Journal of Machine Learning Research 27 (2026) 1-57 Submitted 12/24; Published 1/26

Boosted Control Functions: Distribution Generalization and Invariance in Confounded Models

Nicola Gnecco* N.GNECCO@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, United Kingdom

Jonas Peters* JONAS.PETERS@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zurich
Zurich, Switzerland

Sebastian Engelke SEBASTIAN.ENGELKE@UNIGE.CH
Research Center for Statistics
University of Geneva
Geneva, Switzerland

Niklas Pfister NIKLAS.PFISTER@GMAIL.COM
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Editor: Jin Tian

Abstract

Modern machine learning methods and the availability of large-scale data have significantly advanced our ability to predict target quantities from large sets of covariates. However, these methods often struggle under distributional shifts, particularly in the presence of hidden confounding. While the impact of hidden confounding is well-studied in causal effect estimation, e.g., instrumental variables, its implications for prediction tasks under shifting distributions remain underexplored. This work addresses this gap by introducing a strong notion of invariance that, unlike existing weaker notions, allows for distribution generalization even in the presence of nonlinear, non-identifiable structural functions. Central to this framework is the Boosted Control Function (BCF), a novel, identifiable target of inference that satisfies the proposed strong invariance notion and is provably worst-case optimal under distributional shifts. The theoretical foundation of our work lies in Simultaneous Equation Models for Distribution Generalization (SIMDGs), which bridge machine learning with econometrics by describing data-generating processes under distributional shifts. To put these insights into practice, we propose the ControlTwicing algorithm to estimate the BCF using nonparametric machine-learning techniques and study its generalization performance on synthetic and real-world datasets compared to robust and empirical risk minimization approaches.

Keywords: distribution generalization, invariance, causality, under-identification, simultaneous equation models.

*. Part of this work was conducted while NG and JP were at the University of Copenhagen.

1. Introduction

Prediction and forecasting methods are fundamental in describing how a target quantity behaves in the future or under different settings. With recent advances in machine learning and the availability of large-scale data, prediction has become reliable in many applications, such as macroeconomic forecasting (Stock and Watson, 2006), and predicting effects of policies (Hill, 2011; Kleinberg et al., 2015; Athey and Imbens, 2016; Künzel et al., 2019). However, it is also well-known that focusing solely on prediction when reasoning about a system under changing conditions can be misleading, especially in the presence of unobserved confounding. While there are several well-established methods for dealing with unobserved confounding in causal effect estimation, less research has focused on comparable approaches for prediction tasks. In this work, we consider the problem of predicting a response when the training and testing distributions differ in the presence of unobserved confounding.

In the literature on causal effect estimation from observational data, the main approaches to deal with unobserved confounding are instrumental variables (Angrist et al., 1996; Ng and Pinkse, 1995; Newey et al., 1999), regression discontinuity (Angrist and Lavy, 1999), and difference-in-differences (Angrist and Krueger, 1991). Instrumental variable approaches, for example, use specific types of exogenous variables (called instruments) that can be seen as natural experiments to identify and estimate causal effects. Most of the existing methods require the causal effect to be identifiable. In simple scenarios such as linear or binary models, these identifiability conditions are well understood (Angrist et al., 1996; Amemiya, 1985). Identification strategies in more complex scenarios impose specific structures on the causal effect, e.g., sparsity (Pfister and Peters, 2022), independence between the instruments and the residuals (Dunker et al., 2014; Dunker, 2021; Saengkyongam et al., 2022; Loh, 2023), or the milder mean-independence condition (Newey and Powell, 2003). However, when there are many endogenous covariates, finding a sufficiently large number of valid instruments to identify and estimate the causal effect is often not feasible – even when the effect is sparse.

In this work, we show that even when the causal function is not identifiable, it is possible to exploit the heterogeneity induced by a set of exogenous variables to learn a prediction function that yields valid predictions under a class of distributional shifts. Formally, we consider the task of predicting a real-valued outcome $Y \in \mathbb{R}$ from a large set of (possibly) endogenous covariates $X \in \mathbb{R}^p$ when the training and testing data follow a different distribution. Distribution generalization has received much attention particularly in the machine learning community and is usually tackled from a worst-case point of view, which is particularly relevant in high-stakes applications. Given a set of potential distributions $\mathcal{P}$ of a random vector that contains the components (X, Y), we aim to find a predictive function f^* minimizing the worst-case risk over the set of distributions $\mathcal{P}$, that is

$$\sup_{P \in \mathcal{P}} \mathrm{E}_P[(Y - f^*(X))^2] = \inf_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathrm{E}_P[(Y - f(X))^2],$$

where $\mathcal{F}$ is a class of measurable functions.

The task of distribution generalization is intractable without characterizing the set of potential distributions $\mathcal{P}$. For example, one can model $\mathcal{P}$ by changing the marginal $P_{\mathrm{tr}}^X \neq P_{\mathrm{te}}^X$, known as covariate shift (Shimodaira, 2000; Sugiyama et al., 2007), the conditional $P_{\mathrm{tr}}^{Y|X} \neq P_{\mathrm{te}}^{Y|X}$, known as concept shift (Quiñonero-Candela et al., 2009; Gama et al., 2014), or a combination of both, (Arjovsky et al., 2020; Krueger et al., 2021). Alternatively, one can model $\mathcal{P}$ as a set of distributions that are within a neighborhood of the training distribution $P_{\text{tr}}^{X,Y}$ with respect to the Wasserstein distance (Sinha et al., 2018) or f-divergence (Bagnell, 2005; Hu et al., 2018). Furthermore, one can model $\mathcal{P}$ as the convex hull of different distributions $P_{\text{tr},1}^{X,Y}, \dots, P_{\text{tr},m}^{X,Y}$ that are observed at training time (Meinshausen and Bühlmann, 2015; Sagawa et al., 2020).

Here, we model the heterogeneity in the distributions $\mathcal{P}$ via a vector of exogenous variables $Z \in \mathbb{R}^r$ which are only observed at training time and induce shifts in the conditional mean of the response given the covariates. Our framework builds upon the literature of invariance-based methods for distribution generalization. Invariance (also known as autonomy or modularity (Haavelmo, 1944; Aldrich, 1989; Pearl, 2009)) is one of the key ideas in causality and states that predictions made with causal models are unaffected by interventions on the covariates. Invariance has been used for causal structure search (e.g., Peters et al., 2016) but is also implicitly used in causal effect estimation methods such as instrumental variables. More recently, invariance has also been employed as a concept to achieve distribution generalization via the invariant most predictive (IMP) function (Magliacane et al., 2018; Rojas-Carulla et al., 2018; Arjovsky et al., 2020; Bühlmann, 2020; Christiansen et al., 2022; Krueger et al., 2021; Jakobsen and Peters, 2022; Saengkyongam et al., 2022).

In this work, we propose a strong notion of invariance that, unlike weaker notions (e.g., Arjovsky et al., 2020; Bühlmann, 2020; Christiansen et al., 2022; Krueger et al., 2021; Jakobsen and Peters, 2022), ensures distribution generalization when the structural function is nonlinear and possibly not identifiable. Building upon the control function approach (Ng and Pinkse, 1995; Newey et al., 1999), we propose the boosted control function (BCF), a novel, identifiable target of inference that satisfies the proposed strong invariance notion. The BCF aligns with the invariant most predictive (IMP) function and is provably worst-case optimal under distributional shifts induced by the exogenous variables. Moreover, we provide necessary and sufficient conditions for identifying the BCF under continuous and discrete exogenous variables. Our theoretical results rest upon the simultaneous equation models for distribution generalization (SIMDGs), a novel framework to describe distributional shifts induced by the exogenous variables. SIMDGs establish a connection between distribution generalization in machine learning and simultaneous equation models (Haavelmo, 1944; Amemiya, 1985) in econometrics.

We further develop the ControlTwicing algorithm which estimates the BCF using the 'twicing' idea from Tukey et al. (1977). Our method can be applied using various machine learning methods such as ridge regression, lasso, random forests, boosted trees and neural networks. In a set of numerical experiments, we study the generalizing properties of the BCF estimator and observe its advantage over standard prediction methods when the training and testing distributions differ. Moreover, on the California housing dataset (Pace and Barry, 1997), we show that the BCF is robust to previously unseen distributional shifts and is competitive with existing robust methods.

