial treatment when a variable is treated and all other variables are either common or also treated. Figure 8a–d (3rd column) depicts these conditions as Venn diagrams for the cases in Fig. 1 (main article). The dot-product in the 2m-dimensional Boolean vector space [57] has the interpretation

\[
\begin{align*}
\langle z_i, z_j \rangle &= \frac{1}{2^m} \sum_{a=1}^{m} z(a)z(a) = E_U[z_i \cdot z_j] \\
&= E_U[z_i \cdot \neg z_j],
\end{align*}
\]

which indicates the expected number of treated variables between vectors, when X variables are uniformly distributed (notated U). Non-uniform distributions and continuous treatments are considered in Appendix 2. Due to the (−1, +1) sign convention, the product in the standardized covariate space X leads to the relation

\[
E_U[x_i \cdot \neg x_j] = \begin{cases} 
-\langle x_i, x_j \rangle, & \text{if } \langle x_i, x_j \rangle \leq 0 \\
0, & \text{otherwise}
\end{cases}
\]

When ⟨x_i, x_j⟩ = −1, the probability of drawing a treated variable when comparing i and j is 1.0—i.e., x_i · ¬x_j = +1, ∀a ∈ X. We also associate the pair with a treatment size, φy, and a possible confounding risk in rela-

---

4 For short, we use x to refer to both Boolean vectors and set variables (i.e., the set of variables with value +1).

5 This function is often called a rectifier and is currently the most popular activation function in deep neural networks.
tion to the remaining $n-2$ individuals, $\phi_{ij}^{bl}$. We consider these in a Bayesian framework next.

**Sample Balance and Optimization**

We now turn to conditions $\phi_{ij}$. We will represent these conditions geometrically, while, at the same time, relating them to the density $p_y(x_i, x_j)$. Equation (1) implies the following likelihood over expected effects:

$$p_y(x_i, x_j) \propto \mathcal{N}(x_i - x_j | 0, \phi_{ij}^{bl}).$$

(11)

To consider also learning from pairs with balanced treatments, we introduce a Gaussian prior $\mathcal{N}(x_i \Theta x_j | 0, \phi_{ij}^{bl})$ for the probability $p_y(x_i, x_j)$, where $\phi_{ij}^{bl}$ is a strictly positive scalar for each pair. If not a factorial treatment, the probability that the pair represents the treatment $x_i \Theta x_j$ depends on the likelihood that non-common factors in the pair (2-sample) are balanced in the remainder of the sample.

Combining the likelihood in Eq. (11) with the Gaussian prior, we obtain

$$\prod_{i,j} \mathcal{N}(y_{ij} | p_y(x_i, x_j)f(x_i \Theta x_j), \phi_{ij}^{bl}) \times \mathcal{N}(x_i \Theta x_j | 0, \phi_{ij}^{bl}),$$

(12)

where $\phi_{ij}^{cx} = h(x_i \Theta x_j)$ is the pair’s variance. This formulates Bayesian priors for conditions $\phi_{ij}$ from Eq. (1) in a way similar to a Tikhonov regularization [58].

Considering a single position $x_i$ and treatment $v$, our goal is to transform $x_i$ such that $|x_i - x_j|^2 = p_y(x_i, x_j)f(x_i \Theta x_j)$ for $0 < j \leq n$. Combining likelihoods in Eqs. (12) and (10), taking logarithms and dropping constants we arrive at the objective

$$\Gamma(x_i) = \min_{x_j} \sum_{j=1}^n (1 + \hat{\phi}_{ij}^{cx})(x_i, x_j) + y_{ij}^2 + \hat{\phi}_{ij}^{bl}(x_i, x_j)^2 + b_i,$$

$$\hat{\phi}_{ij}^{cx} = \frac{|x_i + x_j|}{m},$$

$$\hat{\phi}_{ij}^{bl} = \frac{1}{n} \sum_{k=0}^n (x_{ik} \Theta x_j + x_j),$$

$$y_{ij} = \max(0, y_i - y_j).$$

(13)

We consider the overall objective function first, then pairwise penalties estimates, notated $\hat{\phi}_{ij}$, followed by the intercept $b_i$ and outcome differences $y_{ij}$. If we minimize Eq. (13) with respect to the $m$-sized vector $x_i$, we get a maximum a-posteriori likelihood estimate for individuals’ positions.