2. SIMs for Distribution Generalization

Given a vector of covariates $X \in \mathbb{R}^p$, our goal is to identify a function $f : \mathbb{R}^p \to \mathbb{R}$ that predicts a response $Y \in \mathbb{R}$ from X, under distributional shifts induced by exogenous variables $Z \in \mathbb{R}^r$. Let $\mathcal{P}$ denote a collection of distributions of a random vector that contains the components (X, Y, Z) and let $P_{\text{tr}} \in \mathcal{P}$ denote the training distribution, from which we observe $(X, Y, Z) \sim P_{\text{tr}}^{X,Y,Z}$. To evaluate the performance of a candidate function $f \in \mathcal{F}$, where $\mathcal{F}$ is a class of measurable functions, we define the risk function $\mathcal{R} : \mathcal{P} \times \mathcal{F} \to [0, \infty]$ for all $P \in \mathcal{P}$ and $f \in \mathcal{F}$ by

$$\mathcal{R}(P, f) := \mathrm{E}_P[(Y - f(X))^2]. \tag{1}$$

We assume that we observe independent copies of $(X, Y, Z) \sim P_{\text{tr}}^{X,Y,Z}$ and want to predict Y from X under a potentially different but unknown testing distribution $P_{\text{te}}^{X,Y}$, where $P_{\text{te}} \in \mathcal{P}$. In this work, our estimand (or target) is a function $\tilde{f}$ that minimizes the risk (1) under the worst-case distribution in $\mathcal{P}$.

Definition 1 (Distribution generalization) *Denote by $\mathcal{F}$ a subset of measurable functions $f : \mathbb{R}^p \to \mathbb{R}$ and by $\mathcal{P}$ a class of distributions of the random vector (X, Y, Z). Let $\tilde{f} : \mathbb{R}^p \to \mathbb{R}$ be a function satisfying*

$$\sup_{P \in \mathcal{P}} \mathcal{R}(P, \tilde{f}) = \inf_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathcal{R}(P, f) < \infty. \tag{2}$$

Then, we say that distribution generalization is achievable *and $\tilde{f}$ is a* generalizing function.

Without any constraints on the class of distributions $\mathcal{P}$, distribution generalization is generally not achievable. One way to constrain the class of distributions $\mathcal{P}$ is via causal models as done e.g. by Christiansen et al. (2022) who consider structural causal models with independent error terms. Here, we consider semi-parametric simultaneous equation models where

$$\begin{aligned} Y &= f_0(X) + U, \\ X &= M_0 Z + V, \end{aligned} \tag{3a}$$

$$(U, V) \perp\!\!\!\perp Z, \tag{3b}$$

$$\mathrm{E}[(U, V)] = 0, \ \mathrm{E}\big[\, \|(U, V)\|_2^2 \,\big] < \infty, \tag{3c}$$

(see Figure 4 in Appendix A.3 for a visualization) and assume that Z and X are centered to mean zero under P_{tr} (see also Setting 1). The first equation in (3a) is in structural form and describes the mechanism between the dependent variable Y and the endogenous covariates X under distributional shifts of the exogenous variable Z. The function $f_0 : \mathbb{R}^p \to \mathbb{R}$ captures a possibly nonlinear relationship between X and Y. The second equation in (3a) is in reduced form, that is, the dependent variable X depends only on the exogenous variables Z and V. The matrix $M_0 \in \mathbb{R}^{p \times r}$ describes the linear dependence between X and Z and is not required to be full-rank. The hidden variables U and V are not required to be independent, allowing for unobserved confounding between X and Y. This confounding, together with shifts in Z, can induce both covariate shifts ($P_{\text{tr}}^X \neq P_{\text{te}}^X$) and concept shifts ($P_{\text{tr}}^{Y|X} \neq P_{\text{te}}^{Y|X}$) within model (3a).

Remark 2 (Linear dependence between X and Z) *For the sake of simplicity, in* (3a) *we assume a linear functional dependence between Z and X. In principle, however, it is possible to consider nonlinear maps of the form $z \mapsto G(z) := \Theta_0 \phi(z)$, for some real-valued matrix $\Theta_0 \in \mathbb{R}^{p \times q}$ with a (known) basis $\phi : \mathbb{R}^r \to \mathbb{R}^q$. Without parametric restrictions on the function ϕ, however, generalization to shifts in Z is impossible (Christiansen et al., 2022).*

Remark 3 (Z categorical) *If Z is an exogenous categorical variable (which often occurs in practical applications), our model may still apply after encoding Z appropriately. Suppose that Z takes values in $\{a_0, \ldots, a_r\}$ with probabilities $\pi_0, \ldots, \pi_r$. We then encode for all $j \in \{1, \ldots, r\}$ the category a_j as e_j, where $e_j \in \mathbb{R}^r$ is the j-th standard basis vector and a_0 as $-[\frac{\pi_1}{\pi_0}, \ldots, \frac{\pi_r}{\pi_0}]^\top$. The newly encoded variable $Z \in \mathbb{R}^r$ satisfies* $\mathrm{E}[Z] = 0$ *by construction. Moreover, under the assumptions that (a)* $\mathrm{E}[X] = 0$ *and that (b) the conditional distributions of X given the categories are shifted versions of each other, we can express the relation between X and Z as in* (3a)*. More concretely, define $\mu_j \coloneqq \mathrm{E}[X \mid Z = j]$ for all $j \in \{0, \ldots, r\}$ and the matrix $M_0 \coloneqq [\mu_1, \ldots, \mu_r] \in \mathbb{R}^{p \times r}$. Then, (b) formally requires for all $a, b \in supp(P^Z_{\mathrm{tr}})$*[1] *that $X - \mu_a$ conditioned on $Z = a$ has the same distribution as $X - \mu_b$ conditioned on $Z = b$. Hence, together with (a) using $V = X - M_0 Z$ implies that* $\mathrm{E}[V] = 0$ *and $X = M_0 Z + V$.*

When the exogenous variable is categorical, there exist different representations of Z to express the predictor vector as in (3a). The different representations of Z, however, are all equivalent with respect to the generalization guarantees (see Section 4.2) since they describe the same linear span of the conditional means $\mathrm{E}[X \mid Z = a_j]$, where $j \in \{0, \ldots, r\}$.

The following remark shows that under certain conditions, the proposed model (3a)–(3c) allows the exogenous variable Z to directly affect the response Y.

Remark 4 (Allowing Z to affect Y) *In this work, instead of identifying the structural function f_0, we aim at predicting Y from X under distributional shifts on Z. In contrast to widely-used assumptions on IVs, our model* (3a)–(3c) *allows Z to have a direct effect on Y, provided that Z can be expressed as a function of X and V. More precisely, we allow for $Y = g_0(X) + \beta_0^\top Z + U$, for some $g_0 : \mathbb{R}^p \to \mathbb{R}$ and $\beta_0 \in \mathrm{im}(M_0^\top)$. The assumption that $\beta_0 \in \mathrm{im}(M_0^\top)$ has been termed* projectability condition *(Rothenhäusler et al., 2021) and ensures that $\beta_0^\top Z$ can be expressed as a linear combination of the covariates X and hidden variables V. This condition is automatically satisfied when* $\mathrm{rank}(M_0) = r$ *because then* $\mathrm{im}(M_0^\top) = \mathbb{R}^r$*. When* $\mathrm{rank}(M_0) < r$*, however, the projectability condition constrains the allowed models. Given the projectability condition, the exogenous variable satisfies $Z = M_0^\dagger(X - V)$ via equations* (3a)*, where $M_0^\dagger \in \mathbb{R}^{r \times p}$ is the Moore–Penrose inverse of $M_0 \in \mathbb{R}^{p \times r}$. Therefore, we can rewrite the structural equation $Y = g_0(X) + \beta_0^T Z + U$ as*

$$Y = \Big(g_0(X) + \beta_0^\top M_0^\dagger X\Big) + \Big(U - \beta_0^\top M_0^\dagger V\Big) =: f_0(X) + \tilde{U}.$$

By construction, the two structural equations for Y induce the same distribution $P^{X,Y,Z}$ for any distributional shift on Z, and, therefore, they are equivalent for our purpose to solve (2)*.*

To construct the set $\mathcal{P}$ of potential distributions, we introduce the following simultaneous equation model (SIM) for distribution generalization (SIMDG).

Definition 5 (SIM for Distribution Generalization (SIMDG)) *Let $\mathcal{Q}_0$ be a set of distributions over $\mathbb{R}^r$ and Λ_0 be a distribution over $\mathbb{R}^{1+p}$ such that if $(U, V) \sim \Lambda_0$ then* (3c)

1. Let $X : \Omega \to \mathcal{X}$ be a random variable where Ω denotes the sample space and $\mathcal{X}$ is a Euclidean space. Let P^X denote the distribution of X. The support of X, denoted by $\mathrm{supp}(P^X)$, is the set of all $x \in \mathcal{X}$ such that every open neighborhood $N_x \subseteq \mathcal{X}$ of x has positive probability.

holds. Let further $f_0 : \mathbb{R}^p \to \mathbb{R}$ *be a measurable function and* $M_0 \in \mathbb{R}^{p\times r}$ *a matrix. We call the tuple* $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$ *a* SIMDG*. For all* $Q \in \mathcal{Q}_0$ *the model* (f_0, M_0, Λ_0, Q) *induces a unique distribution over* (U, V, X, Y, Z) *via* $Z \sim Q$*,* $(U, V) \sim \Lambda_0$*,* (3b)*, and the simultaneous equations* (3a)*. We define the set of induced distributions by*

$$\mathcal{P}_0 := \left\{ P \text{ distr. over } \mathbb{R}^{2p+r+2} \mid \exists Q \in \mathcal{Q}_0 \text{ s.t. } (f_0, M_0, \Lambda_0, Q) \text{ induces } P \text{ via (3a)} \right\}. \quad (4)$$

A SIMDG $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$ defines a collection of distributions $\mathcal{P}_0$. In particular, the training and testing distribution P_{tr} and P_{te} are (potentially different) distributions induced by (potentially different) $Q_{\mathrm{tr}}, Q_{\mathrm{te}} \in \mathcal{Q}_0$; the changes in $Q \in \mathcal{Q}_0$, in turn, induce mean shifts of X in the directions of the columns of the matrix M_0. In Appendix A.1 we provide further technical details on how a SIMDG generates the class of distributions $\mathcal{P}_0$.

Even by constraining the set of distributions by a SIMDG, distribution generalization is achievable only if we further impose specific assumptions on either $\mathcal{P}_0$ or the function class $\mathcal{F}$ (Christiansen et al., 2022). Assumptions on $\mathcal{P}_0$ usually require that the training distribution dominates all distributions, while assumptions on the function class $\mathcal{F}$ usually ensure that the target function extrapolates outside the training support in a known way.

Assumption 1 (Set of distributions $\mathcal{P}_0$) *For all* $P \in \mathcal{P}_0$ *it holds that* $P \ll P_{\mathrm{tr}}$*.*

Assumption 2 (Function class $\mathcal{F}$) *The function class* $\mathcal{F}$ *is such that for all* $f, g \in \mathcal{F}$ *and all* $P \in \mathcal{P}_0$ *it holds*

$$f(X) = g(X),\ P_{\mathrm{tr}}\text{-a.s.} \implies f(X) = g(X),\ P\text{-a.s.}$$

Note that Assumption 1 implies Assumption 2. Throughout the paper, we will assume that either Assumption 1 or Assumption 2 are satisfied. Furthermore, we will use the following data-generating process.

Setting 1 (Data-generating process) *Fix a SIMDG* $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$ *where* $\mathcal{Q}_0$ *induces a set of distributions* $\mathcal{P}_0$*. Moreover, assume that* $\sup_{P\in\mathcal{P}_0} \mathrm{E}_P[f_0(X)]^2 < \infty$*. Let* $Q_{\mathrm{tr}} \in \mathcal{Q}_0$ *such that* $Z \sim Q_{\mathrm{tr}}$ *satisfies* $\mathrm{E}_{Q_{\mathrm{tr}}}[Z] = 0$ *(Z can but does not have to be categorical, see Remark 3) and* $\mathrm{E}_{Q_{\mathrm{tr}}}[ZZ^\top] \succ 0$*, and let* $Q_{\mathrm{te}} \in \mathcal{Q}_0$ *be an arbitrary distribution. Denote the* training distribution *by* $P_{\mathrm{tr}} \in \mathcal{P}_0$ *and the* testing distribution *by* $P_{\mathrm{te}} \in \mathcal{P}_0$*; both are distributions over* (U, V, X, Y, Z) *induced by* $(f_0, M_0, \Lambda_0, Q_{\mathrm{tr}})$ *and* $(f_0, M_0, \Lambda_0, Q_{\mathrm{te}})$*, respectively. We now consider the following two phases.*

(1) (Training) *Observe an i.i.d. sample* $(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)$ *of size* n *with distribution* $P_{\mathrm{tr}}^{X,Y,Z}$*.*

(2) (Testing) *Given an independent draw* $X \sim P_{\mathrm{te}}^X$ *predict the response* Y*.*

In practical applications, a categorical exogenous variable Z can represent the different environments from which the data was collected, such as hospitals (Bandi et al., 2018). In such cases, using Z as an additional covariate does not enhance the prediction accuracy for environments not observed during training time. A continuous exogenous variable Z can represent geospatial information, such as latitude and longitude. As we discuss in Section 5.3 based on the example of predicting house prices, using the geospatial information directly as covariates can lead to inaccurate predictions due to extrapolating to areas not included in the training data. However, using latitude and longitude as exogenous variables that model distribution shifts in the other covariates can improve predictive accuracy in areas that are not covered by the training data.

SIMDGs describe a set of models that are closely related to the SIMs from the instrumental variable literature (e.g., Newey et al., 1999) and the structural causal models (SCMs) from the causality literature in statistics (e.g., Pearl, 2009), as we discuss in Appendices A.2 and A.3.

3. Invariant Most Predictive Functions

The concept of invariant most predictive (IMP) function has been recently proposed as a guiding principle to identify a generalizing function (Magliacane et al., 2018; Rojas-Carulla et al., 2018; Arjovsky et al., 2020; Bühlmann, 2020; Christiansen et al., 2022; Krueger et al., 2021; Jakobsen and Peters, 2022; Saengkyongam et al., 2022). To provide a motivating example, consider Setting 1 with SIMDG $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$, where f_0 belongs to the class of linear functions $\mathcal{F}$ and $\mathcal{Q}_0$ consists of arbitrary distributions on $\mathbb{R}^r$. For a fixed $z \in \mathbb{R}^r$, define the point mass distribution $Q_z := \delta_z \in \mathcal{Q}_0$, and denote by P_z the distribution induced by $(f_0, M_0, \Lambda_0, Q_z)$. Then, using the simultaneous equations (3a), for all functions $g \in \mathcal{F}$ the risk of g under the perturbed distribution P_z can be expressed as

$$\begin{aligned}\mathcal{R}(P_z, g) = \mathrm{E}_{P_z}\left[\big(Y - g(X)\big)^2\right] &= \mathrm{E}_{P_z}[U^2] + \delta^\top \mathrm{E}_{P_z}[XX^\top]\delta - 2\delta^\top \mathrm{E}_{P_z}[XU] \\ &= \mathrm{E}_{P_z}[(U - \delta^\top V)^2] + \delta^\top M_0 z z^\top M_0^\top \delta,\end{aligned} \tag{5}$$

where $\delta^\top x := f_0(x) - g(x)$. When $M_0^\top \delta \neq 0$, the risk $\mathcal{R}(P_z, g)$ can be made arbitrarily large by increasing the magnitude of $z \in \mathbb{R}^r$, and therefore g is then not a generalizing function. The reason is that the distribution of residuals $Y - g(X) = U - \delta^\top X = U - \delta^\top V - \delta^\top M_0 Z$ is not invariant to changes in the marginal distribution of Z. The IMP function identifies a function minimizing (5) among those yielding an invariant distribution of residuals. In the above example with a linear function class, it is clear that the invariance of the residual distribution is a necessary condition to achieve distribution generalization. However, for more general function classes the relation between invariance and distribution generalization is more intricate. Even more so, when the function class is not constrained to the linear setting, existing IMP-based approaches can fail at identifying a function that is invariant in the sense of Definition 6 (see Example 1), and therefore, may not yield a generalizing function (see Example 3).

The goal of this section is to investigate the relation between invariance and distribution generalization in the more general setting when $\mathcal{F}$ is not constrained to be a parametric function class. To do so, we first formally define the notion of invariant function.

Definition 6 (Invariant function) *Assume Setting 1 and for all $P \in \mathcal{P}_0$ and all $f \in \mathcal{F}$ denote by $P^{Y-f(X)}$ the distribution of the random variable $Y - f(X)$ under the probability measure P. We say $f \in \mathcal{F}$ is* invariant *w.r.t. $\mathcal{P}_0$ if*

$$P_{\mathrm{tr}}^{Y-f(X)} = P^{Y-f(X)}, \ \textit{for all } P \in \mathcal{P}_0.$$

Furthermore, we define the set of invariant functions

$$\mathcal{I}_0 := \{f \in \mathcal{F} \mid f \text{ is invariant w.r.t. } \mathcal{P}_0\}.$$

Given a SIMDG $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$ and the set of induced distributions $\mathcal{P}_0$, the structural function f_0 is always invariant since the distribution of its residuals $Y - f_0(X) = U$ does not depend on $P \in \mathcal{P}_0$. Additionally, depending on the relation between the exogenous variables and the covariates, there may exist further invariant functions other than f_0. More precisely, the size of the set of invariant functions $\mathcal{I}_0$ depends on the rank and order conditions of identifiability (Amemiya, 1985). For example, under Setting 1, if $\text{rank}(M_0) = p$, then the set of invariant functions is a singleton $\mathcal{I}_0 = \{f_0\}$. If, instead $q := \text{rank}(M_0) < p$, then the set of invariant functions $\mathcal{I}_0$ can contain infinitely many elements, as we now argue. Let $\ker(M_0^\top)$ denote the null space of $M_0^\top$ with dimension $p - q > 0$. Define the matrix $R := (r_1, \ldots, r_{p-q}) \in \mathbb{R}^{p \times (p-q)}$, where $r_1, \ldots, r_{p-q} \in \mathbb{R}^p$ is a basis of $\ker(M_0^\top)$, let $h : \mathbb{R}^{p-q} \to \mathbb{R}$ be an arbitrary function satisfying $h \circ R^\top \in \mathcal{F}$, and define $f^* := f_0 + h \circ R^\top \in \mathcal{F}$ (which holds true if $\mathcal{F}$ is closed under addition). By using the reduced form equation for X in (3a), we have that $R^\top X = R^\top M_0 Z + R^\top V = R^\top V$, and so the distribution of $Y - f^*(X) = U - h(R^\top V)$ remains the same for all $P \in \mathcal{P}_0$. Since there may be infinitely many such functions h, the set $\mathcal{I}_0$ can contain infinitely many functions in addition to $f_0 \in \mathcal{I}_0$. This motivates the following definition.