---

6 The typical likelihood notation $\mathcal{N}(y \mid \mu, \sigma^2)$, for an observation $y$ with mean $\mu$ and variance $\sigma^2$, is used.
The objective function argument is an individual’s position $x_i$ (rowspace vectors) and not factor positions (column space vectors). More specifically, Eq. (13) leads to an iterative gradient minimization procedure for each individual, $x_i^{t+1} = x_i^t - \eta \Gamma(x_i^t)$, with \(\eta\) as learning rate and $x_i^{t=0} = x_i$. Considering the entire sample population, we iteratively minimize their gradient sum, $\sum [x_i^t - \Gamma(x_i^t)]$. Both Statistics and Machine Learning researchers have considered the problem of minimizing an objective function in the form of a sum of gradients. We use a Stochastic Gradient Descent (SGD) [59] which samples a subset of summand functions at every step and has found wide-spread use in Machine Learning [60, 61]. The scheme allows us to consider billions of observation pairs when estimating effects. We discuss other implementation details in “Implementation”.

The resulting optimization transforms the original space \(X\) into \(T(X)\). It transforms treatment vector differences, until they reflect difference in outcomes that approximate, according to the defined costs, those that would be observed in factorial experiments.

Equation (13) defines \(y_{ij}\) as nonnegative outcome differences. The scalar term \(b_i\) is an individual’s intercept with expected zero mean that is also minimized. Terms $\phi_{ij}^y$ and $\phi_{ij}^x$ reflect difference-of-means balance and treatment size conditions for pairs of individuals \(i\) and \(j\). Pairs with both zero penalties (balanced and univariate treatments) reproduce, according to the previous assumptions, factorial or randomized treatments. In this case, \(x_i, x_j\) is made to reflect \(y_{ij}\).

Calculations will run over thousands of iterations for large observation matrices \(X\). Therefore, it is important to define simple penalties \(\phi_{ij}\). We defined treatments by dividing pairs’ variables into treated and common variable subsets. With the [-1, +1] sign convention, the vector $x_i + x_j$ has non-zero values for non-treated variables. In Eq. (13), the penalty $\phi_{ij}^x$ is therefore a normalized estimate for the number of non-treated variables. Non-treated (i.e., non-zero) coordinates in $x_i + x_j$ can confound outcome effect observations, \(y_{ij}\), Fig. 7c. We did not deem pairs under these conditions as necessarily unsuitable for estimation. Instead, we considered that the pair has coordinates that need to be balanced in individuals \(k\) that do not belong to the pair, \(k \neq i, j\). For an out-of-pair individual \(k\), $\langle x_k, x_i + x_j \rangle$ is the projection of that individual’s vector $x_k$ onto $x_i + x_j$. The penalty $\phi_{ij}^y$ is a sum of such projections from all other \(n - 2\) individuals (signed, due to the same convention). Orthogonal vectors have null projections, and, balanced vector sets have null sums.

### Appendix 2: Supporting Material

#### Continuous Treatments

We can also use the previous method with continuous variables, when it is assumed that there is uncertainty over the intensity of treatments. This can be carried out either by extending $p_i(x_i, x_j)$ directly or by considering a third Bayesian factor for treatment intensity in Eq. (11) (together with treatment balance and size). We consider the former. In Computational Learning Theory, a product distribution [62, 63] is a distribution over \([0, 1]^m\) which generalizes the relationship in Eq. (8) to the non-uniform case. We define a distribution

$$D_{ij} = \prod_{p_i(x_i, x_j) \in [0, 1]} \prod_{a \in X} p_i(+a)p_i(-a)$$

where $p_i(+a)$ is the probability that individual \(i\) has factor \(a\) and $p_i(-a)$ that he doesn’t, $a \in X$. The first therefore indicates certainty of positive treatment status and the second of negative. Any continuous value in between corresponds to individuals with uncertain treatment statuses.