Definition 7 (Invariant most predictive (IMP) function) *Assume Setting 1 and denote by $\mathcal{I}_0$ the set of invariant functions (see Definition 6). We call $f_{\mathcal{I}_0} \in \mathcal{F}$ an* invariant most predictive (IMP) *function if*

$$\mathcal{R}(P_{\text{tr}}, f_{\mathcal{I}_0}) = \inf_{f \in \mathcal{I}_0} \mathcal{R}(P_{\text{tr}}, f). \tag{6}$$

The notion of invariant most predictive (IMP) function provides a constructive method to tackle the distribution generalization problem. A potential approach to identify an IMP function $f_{\mathcal{I}_0}$ is to (i) identify the set of invariant functions $\mathcal{I}_0$, and (ii) solve the optimization problem in (6) constrained to the set $\mathcal{I}_0$. For instance, if $\mathcal{F}$ is the class of linear functions, then one can show that any invariant function f is identified by the moment condition $\text{E}_{P_{\text{tr}}}[(Y - f(X))Z] = 0$ and that $\mathcal{I}_0 = \{f \in \mathcal{F} \mid \text{E}_{P_{\text{tr}}}[(Y - f(X))Z] = 0\}$. Furthermore, one can show that any function f is invariant if and only if $f(x) = f_0(x) + \delta^\top x$, where $\delta \in \ker(M_0^\top)$ (Jakobsen and Peters, 2022).

Though in the linear setting the identification and characterization of the set of invariant functions $\mathcal{I}_0$ is straightforward, this is not true in a more general setting. In particular, when $\mathcal{F}$ is a flexible function class, invariance-based methods for distribution generalization consider weaker notions of invariance compared to Definition 6 and, in some cases, do not identify a generalizing function.

3.1 Relaxing Invariance and its Implications

The IMP function is defined as the solution to the constrained optimization problem (6) over the set of invariant functions $\mathcal{I}_0$. Existing invariance-based methods usually tackle this problem in two steps. First, they solve a relaxation of (6) by optimizing over a larger set

$\mathcal{H} \supseteq \mathcal{I}_0$; common examples of such sets are $\mathcal{I}_0 \subseteq \mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3$ defined as

$$\mathcal{H}_1 := \left\{ f \in \mathcal{F} \mid Y - f(X) \perp\!\!\!\perp Z \text{ under } P_{\text{tr}} \right\},$$
$$\mathcal{H}_2 := \left\{ f \in \mathcal{F} \mid \mathrm{E}_{P_{\text{tr}}}[Y - f(X) \mid Z] = 0 \right\},$$
$$\mathcal{H}_3 := \{ f \in \mathcal{F} \mid \mathrm{E}_{P_{\text{tr}}}[(Y - f(X))Z] = 0 \}.$$

Second, they assume that the statements defining $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ also hold for all $P \in \mathcal{P}_0$. The set $\mathcal{H}_1$ is considered by (Magliacane et al., 2018; Rojas-Carulla et al., 2018) in a linear unconfounded setting and by (Saengkyongam et al., 2022) in a nonlinear underidentified IV setting. The set $\mathcal{H}_2$ defines a conditional mean independence restriction and is considered by (Arjovsky et al., 2020) in a nonlinear unconfounded setting. More recently, in the same setting as (Arjovsky et al., 2020), (Krueger et al., 2021) consider the slightly stronger restriction $\mathcal{H}_2 \cap \{f \in \mathcal{F} \mid \mathrm{E}_{P_{\text{tr}}}[(Y - f(X))^2] = c \text{ for some } c \in \mathbb{R}\}$. The set $\mathcal{H}_3$ imposes an unconditional moment constraint and is studied by (Jakobsen and Peters, 2022) in a linear underidentified IV setting and by (Bühlmann, 2020; Christiansen et al., 2022) in a nonlinear underidentified IV setting. We now investigate the relationship between the sets $\mathcal{I}_0$, $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ by showing that the notion of invariance proposed in this work is stronger than the stochastic independence of the residuals. In particular, we show that some functions in $\mathcal{H}_1$ can fail to be invariant.

Proposition 8 (Invariance implies independence of residuals) *Assume Setting 1 and that for all $z \in \mathbb{R}^r$ it holds that $\delta_z \in \mathcal{Q}_0$. Then, it holds that $\mathcal{I}_0 \subseteq \mathcal{H}_1$. Moreover, if Assumption 1 holds, it follows that $\mathcal{H}_1 \subseteq \mathcal{I}_0$. If Assumption 1 is not satisfied, then even under Assumption 2, it may happen that $\mathcal{H}_1 \neq \mathcal{I}_0$.*

The following example, whose construction is in spirit related to (Tsai et al., 2024, Example 4.4), shows that $\mathcal{H}_1 \neq \mathcal{I}_0$.

Example 1 (Independence of residuals does not imply invariance) *Consider the SIMDG $(f_0, 1, \Lambda_0, \mathcal{Q}_0)$, where the vector of observed variables $(X, Y, Z) \in \mathbb{R}^3$ is defined as*

$$X = Z + V, \quad Y = f_0(X) + U,$$

$Z \sim Q_{\text{tr}}$ is supported on $\{-2\pi, 2\pi\}$ with $\mathbb{E}_{Q_{\text{tr}}}[Z] = 0$ and $(U, V) \sim \Lambda_0$ follows a multivariate centered Gaussian distribution such that $V \sim N(0, \sigma^2)$, $U \sim N(0, 1+\sigma^2)$, and $\mathrm{E}[U \mid V] = V$. Under P_{tr}, X is a Gaussian mixture with strictly positive density on $\mathbb{R}$; hence Assumption 2 is satisfied. Consider the function $f(x) := f_0(x) + \sin(x)$. Under P_{tr}, the residuals

$$Y - f(X) = U - \sin(X) = U - \sin(Z + V) = U - \sin(V),$$

since $\sin(Z + V) = \sin(V)$ for any $Z \in \{-2\pi, 2\pi\}$. Therefore, we have that $Y - f(X) \perp\!\!\!\perp Z$ under P_{tr}, i.e., $f \in \mathcal{H}_1$. Now, consider the distribution $P_{\text{te}} \in \mathcal{P}_0$ induced by replacing Q_{tr} with a point mass distribution at π, i.e., $Q_{\text{te}} = \delta_\pi$. Under P_{te}, the residuals $Y - f(X) = U - \sin(\pi + V)$ are distributed differently from $U - \sin(V)$. Therefore, $P_{\text{tr}}^{Y-f(X)} \neq P_{\text{te}}^{Y-f(X)}$, i.e., $f \notin \mathcal{I}_0$.

Under the setting of Example 1, we show in Section 4.2 an even stronger result than $\mathcal{H}_1 \neq \mathcal{I}_0$: optimizing the training risk over the larger set $\mathcal{H}_1 \supseteq \mathcal{I}_0$ can yield a function that is not the IMP.

4. Boosted Control Functions

Assume Setting 1 with the SIMDG $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$. In the conventionally considered case in which f_0 is identifiable (which implies $\operatorname{rank}(M_0) = p$) and nonlinear, there are two prevalent categories of methods for identifying f_0; nonparametric IV methods (Newey and Powell, 2003) and nonlinear control function approaches (Ng and Pinkse, 1995; Newey et al., 1999). The nonparametric IV (Newey and Powell, 2003) identifies the structural function f_0 by solving the inverse problem

$$\mathrm{E}_{P_{\mathrm{tr}}}[Y \mid Z] = \mathrm{E}_{P_{\mathrm{tr}}}[f_0(X) \mid Z]. \tag{7}$$

The nonparametric IV estimator of f_0 is then given as the solution to a finite-sample version of (7), where f_0 is approximated by power series or splines, for example. The nonlinear control function approach (Ng and Pinkse, 1995; Newey et al., 1999) identifies the structural function f_0 by computing a conditional expectation of the response Y given the predictors X and a set of control variables $V \coloneqq X - \mathrm{E}_{P_{\mathrm{tr}}}[X \mid Z]$. We then have

$$\begin{aligned} \mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X, V] &= \mathrm{E}_{P_{\mathrm{tr}}}[f_0(X) + U \mid X, V] \\ &= f_0(X) + \mathrm{E}_{P_{\mathrm{tr}}}[U \mid V] \\ &= f_0(X) + \gamma_0(V), \end{aligned} \tag{8}$$

where $\gamma_0 : v \mapsto \mathrm{E}_{P_{\mathrm{tr}}}[U \mid V = v]$ is the control function (hence the name). Later we will assume that the control function γ_0 belongs to a class $\mathcal{G}$ of measurable functions. In contrast to nonparametric IV methods, the nonlinear control function approach has the advantage of identifying the structural function f_0 via the conditional expectation in (8), and therefore can estimate f_0 with flexible nonparametric estimators, such as nearest-neighbor regression or regression trees.