This generalization preserves the relationship in Eq. (10), where $x_i$ becomes the observational random vector with $x_i(a) = p_i(+a) - p_i(-a)$. In this case, the dot-product reflects the expectation over $D_{ij}$ instead of \(U\) [62]. With a single observation per individual, a simple way of obtaining these vectors is unity-base normalizing (i.e., feature scaling) the observation matrix \(X\) and assuming any applicable prior for values in-between, [-1, +1]. This makes maximum and minimum correspond to treated and nontreated statuses, with intermediary treatments having, for example, exponentially decreasing intensities.

#### Implementation

The method can be carried out for all individuals in parallel with matrix operations. For results in this article, we first unity-base normalize \(X\),

$$T^0(X) = 2(X - X_{\min}) \odot (X_{\max} - X_{\min}) - 1.0,$$

where $X_{\min}$ and $X_{\max}$ are \(m \times n\) matrices with per-column maximum and minimum values of \(X\) and $\odot$ is the element-wise (schur) division.
Data: $X, y$
Parameter: $\eta$ (learning rate)
Result: $X_y = T_y(X)$
calculate matrices $X_{\text{min}}$ and $X_{\text{max}}$;
calculate matrix $Y_{ij}$;  \hspace{1cm} // \hspace{0.5cm} y_{ij} = \max(0, y_i - y_j)$
$X_0 := T_0(X)$
\begin{algorithm}
\textbf{while} approximate minimum not obtained \textbf{do}
\begin{align*}
\text{randomly shuffle } m \text{ examples from } X; \\
\text{for } i := 1 \text{ to } m \text{ do}
&\quad \text{calculate matrices } \Phi_{ij}^{ce}, \Phi_{ij}^{bl} \\
&\quad A := \frac{1}{m} X_i \ast X_i \\
&\quad A := (A \otimes \Phi_{ij}^{ce} - Y_{ij})^2 + A^2 \otimes \Phi_{ij}^{bl}; \hspace{1cm} // \hspace{0.5cm} \otimes \hspace{0.5cm} \text{is the schur product} \\
&\quad \nabla X := \frac{1}{m} (A \ast X_i^T)^T \\
&\quad X_i := X_{i-1} - \eta \nabla X \\
\end{align*}
\text{end}
\text{end}
\text{return } X_0
\end{algorithm}

\hspace{1cm} \textbf{Algorithm 1: Space $X_y$ estimation.}

Equation (15) calculates the initial space $T_0(X)$. Subsequent transformations are performed by gradient descent over the sample population. The resulting method is summarized in Algorithm 1. All results in this article use 10,000 iterations and a learning rate of $\eta = 0.025$. An optimized C++ version estimates a space $T_y(X)$ for the NSW dataset in under 5 minutes on a Macbook laptop.

\hspace{1cm} \textbf{NSW Study Details}

The NSW was a 1970s subsidized work program, running in 15 cities across the US for 4 years. It targeted individuals with longstanding employment problems: ex-offenders, former drug addicts, recipients of welfare benefits, and school dropouts. At the time of enrollment, each NSW participant was given a retrospective baseline interview, generally covering the previous two years, followed by up to four follow-up interviews scheduled at nine-month intervals. Survey questions covered demographic and behavior topics such as age, sex, race, marital status, education, number of children, employment history, job search, job training, mobility, housing, household, welfare assistance, military discharge status, drug use and extralegal activities. Most questions were objective and probed for specific information loosely around the previous themes (e.g., ‘what kind of school are you going to? 1 = high school, 2 = vocational, 3 = college, 99 = other’, ‘was heroin used in the last 30 days?’ etc.) Some questions were subjective (e.g., ‘tell me how important each one is to you. knowing the right people, education, luck, hard work,...’)

To assemble control surrogates for the NSW, Lalonde used the \textit{Panel Study of Income Dynamics} (PSID), a household survey, and the Westat’s matched \textit{Current Population Survey-Social Security Administration} file (CPS). He drew 3 subsamples from each the PSID and CPS (6 in total). Control groups had 450, 550, 726, 2666, 2787 and 16289 individuals.\footnote{Thus approximately 100K-200M treatments.} Lalonde ex ante assumptions for the NSW, PSID and CPS regarded mainly participants’ assignment date, gender, retirement status, age and prior wages. DW added further assumptions regarding prior wages for the NSW and used Lalonde’s control groups.

For the ‘missing causes’ study, we first estimated a model $T_y(X)$ where

$$X = \{ \text{all 1231 variables in the NSW dataset} \}. \hspace{1cm} (16)$$

We use the same outcome variable $y$ as Lalonde, post-program annual earnings (in 1982 dollars). While using all NSW variables (i.e., the answer to every survey question), we only restrict them in one way. The restriction doesn’t reduce the participant and variable counts. We ignore any
Table 1 20 top-variables according to $T_s(X)$ and LASSO regression in the complete NSW, ordered by estimated effect or correlation.

| Select by | NSW variables |
|-----------|---------------|
| Equation (6) | Importance getting ahead in life: hard work, African American, treatment, not in school last 6 months, worked < 40 h last 4 weeks, alcohol and hard-liquor consumption, target NSW group (AFDC, ex-offender, ex-addict, youth, other), importance getting ahead in life: education, age > 18, technical eligibility flags, site location (New York City), looking for work last in the last 4 weeks, site location (Philadelphia), no job in the last 6 months, searched for jobs directly from employer, earing less than program minimum (eligibility criteria), other gross eligibility flag (not revealed in the public file), youth group, employed < 9 months last year, program assignment year 1976 |
| LASSO regression, $\beta$ | amount of alimony money, unemployed in 8th pre-program month (timeline), amount of money from training, NSW program, amount of money from social security, amount of SSI dollars, consumed drugs other than marijuana, amount of money from alimony/child support, money from welfare, receive workman’s compensation, amount money from other welfare programs, gender, gender (eligibility flag), any money from workman’s compensation, other gross eligibility flag (not revealed in public file), ever gone to school, how related to the person living with the participant (one of max. 12 persons living with participant), number of children, holds bachelor’s degree, amount of money from AFDC program, how old is relative (one of max. 12 persons living with participant) |

Bolded variables were selected in the employed model selection criteria (see “NSW study details”)

variable values that are negative or ‘99’, taking them as omitted—these values are then mapped to $x(a) = 0$ values according to Eq. (15). These correspond to unknown, not specified or exceptional values in the survey. We assume SFE should be able to handle other types of entry. Most variables are binary and naturally normalized to $[−1, +1]$ according to Eq. (15). Other variables are coded to reflect a spectrum (e.g., ‘even though the 1000 could result in arrest, how likely is that you would take the chance? 1 = very likely, 2 = somewhat likely, 3 = somewhat unlikely, 4 = not likely at all’) and they are accordingly mapped to $[−1, +1]$. Continuous and count variables are similarly linearly normalized to fit the interval (with maxima mapped to +1 and minima to -1). Following Lalonde’s protocol, we ‘annualized’ the

8 Remember that the cosine of angles between pairs of vectors in standardized datasets correspond to their Pearson correlation.
(prior to assignment). The drink variable indicates the participant’s answer to ‘do you ever drink beer, wine, gin or other hard liquor?’ The nyc variable indicates that the participant’s NSW site location was New York City. Another location (Philadelphia) and an assignment year (1976) also appears in the top-20 list.

Similar to Lalonde, we established a correspondence between the NSW and the PSID for these variables. We ignored nyc as the PSID has no public location information. We mapped hardwork to the ‘earning acts’ PSID variable (V2941). It is an aggregate of indicators: ‘[Family] head seldom or never late for work, head rarely or never fails to go to work when not sick, head has extra jobs, head likes to do difficult or challenging things, etc.’ And we mapped drink to PSID’s annual expenditures on alcoholic beverages variable (V2472) divided by income.

The estimates for this alternative model specification across the previous two samples are depicted in Fig. 5 (main text). These results confirm some of the reasons Heckman et al. [17, 56] put forward to explain the poor performance of matching estimators in Lalonde’s NSW subsample: ‘locations in different labor markets’ appear as an effective factor and that the expansion of the ‘limited selected observed variables’ can improve methods’ accuracy. Results suggest that issues like these can, however, be overcome by observational methods by considering missing causes, non-causes, and how to identify them. They also suggest that SFE can be used for both estimate causal effects and help with model specifications for other estimators.