In this work, we are not interested in the structural function f_0 directly but in a function that achieves distribution generalization. The key idea is to adapt the control function approach above in a way that allows us to identify the IMP. In contrast to f_0, the IMP can be identifiable even in settings where $\operatorname{rank}(M_0) < p$. The following example illustrates the non-identifiability of f_0 based on the standard control function approach.

Example 2 (Non-identifiability of f_0 and γ_0) *Consider a SIMDG over the variables $(X_1, X_2, Y, Z) \in \mathbb{R}^4$ with the structural equations*

$$\begin{aligned} X_1 &= Z + V_1, \\ X_2 &= Z + V_2, \\ Y &= f_0(X_1, X_2) + U, \end{aligned}$$

where $f_0(x_1, x_2) \coloneqq x_1$, Λ_0 is a zero mean Gaussian distribution with $(U, V_1, V_2) \sim \Lambda_0$ satisfies $\mathrm{E}[U|V_1, V_2] = V_1$ and $\mathcal{Q}_0$ the set of all distributions on $\mathbb{R}$ with full support. Expanding the conditional expectation and using that P_{tr}-a.s. it holds that $X_1 - X_2 + V_2 = V_1$ we get

$$\begin{aligned} \mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X_1, X_2, V_1, V_2] &= f_0(X_1, X_2) + \gamma_0(V_1, V_2) = X_1 + V_1 \\ &= 2X_1 - X_2 + V_2 = \tilde{f}(X_1, X_2) + \tilde{\gamma}(V_1, V_2), \end{aligned}$$

where $\gamma_0(v_1, v_2) \coloneqq v_1$, $\tilde{f}(x_1, x_2) \coloneqq 2x_1 - x_2$ *and* $\tilde{\gamma}(v_1, v_2) \coloneqq v_2$. *Therefore, while the conditional expectation* $\mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X, V]$ *is identifiable, the separation into the structural function* f_0 *and the control function* γ_0 *is not.*

While it may happen that several functions f and γ satisfy P_{tr}-a.s. that $f(X) + \gamma(V) = \mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X, V]$, we will show that any such pair f and γ can be used to construct a specific target function that achieves distribution generalization. To construct this target function, let $q \coloneqq \mathrm{rank}(M_0)$ and define

$$R \coloneqq \begin{cases} (r_1, \dots, r_{p-q}) & \text{if } q < p \\ \mathbf{0} & \text{if } q = p, \end{cases} \tag{9}$$

where $(r_1, \dots, r_{p-q}) \in \mathbb{R}^{p \times (p-q)}$ is an orthonormal basis of $\ker(M_0^\top)$ if $q < p$ and $\mathbf{0} \in \mathbb{R}^{p \times 1}$ is the zero map. The matrix R allows us to extract invariant parts of X since $R^\top X = R^\top M_0 Z + R^\top V = R^\top V$, which has a fixed distribution for all $P \in \mathcal{P}_0$. We use it to construct the target function as follows.

Definition 9 (Boosted control function (BCF)) *Assume Setting 1, define* $\gamma_0(V) \coloneqq \mathrm{E}_{P_{\mathrm{tr}}}[U \mid V]$ *and* R *as in* (9). *Then, we define the* boosted control function (BCF) $f_\star \in \mathcal{F}$ *for* P_{tr}*-a.e.* $x \in \mathbb{R}^p$ *by*

$$f_\star(x) \coloneqq f_0(x) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid R^\top X = R^\top x]. \tag{10}$$

In particular, if $p = \mathrm{rank}(M_0)$*, the BCF is* $f_\star(\cdot) \coloneqq f_0(\cdot) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V)] = f_0(\cdot)$*, since* $\mathrm{E}_{P_{\mathrm{tr}}}[U] = 0$.

The BCF $f_\star$ is motivated by the IMP introduced in Definition 7. In particular, $f_\star$ is defined as the sum of the structural function f_0 and a term depending on X that is as predictive as possible and, at the same time, invariant. The name boosted control function alludes to the second component which extracts invariant predictive information from the remainder term $\gamma_0(V)$. In the following section, we provide conditions under which the control function is identifiable.

4.1 Identifiability Conditions

For the BCF $f_\star$ to be useful, we first need to ensure that it is indeed identifiable from the training distribution P_{tr}, that is, if f_0 and γ_0 are replaced by any other functions f and γ satisfying $f(X) + \gamma(V) = \mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X, V]$ in (10) the BCF $f_\star$ does not change. Formally, identifiability is defined as follows.

Definition 10 (Identifiability of BCF) *Assume Setting 1 and define* R *as in* (9)*. Assume* $f_0 \in \mathcal{F}$ *and* $\gamma_0 \in \mathcal{G}$*, where* $\mathcal{F}$ *and* $\mathcal{G}$ *are classes of measurable functions. The BCF* $f_\star$ *is identifiable (with respect to* $\mathcal{F}$ *and* $\mathcal{G}$*) from the observational distribution* P_{tr} *if and only if for all* $f \in \mathcal{F}$ *and* $\gamma \in \mathcal{G}$ *the following statement holds*

$$\mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X, V] = f(X) + \gamma(V),\ P_{\mathrm{tr}}\text{-a.s.}$$

$$\Longrightarrow$$

$$\begin{cases} f_\star(X) = f(X) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma(V) \mid R^\top X]\ P_{\mathrm{tr}}\text{-a.s.}, & \text{if } q < p, \\ f_\star(X) = f(X) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma(V)]\ P_{\mathrm{tr}}\text{-a.s.} & \text{if } q = p. \end{cases}$$

An equivalent definition of identifiability of $f_\star$ states that for any additive function $f(X) + \gamma(V)$ satisfying (8) the difference $f(X) - f_0(X)$ can be written as a function $\delta(R^\top X)$ depending on X only via the null space of $M_0^\top$. This is formalized in the following proposition.

Proposition 11 (Equivalent condition for identifiability) *Assume Setting 1, let* $q = \text{rank}(M_0)$ *and define* R *as in* (9). *The BCF* $f_\star$ *is identifiable from the observational distribution* P_{tr} *if and only if for all measurable functions* $h, g : \mathbb{R}^p \to \mathbb{R}$ *the following statement holds,*

$$
\begin{gathered}
h(X) + g(V) = 0, \;\; P_{\text{tr}}\text{-}a.s. \\
\Longrightarrow \\
\begin{cases} \exists \delta : \mathbb{R}^{p-q} \to \mathbb{R} \text{ such that } h(X) = \delta(R^\top X), \;\; P_{\text{tr}}\text{-}a.s., & \text{if } q < p, \\ \exists c \in \mathbb{R} \text{ such that } h(X) = c, \;\; P_{\text{tr}}\text{-}a.s., & \text{if } q = p. \end{cases}
\end{gathered}
\tag{11}
$$

A proof can be found in Appendix B.2. Proposition 11 can be seen as an extension of Newey et al. (1999)'s identifiability condition to the underidentified setting: Newey et al. (1999) consider only the case $q = p$; a further difference to that work is that our goal is to achieve distribution generalization rather than identifying the structural function f_0. Identifiability of the BCF $f_\star$ depends on the assumptions we are willing to make on the class of structural functions f and the control functions γ. Depending on whether the exogenous variable Z is categorical or continuous, we now provide two different sufficient conditions for identifiability of the BCF.

Assumption 3 (Categorical Z and linear γ_0) *Let* $\mathcal{G}$ *be the class of linear functions. The exogenous variable* Z *is categorical with values in a finite subset* $\mathcal{Z} \subseteq \mathbb{R}^r$ *(see, e.g., the construction proposed in Remark 3) such that for all* $z \in \mathcal{Z}$ $P_{\text{tr}}(Z = z) > 0$, $\text{E}_{P_{\text{tr}}}[Z] = 0$ *and* $\text{E}_{P_{\text{tr}}}[ZZ^\top] \succ 0$ *(which implies that* Z *takes at least* $r+1$ *different values). Moreover, the control function* $\gamma_0 : \mathbb{R}^p \to \mathbb{R}$ *is linear. Let* $q = \text{rank}(M_0)$ *and assume that there exists* $\{z_1, \tilde{z}_1, \dots, z_q, \tilde{z}_q\} \subseteq \mathcal{Z}$ *such that for all* $j \in \{1, \dots, q\}$ *the distributions* $P_{\text{tr}}^{V+M_0 z_j}$ *and* $P_{\text{tr}}^{V+M_0 \tilde{z}_j}$ *are not mutually singular*[2]*, and* $\text{span}(\{M_0(z_j - \tilde{z}_j) \mid j \in \{1, ..., q\}\}) = \text{im}(M_0)$.