**Analysis: Bias-Variance Tradeoff**

We now motivate the choice to model pairwise individual differences with an alternative analytic argument. Consider an outcome difference predictor \( \hat{y}_{ij} \) for individual \( i \) (i.e., for outcome differences from others). Most effect estimators consider the least-biased estimate for a population. We consider, instead, what would be the least-biased estimate for an individual. Repeated samples of \( x_i, y_j \) can increase the estimator’s accuracy. As assumed in Rubin’s framework [29], these are rarely available, while observations from other individuals are often abundant. We, therefore, consider the error incurred by \( i \) when using an observation from a second individual \( j \). Properties for the following dot-product-based estimator are well known [48](2001, p. 50), as well as the relationship to the Gram-Schmidt procedure (the same results can also be derived through product distributions [62]). What distinguishes the following is the formulation of an estimator for outcome differences, \( y_j \), as opposed to outcomes, \( y_i \).

For \( i \) and an observational pair \((ij)\), the observed outcome difference \( y_{ij} \) can only be due to attributes in \( x_i \) not present in \( x_j \). This leads to a ‘counterfactual’ estimator for effects at the individual-level. According to the estimator, the observed outcome difference between individuals is due to the effect of attributes that only \( i \) has, minus the effect of attributes that only \( j \) has, \( y_{ij} \sim f_{ij}(x_i \Theta x_i) - f_{ij}(x_j \Theta x_j) \).

For an individual \( i \), observations from other individuals lead to the effect predictor

\[
\hat{y}_{ij} = \frac{1}{|x_j - x_i|} \sum_{a \in X} f_{ij}(a) x_i(a) + \epsilon_{ij} = \frac{\langle f_{ij}, x_i - x_j \rangle}{|x_i - x_j|} + \epsilon_{ij},
\]

(19)

where \( f_{ij} \in \mathbb{R}^m \) is an individual vector of effects in \( y \)-scale, \( x \in [-1, +1]^m \), \( E(\epsilon) = 0 \) and \( Var(\epsilon) = 0 \). Due to the sign convention, the estimator sums the effects of variables in \( x_i \) but not in \( x_j \), subtracts the effects of variables in \( x_j \) but not in \( x_i \) and cancels out the effect of variables in both.

The estimator’s squared error loss can be decomposed in 3 components corresponding to a heterogeneity bias, variance and irreducible error \( \epsilon_{ij}^2 \):

\[
Err_{ij} = E[(y_{ij} - \hat{y}_{ij})^2],
\]

\[
= E\left[ (y_{ij} - \frac{\langle f_{ij}, (x_i - x_j) \rangle}{|x_i - x_j|})^2 \right],
\]

\[
= \left[ \text{Heter}_{ij}^2 + \text{Var}_{ij}^2 + \epsilon_{ij}^2 \right].
\]

\[
\text{Heter}_{ij} = E[\hat{y}_{ij}] - y_{ij},
\]

\[
= \left[ \frac{1}{2} \left( \frac{\langle f_{ij}, x_i - x_j \rangle}{|x_i - x_j|} - \frac{\langle f_{ij}, x_i - x_j \rangle}{|x_i - x_j|} \right) \right]^2 - y_{ij},
\]

\[
= \left[ \frac{1}{2} \frac{(f_{ij} - f_{ij})^2}{|x_i - x_j|^2} \right] - y_{ij},
\]

\[
= \frac{1}{4} |f_{ij} - f_{ij}|^2 \cos(\theta_{ij})^2 - y_{ij},
\]

\[
\text{Var}_{ij} = E[\hat{y}_{ij} - E[\hat{y}_{ij}]],
\]

\[
= \frac{\epsilon_{ij}^2}{|x_i - x_j|^2}.
\]

where \( \theta_{ij} \) is the angle between vectors \( f_{ij} \) and \( x_i - x_j \). The second term is a squared heterogeneity bias, the \( y \)-amount by which the estimate differs from the mean using other individuals’ effects. The last term is the variance, the expected squared deviation around the estimated mean in \( y \)-amounts.