Assumption 3 ensures that we observe the same realization of the predictor X under distinct environments, and these environments span a space that is rich enough. The following assumption is a slight modification of (Newey et al., 1999, Theorem 2.3) with the difference that we do not require $\text{rank}(M_0) = p$ since we allow for underidentified settings, too.

Assumption 4 (Differentiable f_0 and γ_0) *Let* $\mathcal{F}$ *and* $\mathcal{G}$ *each be the class of differentiable functions. The boundary of the support of* (V, Z) *has zero probability under* P_{tr} *and the interior of the support of* (V, Z) *is convex under* P_{tr} *(this implies that* Z *is not discrete). Furthermore, the control and structural functions* $\gamma_0, f_0 : \mathbb{R}^p \to \mathbb{R}$ *are differentiable.*

Under either Assumption 3 or 4, the observational distribution P_{tr} contains enough information to identify the BCF $f_\star$.

2. Two probability measures μ, ν on a space $(\Omega, \mathbb{F})$ are mutually singular if there exists a measurable set $E \in \mathbb{F}$ such that $\mu(E^c) = \nu(E) = 0$.

Algorithm 1 Boosted control function identification

Input: Observational distribution $P_{\text{tr}}^{X,Y,Z}$ over (X, Y, Z).

Output: BCF $f_\star : \mathbb{R}^p \to \mathbb{R}$.

1: Compute conditional expectation $\mathrm{E}_{P_{\text{tr}}}[X \mid Z] = M_0 Z$ to identify $M_0 \in \mathbb{R}^{p\times r}$.

2: Compute the control variable $V = X - M_0 Z$ and a basis $R \in \mathbb{R}^{p\times(p-q)}$ for the left null space $\ker(M_0^\top)$.

3: Compute the additive conditional expectation $\mathrm{E}_{P_{\text{tr}}}[Y \mid X, V] = f_0(X) + \gamma_0(V)$.

4: Compute the conditional expectation $\mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top X]$.

5: Return $f_\star(x) \coloneqq f_0(x) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top X = R^\top x]$.

Proposition 12 (BCF $f_\star$ is identifiable) *Assume Setting 1. Suppose that either Assumption 3 or Assumption 4 holds. Then, the BCF $f_\star$ is identifiable from P_{tr}.*

A proof can be found in Appendix B.3. Algorithm 1 provides a procedure for identifying the BCF $f_\star$ from the observed distribution if it is identifiable. We will now see that $f_\star$ indeed comes with generalization guarantees and is invariant most predictive.

4.2 Generalization Guarantees

In this section, we study under which conditions the BCF $f_\star$ is a generalizing function, that is

$$\sup_{P\in\mathcal{P}_0} \mathcal{R}(P, f_\star) = \inf_{f\in\mathcal{F}} \sup_{P\in\mathcal{P}_0} \mathcal{R}(P, f),$$

where $\mathcal{P}_0$ denotes the set of distributions induced by the SIMDG as defined in (4). Clearly, if $\mathcal{P}_0$ only contains P_{tr}, then the least squares solution is optimal in terms of distribution generalization. In particular, it is possible to quantify the additional risk of the BCF $f_\star$ on the training distribution P_{tr} compared to the least squares prediction $f_{\text{LS}} \colon x \mapsto \mathrm{E}_{P_{\text{tr}}}[Y \mid X = x]$.

Proposition 13 *Assume Setting 1, let $\gamma_0(V) = \mathrm{E}_{P_{\text{tr}}}[U \mid V]$ and define R as in (9) Then, the BCF $f_\star$ is invariant according to Definition 6. Moreover, the mean squared error risks of $f_\star$ and the least squares predictor f_{LS} are related by*

$$\mathcal{R}(P_{\text{tr}}, f_\star) = \mathcal{R}(P_{\text{tr}}, f_{\text{LS}}) + \mathrm{E}_{P_{\text{tr}}}\left[\left(\mathrm{E}_{P_{\text{tr}}}[\gamma_0(V)|R^\top X] - \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V)|X]\right)^2\right]. \tag{12}$$

A proof can be found in Appendix B.4. The additional term in (12), is the price one needs to pay in terms of prediction performance to be invariant. That is, if $P_{\text{te}} = P_{\text{tr}}$ then the BCF is outperformed by the least squares predictor by exactly this term. In contrast, if there is sufficient heterogeneity in $\mathcal{P}_0$, then the BCF $f_\star$ is a generalizing function. Formally, we use the following assumption to quantify what is meant by sufficient heterogeneity.

Assumption 5 *Assume Setting 1, let $\gamma_0(V) = \mathrm{E}_{P_{\text{tr}}}[U \mid V]$ and define R as in (9). Then, assume that*

$$\inf_{P\in\mathcal{P}_0} \mathrm{E}_P\left[\left(\mathrm{E}_P[\gamma_0(V) \mid R^\top X] - \mathrm{E}_P[\gamma_0(V) \mid X]\right)^2\right] = 0.$$

Whether this assumption holds depends on the class of distributions $\mathcal{P}_0$ and on the control function γ_0. From the SCM perspective, the assumption requires that the interventions on Z are strong enough so that the only part of X that is relevant to explain the confounder V is the invariant one, i.e., $R^\top X$. Assumption 5 holds, for example, when the noise terms are jointly Gaussian or when the control function $\gamma_0(V)$ is bounded, as long as the class of distributions $\mathcal{Q}_0$ contains standard Gaussian distributions with arbitrarily large variance.

Proposition 14 *Assume Setting 1 and suppose that*

$$\{N(0, k^2 I_r) \mid k \in \mathbb{N}\} \subseteq \mathcal{Q}_0.$$

Suppose the joint distribution Λ_0 of (U, V) satisfies one of the following conditions.

(a) V has a density w.r.t. Lebesgue, and the control function $\gamma_0(V) = \mathrm{E}_{P_{\mathrm{tr}}}[U \mid V]$ is almost surely bounded.

(b) Λ_0 is a multivariate centered Gaussian distribution (with a non-degenerate covariance matrix).

Then, Assumption 5 holds.

A proof can be found in Appendix B.5. We thank Alexander M. Christgau for pointing us to their trigonometric argument in Christgau and Hansen (2024) that turned out to be helpful to prove part (a) of Proposition 14. Condition (a) in Proposition 14 allows for a wide range of distributions, including those inducing bounded random variables with densities, for example.

Theorem 15 *Assume Setting 1. Let $f_\star \in \mathcal{I}_0$ denote the BCF. Suppose that Assumptions 2 and 5 hold. Then,*

$$\sup_{P \in \mathcal{P}_0} \mathcal{R}(P, f_\star) = \inf_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}_0} \mathcal{R}(P, f).$$

A proof can be found in Appendix B.6. As a corollary of Theorem 15, we have that the BCF $f_\star$ is indeed the IMP.

Corollary 16 *Assume Setting 1 and suppose that Assumptions 2 and 5 hold. Let $f_\star \in \mathcal{I}_0$ denote the BCF. Then,*

$$\mathcal{R}(P_{\mathrm{tr}}, f_\star) = \inf_{f \in \mathcal{I}_0} \mathcal{R}(P_{\mathrm{tr}}, f).$$

A proof can be found in Appendix B.7. We conclude the section by revisiting Example 1, which showed that the set $\mathcal{H}_1 = \{f \in \mathcal{F} \mid Y - f(X) \perp\!\!\!\perp Z \text{ under } P_{\mathrm{tr}}\}$ may contain non-invariant functions. Here, we provide an even stronger result, namely that optimizing the training risk over $\mathcal{H}_1 \supseteq \mathcal{I}_0$ may not yield the IMP.

Example 3 (Optimizing the training risk over $\mathcal{H}_1$ may not return the IMP) *Consider the setting of Example 1 and recall that the structural function f_0 is invariant, i.e., $f_0 \in \mathcal{I}_0$. By definition, the BCF coincides with the structural function f_0. As we argued in Example 1, this setup satisfies Assumption 2. Moreover, it satisfies Assumption 3:*

Z is a centered, non-degenerate discrete random variable, the function $v \mapsto \gamma_0(v) := v$ is linear, $P_{\text{tr}}^{V+2\pi} = N(2\pi, \sigma^2)$ and $P_{\text{tr}}^{V-2\pi} = N(-2\pi, \sigma^2)$ are not mutually singular, and $\text{span}(\{2\pi - (-2\pi)\}) = \mathbb{R}$. Hence, by Proposition 12 the BCF f_0 is identifiable from P_{tr}.

Suppose that the set of perturbed distribution satisfies $\{N(0, k^2) \mid k \in \mathbb{N}\} \subseteq \mathcal{Q}_0$. Then, by Proposition 14, Assumption 5 is fulfilled and by Corollary 16 the BCF f_0 is the IMP function.

Now, assume that the noise variance of V is $\sigma^2 = 2$. We show that the non-invariant function $f(x) := f_0(x) + \sin(x) \in \mathcal{H}_1 \setminus \mathcal{I}_0$ achieves a lower risk than f_0; hence, optimizing over $\mathcal{H}_1$ does not yield the IMP. By Lemma 19,

$$\mathcal{R}(P_{\text{tr}}, f) = 1 + \sigma^2 + \frac{(1 - e^{-2\sigma^2})}{2} - 2\sigma^2 e^{-\sigma^2/2}.$$

For $\sigma^2 = 2$, one can verify that $\mathcal{R}(P_{\text{tr}}, f) < 1 + \sigma^2 = \mathcal{R}(P_{\text{tr}}, f_0)$, and therefore

$$\inf_{f \in \mathcal{H}_1} \mathcal{R}(P_{\text{tr}}, f) < \mathcal{R}(P_{\text{tr}}, f_0).$$

4.3 Estimating Boosted Control Functions

Consider Setting 1 and let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ be the i.i.d. training sample drawn from the training distribution $P_{\text{tr}} \in \mathcal{P}_0$. We denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{Z} \in \mathbb{R}^{n \times r}$ the respective design matrices resulting from row-wise concatenations of the observations. The algorithm to identify the BCF from P_{tr}, Algorithm 1, consists of two parts. First (lines 1–2), it computes the matrix $M_0 \in \mathbb{R}^{p \times r}$, the control variables $V := X - M_0 Z$, and a basis $R \in \mathbb{R}^{p \times (p-q)}$ for the null space of $M_0^\top$. Second (lines 3–4), it computes the additive function $(x, v) \mapsto \mathrm{E}_{P_{\text{tr}}}[Y \mid X = x, V = v]$ and the resulting BCF. We now convert each part into an estimation procedure.

For the first part of Algorithm 1, we consider the framework of reduced-rank regression (Reinsel and Velu, 1998) to estimate the matrix $\hat{M}_0$ and its rank $\hat{q} \leq \min\{p, r\}$, since $M_0 \in \mathbb{R}^{p \times r}$ is not necessarily full rank. More precisely, we adopt the rank selection criterion proposed by Bunea et al. (2011). For a fixed penalty $\lambda > 0$, they propose to estimate M_0 by

$$\hat{M}_0(\lambda) := \arg\min_{M} \left\{ \left\| \mathbf{X} - \mathbf{Z} M^\top \right\|_F^2 + \lambda \operatorname{rank}(M) \right\}.$$

To do so, they observe

$$\min_{M} \left\{ \left\| \mathbf{X} - \mathbf{Z} M^\top \right\|_F^2 + \lambda \operatorname{rank}(M) \right\} = \min_{k} \left\{ \min_{M,\ \operatorname{rank}(M) = k} \left[\left\| \mathbf{X} - \mathbf{Z} M^\top \right\|_F^2 + \lambda k \right] \right\},$$

which suggests that one can estimate a rank-k matrix $\hat{M}_{0,k}(\lambda)$ for each $k \in \{1, \ldots, \min\{p, r\}\}$ and choose the optimal $\hat{q}(\lambda) := \arg\min_k \|\mathbf{X} - \mathbf{Z}\hat{M}_{0,k}(\lambda)^\top\|_F^2 + \lambda k$. As mentioned by Bunea et al. (2011), $\hat{M}_{0,k}(\lambda)$ and $\hat{q}(\lambda)$ can be computed in closed form and efficiently. Moreover, one can select the optimal λ^* by cross-validation. This results in the final estimators $\hat{M}_0 := \hat{M}_{0,\hat{q}(\lambda^*)}(\lambda^*)$ and $\hat{q} := \hat{q}(\lambda^*)$. Bunea et al. (2011) show that their rank selection criterion consistently recovers the true rank q of M_0 if V is multivariate standard Gaussian and

Algorithm 2 ControlTwicing

Input: data $(\mathbf{X}, \mathbf{Y}, \hat{\mathbf{V}}) \in \mathbb{R}^{n\times(p+1+p)}$; nonparametric regressors $\hat{f} \in \mathcal{F}$ and $\hat{\gamma} \in \mathcal{G}$; max passes J.
Output: $\hat{f} \in \mathcal{F}$, $\hat{\gamma} \in \mathcal{G}$.
1: $\widetilde{\mathbf{Y}} \leftarrow \mathbf{Y} - \bar{\mathbf{Y}}$
2: **for** $j \in \{1, \dots, J\}$ **do**
3: Estimate $\hat{\gamma}$ based on $\widetilde{\mathbf{Y}} \sim \hat{\mathbf{V}}$
4: $\widetilde{\mathbf{Y}} \leftarrow \mathbf{Y} - \hat{\gamma}(\hat{\mathbf{V}})$
5: Estimate $\hat{f}$ based on $\widetilde{\mathbf{Y}} \sim \mathbf{X}$
6: $\widetilde{\mathbf{Y}} \leftarrow \mathbf{Y} - \hat{f}(\mathbf{X})$

if the q-th largest singular value of the signal $M_0 Z$ is well separated from the smaller ones. Based on these estimators $\hat{M}_0$ and $\hat{q}$, we then estimate the control variables $\hat{\mathbf{V}} := \mathbf{X} - \mathbf{Z}\hat{M}_0^\top$ and a basis of the null space by $\hat{R} := (r_1, \dots, r_{\hat{q}}) \in \mathbb{R}^{p\times(p-\hat{q})}$, where $r_1, \dots, r_{p-\hat{q}}$ are the $p - \hat{q}$ left singular vectors of $\hat{M}_0$ associated to zero singular values.

For the second part, given the estimated control variables $\hat{\mathbf{V}}$ and the basis of the estimated null space $\hat{R}$, we then estimate the BCF. As shown in Algorithm 1, the BCF is obtained by performing two separate regressions such that

(i) $(x, v) \mapsto \hat{f}(x) + \hat{\gamma}(v)$ estimates $(x, v) \mapsto \mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X = x, V = v]$,

(ii) $x \mapsto \hat{\delta}(\hat{R}^\top x)$ estimates $x \mapsto \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid R^\top X = R^\top x]$.

For the additive function in step (i), we devise a practical estimator inspired by the alternating conditional expectation (ACE) algorithm by Breiman and Friedman (1985). Unlike Breiman and Friedman (1985), here we do not assume full additivity of the functions, but allow $\tilde{f}$ and $\tilde{\gamma}$ to belong to general function classes $\mathcal{F}$ and $\mathcal{G}$, respectively. The procedure estimates $\tilde{f}$ and $\tilde{\gamma}$ by alternating between the two regressions. Formally, we assume we have two arbitrary nonparametric regression methods that result in estimates $\hat{f} \in \mathcal{F}$ and $\hat{\gamma} \in \mathcal{G}$. We then start by estimating $\hat{\gamma}$ based on a regression of $\mathbf{Y} - \frac{1}{n}\sum_{i=1}^n Y_i$ on $\hat{\mathbf{V}}$ after which we estimate $\hat{f}$ based on a regression of $\mathbf{Y} - \hat{\gamma}(\hat{\mathbf{V}})$ on $\mathbf{X}$, and then iterate this after updating $\mathbf{Y}$ at each step (see Algorithm 2). We call the algorithm ControlTwicing; the first part of the name alludes to the fact that we deal with a regression problem in a control function setup. The second part of the name refers to the twicing idea (see Tukey et al., 1977, Chapter 16) consisting of fitting an additive model over repeated iterations. Once we estimate $\hat{f}$ and $\hat{\gamma}$, in step (ii), we again use a nonparametric regression procedure (generally the same as the one used in step (i) to estimate $\hat{f}$) to obtain an estimate $\hat{\delta}$. More specifically, we regress the pseudo response $\hat{\gamma}(\hat{\mathbf{V}})$ on the pseudo covariates $\mathbf{X}\hat{R}$. The final estimated BCF is then defined for all $x \in \mathbb{R}^p$ by $\hat{f}_\star(x) := \hat{f}(x) + \hat{\delta}(\hat{R}^\top x)$.

In Proposition 21 in Appendix B.8, we show that when using sampling-splitting the overall convergence rate of the BCF estimator is at least as fast as the slower of the two rates from steps (i) and (ii). For simplicity, the result assumes the first-step quantities (M_0, R, V) are known.

5. Numerical Experiments

We now study the properties of the estimated BCF on simulated data. In our first experiment, we analyze how well the BCF and the oracle IMP generalize to testing distributions. In the second experiment, we consider how well the reduced-rank regression estimates the matrix M_0. The last experiment considers the California housing dataset (Pace and Barry, 1997) to evaluate the robustness of the BCF estimator to distributional shifts compared to the least squares estimator. The code to reproduce the results can be found at `https://github.com/nicolagnecco/bcf-numerical-experiments`.