According to this, variance and heterogeneity are related to two distances (norms of position differences) between individuals \( i \) and \( j \). These are, in turn, related to spaces \( X \) and \( T_X(X) \). Variance is related to distances in the \([-1, +1]^m \) covariate space, \( |x_i - x_j| \), and heterogeneity to distances over effects, \( |f_{ij} - f_{ij}| \). Larger distances in \( X \) correspond to treatments over more variables, decreasing the estimator’s variance in Eq. (19). Larger distances in \( T_X(X) \)
correspond to estimates between more heterogeneous individuals, increasing the heterogeneity bias. Decreasing this bias increases the estimate’s external validity (for the individual, not the sample population), while decreasing variance increases its internal validity.

Particularly, Eq. (20) suggest that, for a given \( \theta_y \), there are two sources of bias: the difference in variable effects among individuals, \( f_i(a) - f_j(a) \), and the covariate space dimension, \( m \). For the former, an estimate with minimal \( \text{Heter}_{ij} \) must have \( y_{ij} = \frac{1}{2} |f_i - f_j| \). This corresponds to the minimized residual illustrated in Fig. 7c and implemented by Eq. (13). For \( m \), decreasing the space to a dimension \( d < m \) can increase the variance, \( \text{Var}_{ij} \), in Eq. (19) but decrease \( \text{Heter}_{ij} \). This motivated the introduction of the treatment size penalty \( \phi^m_{ij} \). Both the decision to use only variables that differ among pairs and the proposed optimization procedure can therefore be seen as attempts to reduce individual heterogeneity bias, \( \text{Heter}_{ij} \).

Heterogeneity is also decreasing with \( \cos^2 \theta_y \), which, in turn, reflects statistical correlation. This indicates that heterogeneity is maximal for individuals with highly correlated variables (e.g., sharing many attributes) that observe different effects. This motivated the introduction of the treatment balance penalty \( \phi^b_{ij} \), which penalized non-orthogonal pairs, as well as the variable selection criteria in Eq. (17). Together, these considerations suggest a metric space as representation for a sample population, consisting of a set of orthogonal dimensions with correlated covariates (\( \cos^2 \theta_y \approx 1 \)) that are minimally heterogeneous (\( y_{ij} \approx \frac{1}{2} |f_i - f_j| \)).

Starting with a single individual \( i \) and her individual sample \( \{x_i, y_i\} \), we start with maximal internal validity. As we increase the population scope and consider other individuals’ samples \( \{x_i, y_i\} \), we can increase estimates’ external validity. Learning a representation for individual differences, \( \hat{y}_{ij} \), allowed for more accurate (individual) effect estimates while inter-individual effect differences, \( \text{Heter}_{ij} \), were minimized explicitly. This lead to a space that is ’minimal’ but that still reflects observed outcome differences, \( y_{ij} \).

### Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

### References

1. Pearl J. The seven tools of causal inference, with reflections on machine learning. Commun ACM. 2019;62(3):54–60.

2. Athey S. Beyond prediction: Using big data for policy problems. Science 2017;355(6324):483–485. http://science.sciencemag.org/content/355/6324/483.full.pdf. https://doi.org/10.1126/science.aaq3211

3. Imbens GW. Better late than nothing: some comments on deaton (2009) and heckman and urzua (2009). J Econ Lit. 2010;48(2):399–423. https://doi.org/10.1257/jel.48.2.399.

4. Duflo E, Glennerster R, Kremer M. Using randomization in development economics research: a toolkit. In: Schultz TP, Strauss JA (eds.) Handbook of development economics, 2008;4:3895–3962. Elsevier. Chap. 61. https://ideas.repec.org/h/bee/devchp/5-61.html

5. Deaton A. Instruments, randomization, and learning about development. J Econ Lit. 2010;48(2):424–55. https://doi.org/10.1257/jel.48.2.424.

6. Heckman JJ, Smith JA. Assessing the case for social experiments. J Econ Perspect. 1995;9(2):85–110. https://doi.org/10.1257/jep.9.2.85.

7. Xie Y. Population heterogeneity and causal inference. Proc Natl Acad Sci. 2013;110(16):6262. https://doi.org/10.1073/pnas.1303102110.

8. Morgan SL. Counterfactuals and causal inference : methods and principles for social research. New York (2007). Includes bibliographical references (pp. 291–316) and index.; ID: http://id.lib.harvard.edu/aleph/010910135/catalog

9. Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci Rev J Inst Math Stat. 2010;25(1).

10. Colson KE, Rudolph KE, Zimmerman SC, Goin DE, Stuart EA, Van DLM, Ahern J. Optimizing matching and analysis combinations for estimating causal effects. Nat Sci Rep. 2016;6(1). https://doi.org/10.1038/srep23222.

11. MacKay DJC. Information theory, inference, and learning algorithms. UK; New York: Cambridge University Press, Cambridge; 2003.

12. Hall DL, Liggins ME, Linlins J. Handbook of multisensor data fusion: theory and practice. Boca Raton, Florida: CRC Press; 2008.

13. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. https://doi.org/10.1093/biomet/70.1.41.

14. Reichenbach H. The direction of time. Berkeley: University of California Press; 1956.

15. Suter R, Miladinovic D, Schölkopf B, Bauer S. Robustly disenchanting causal mechanisms: Validating deep representations for interventional robustness. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97. pp. 6056–6065. PMLR, 2019. http://proceedings.mlr.press/v97/suter19a.html.

16. Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. Toward causal representation learning. Proc IEEE. 2021;109(5):612–34. https://doi.org/10.1109/JPROC.2021.3058954.

17. Heckman JJ, Ichimura H, Todd P. Matching as an econometric evaluation estimator. Rev Econ Stud. 1998;65(2):261–94. https://doi.org/10.1111/1467-937X.00044.

18. Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. J Am Stat Assoc. 2008;103(484):1334–44.

19. Wang M, Zhao Y, Zhang B. Efficient test and visualization of multi-set intersections. Sci Rep. 2015;5(1):16923–16923. https://doi.org/10.1038/srep16923.

20. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.) Advances in neural information...
processing systems 30, pp. 6446–6456. Curran Associates, Inc., (2017). http://papers.nips.cc/paper/7223-causal-effect-inference-with-deep-latent-variable-models.pdf
21. Wang Y, Blei DM. The blessings of multiple causes. J Am Stat Assoc. 2020;114(528):1574–96. https://doi.org/10.1080/01621459.2019.1686987.
22. Abadie A, Diamond A, Haimmueller J. Comparative politics and the synthetic control method. Am J Poli Sci. 2015;59(2):495–510. https://doi.org/10.1111/ajps.12116.
23. Ribeiro A. An experimental-design perspective on population genetic variation. Proc Natl Acad Sci (PNAS) (Under Review) 2020.
24. Lalonde RJ. Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev. 1986;76(4):604–20.
25. Angrist JD. Mostly harmless econometrics: an empiricist’s companion. Princeton: Princeton University Press; 2009.
26. A., J.S., E., P.T. Does matching overcome lalonde’s critique of nonexperimental estimators? J Econ 2005;125(1):305–353. https://doi.org/10.1016/j.jeconom.2004.04.011
27. Dehejia R, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. J Am Stat Assoc. 1999;94:1053.
28. Zhao Z. Matching estimators and the data from the national sup.
29. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701. https://doi.org/10.1037/h037350.
30. Fisher R. Arrangement of field experiments. Agric J India 1927;22.
31. Dasgupta T, Pillai NS, Rubin DB. Causal inference from 2k factorial designs by using potential outcomes. J R Stat Soc Ser B (Stat Methodol). 2015;77(4):727–53.
32. Pearl J. 3. the foundations of causal inference. Sociol Methodol 2010;40(1):75–149. https://doi.org/10.1177/1467953109345074.
33. King G, Nielsen R. Why propensity scores should not be used for matching. Polit Anal. 2019;27(4):435–54. https://doi.org/10.1017/pan.2019.11.
34. Imai K, King G, Stuart E. Misunderstandings between experiment.
35. Pearl J. Causality : Models, Reasoning, and Inference, Cambridge, U.K.; New York (2000). Includes bibliographical references (p. 359-373) and indexes.; ID: http://id.lib.harvard.edu/ealph/00837258/catalog.
36. Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol. 2013;66(8):84–901. https://doi.org/10.1016/j.jclinepi.2013.01.013.
37. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34(28):3661–79. https://doi.org/10.1002/sim.6607.
38. Belisier SV, Martens EP, Pestman WR, Groenwold RHH, de Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf. 2011;20(11):1115–29. https://doi.org/10.1002/pds.2188.
39. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083–107. https://doi.org/10.1002/sim.3697.
40. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods. 2004;9(4):403–25. https://doi.org/10.1037/1082-989X.9.4.403.
41. Diamond A, Sekhon JS. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Rev Econ Stat. 2012;95(3):932–45.
42. Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008;95(2):481–8. https://doi.org/10.1093/biomet/asm004.
43. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeewess S. Metrics for covariate balance in cohort studies of causal effects. Stat Med. 2014;33(10):1685–99. https://doi.org/10.1002/sim.6058.
44. Iacus SM, King G, Porro G. Causal inference without balance checking: coarsened exact matching. Polit Anal. 2012;20(1):1–24. https://doi.org/10.1017/panp013.
45. Huling JD, Mak S. Energy balancing of covariate distributions 2020.
46. Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. Ann Stat. 2013;41(5):2263–91. https://doi.org/10.1214/13-AOS1140.
47. Tarantola A. Inverse problem theory and methods for model parameter estimation. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2005.
48. Hastie T. The elements of statistical learning : data mining, inference, and prediction. New York, NY: Springer Series in Statistics. Springer; 2001.
49. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci 2016;113(27):7353–7360. http://www.pnas.org/content/113/27/7353.full.pdf. https://doi.org/10.1073/pnas.1510489113.
50. van Der Laan J, M, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007:6:25.
51. Chatton A, Le Borgne F, Leyrat C, Gillaizeau F, Rousseau C, Barbin L, Laplaud D, Leger M, Giraudneau B, Foucher Y. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. Nat Sci Rep. 2020;10(1):9219–9219. https://doi.org/10.1038/s41598-020-65917-x.
52. Stanley XP, Colleen RM, Marsha RP, Susan SM, Christopher BP, David SP. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. Value Health 2010;13(2):273–277. https://doi.org/10.1111/j.1524-4733.2009.00671.x.
53. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass.) 2000;11(5):550–560. https://doi.org/10.1097/00001648-200009000-00011.
54. Aalen OO, Farewell VT, de Angelis D, Day NE, Nöel Gill O. A markov model for hiv disease progression including the effect of hiv diagnosis and treatment: application to aids prediction in england and wales. Stat Med. 1997;16(19):2191–210. https://doi.org/10.1002/ (SICI) 1097-0258(19971 015) 16: 19< 2191:: AID- SIM64 6>3.0.CO.
55. Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: A tutorial. Stat Med. 2018;37(16):2530–46. https://doi.org/10.1002/ (SICI) 1097-0258(20180715)37:16<2530:: AID- SIM64 5>3.0.CO.
56. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat. 2002;84(1):151–61. https://doi.org/10.1162/003465302317331982.
57. Furst M, Jackson J, Smith S. Improved learning of ac0 functions. In: Annual Workshop on Computational Learning Theory: Proceedings of the Fourth Annual Workshop on Computational Learning Theory.
58. Figueiredo M. Adaptive sparseness for supervised learning. IEEE Trans Pattern Anal Mach Intell. 2003;25(9):1150–9.

59. Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. Ann Math Stat. 1952;23(3):462–6. https://doi.org/10.1214/aoms/1177729392.

60. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Belleville MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529. https://doi.org/10.1038/nature14236.

61. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.

62. Servedio RA. On learning monotone dnf under product distributions. Inf Comput. 2004;193(1):57–74.

63. Bshouty N, Tamon C. On the fourier spectrum of monotone functions. J ACM (JACM). 1996;43(4):747–70.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.