5.1 Experiment 1: Predicting unseen interventions

In the first experiment, we assess the predictive performance of the BCF estimator compared to the (oracle) IMP and the least squares (LS) estimator in a fixed SIMDG. We measure the predictive performance by the mean squared error (MSE) between the response and the predicted values. The BCF is estimated using a random forest with 100 fully-grown regression trees for $\hat{f}$ and $\hat{\delta}$ and ordinary linear least squares for $\hat{\gamma}$. The number of passes in the ControlTwicing algorithm (see Algorithm 2) is set to $J = 2$. The LS method uses a random forest (with the same parameters) to regress $\mathbf{Y}$ on $\mathbf{X}$. The theoretical IMP corresponds to the population version of the BCF (see Corollary 16) and can be computed in closed form. As additional baselines, we consider the MSE of predicting Y with its unconditional mean (constant mean estimator) and with the true structural function. We generate data from a SIMDG $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$ with the following specifications. The function $f_0 : \mathbb{R}^p \to \mathbb{R}$ is a decision tree depending on a subset of the predictors (further details on how the tree is sampled can be found in Appendix D). The matrix $M_0 \coloneqq AB^\top \in \mathbb{R}^{p\times r}$ is a rank-q matrix where $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{r\times q}$ are orthonormal matrices sampled from the Haar measure, i.e., the uniform distribution over orthonormal matrices. The singular values of M_0 are $\sigma_1 = \cdots = \sigma_q = 1$ and $\sigma_{q+1} = \cdots = \sigma_{\min\{p,r\}} = 0$. The distribution Λ_0 over $\mathbb{R}^{p+1}$ is a mean-zero Gaussian with $\mathrm{E}[VV^\top] = I_p$, $\mathrm{E}[U^2] = c^2 + 0.1^2$ and $\mathrm{E}[VU] = c\eta$, where $c > 0$ denotes the confounding strength and $\eta \in \mathbb{R}^p$ is a vector sampled uniformly on the unit sphere. Here, we set the number of predictors to $p = 10$, the number of exogenous variables to $r = 5$, and the confounding strength to $c = 2$. Finally, we define the set of distributions over the exogenous variables by $\mathcal{Q}_0 \coloneqq \{N(0, k^2 I_r) \mid k > 0\}$ and define $Q_{\mathrm{tr}} \coloneqq N(0, I_r)$ and $Q_{\mathrm{te}}^k \coloneqq N(0, k^2 I_r)$ for all $k \geq 1$. Therefore, k specifies the perturbation strength relative to the training distribution.

Figure 1 displays the MSE of the BCF estimator and the competing methods, averaged over 50 repetitions for different perturbation strengths $k \in \{1, \ldots, 10\}$. For each repetition, we generate random instances of f_0 and M_0 and then generate $n = 1000$ i.i.d. observations $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ from the model $(f_0, M_0, \Lambda_0, Q_{\mathrm{tr}})$ on which we train each method. For all $k \in \{1, \ldots, 10\}$, we then generate $n = 1000$ i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the model $(f_0, M_0, \Lambda_0, Q_{\mathrm{te}}^k)$, which we use to evaluate MSE. Each panel consists of a different value of $q = \mathrm{rank}(M_0)$. For all values of q, the results demonstrate that the BCF estimator indeed performs similarly to the theoretical IMP and its performance remains approximately invariant even for large perturbations. Moreover, the BCF estimator outperforms the true structural function since it further uses all the signal in $\gamma_0(V)$ that is invariant to shifts in the exogenous variable. As the value of q increases,

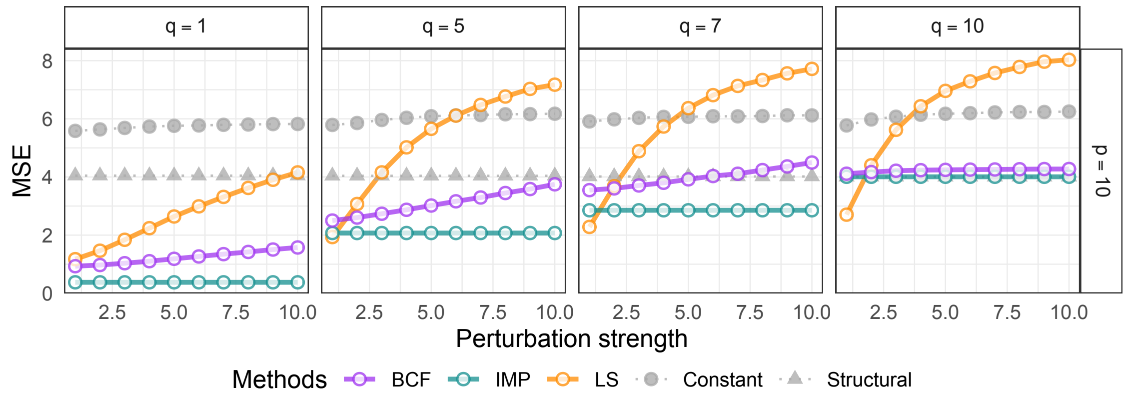

Figure 1: MSE between the response and the predicted values for increasing perturbation strength over different values of $q = \text{rank}(M_0)$. We fix the number of predictors to $p = 10$, the number of exogenous variables to $r = 5$, and the confounding strength to $c = 2$. Each point is an average over 50 repetitions. The BCF estimator achieves a performance close to the theoretical optimally achievable error given by the IMP and is in particular better than the structural function. The LS baseline is clearly non-invariant and performs worse than the competing methods when the perturbation in the testing distribution is large.

for large perturbation strength k, the gain in predictive performance of BCF over LS becomes more pronounced. This is because q determines the dimension of the subspace in the predictor space where perturbations occur. In the case where $q = p$, the perturbations can occur in any direction in the predictor space, and therefore the performance of non-invariant methods deteriorate significantly. Moreover, for increasing values of q, the BCF estimator and the theoretical IMP converge to the structural function f_0. This is because the dimension $p - q$ of the invariant space $\ker(M_0^\top)$ decreases as q increases; in the limit case when $q = p$, we have that $\ker(M_0^\top) = \{0\}$ and therefore the BCF corresponds to the structural function. Finally, as the value of q increases, when $k = 1$ and thus the training and testing distribution are the same, the LS estimator performs better than the BCF estimator and the theoretical IMP. This behavior is expected because the LS estimator can always use all the information in X to predict $\gamma_0(V)$, while the BCF and IMP can only use the information in the invariant space $R^\top X$ to predict $\gamma_0(V)$ and the dimension of $R^\top X$ decreases for increasing q.

5.2 Experiment 2: Estimating M_0

In the second experiment, we assess the performance of the rank selection criterion to estimate a low-rank matrix $\hat{M}_0$ and its left null space $\ker(M_0^\top)$. Given the matrix of the observed predictors $\mathbf{X}$ and exogenous variables $\mathbf{Z}$, we estimate a low-rank matrix $\hat{M}_0 \in \mathbb{R}^{p \times r}$ that is subsequently used to compute a basis for $\ker(M_0^\top)$ and the control variables $\hat{\mathbf{V}}$. The hardness of estimating the column and null spaces of a matrix depends (among other things) on its eigengap, which is defined as the size of the smallest non-zero singular value. When the eigengap of a matrix is close to zero, it is hard to disentangle its column space from its left null space (see Wainwright, 2019; Cheng et al., 2021) because eigenvectors associated with small singular values are sensitive to small perturbations of the original matrix; intuitively, small estimation errors in $\hat{M}_0$ are amplified in the estimation of the column and left null space. In this experiment, we sample a rank-q matrix $M_0 := \tau AB^T \in \mathbb{R}^{p\times r}$, where $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{r\times q}$ are orthonormal matrices sampled from the Haar measure. The singular values of M_0 are $\sigma_1, \dots, \sigma_q = \tau$ and $\sigma_{q+1}, \dots, \sigma_{\min\{p,r\}} = 0$, so that its eigengap is τ. For this experiment, we only need observations of X and Z, which we generate by first sampling $Z \sim N(0, I_r)$ and $V \sim N(0, I_p)$ independently and then generating the covariates according to $X = M_0 Z + V$.

Figure 2 displays the distance between the column space of the true matrix M_0 and the estimated matrix $\hat{M}_0$, defined as $||\Pi_{M_0} - \Pi_{\hat{M}_0}||_F^2$, averaged over 50 repetitions for increasing values of the eigengap and sample size. We estimate $\hat{M}_0$ using the rank selection criterion defined in Section 4.3. We fix the true rank to be $q = 5$ and consider different combinations of number of covariates p and exogenous variables r (which are shown in the four panels). As expected, for larger values of the eigengap τ, the distance between the true and estimated linear subspaces converges to zero much faster, compared to smaller values of τ.

5.3 California Housing Dataset

We consider the California housing dataset (Pace and Barry, 1997) consisting of 20,640 observations derived from the 1990 U.S. census. The unit of analysis is a block group, which is the smallest geographical denomination for which the U.S. Census Bureau publishes sample data. The primary aim of our experiment is to predict the median house value from the following covariates: median income, median house age, average number of rooms per household, average number of bedrooms per household, total population, average number of household members, and average annual temperature between 1991 and 2020. We use latitude and longitude as exogenous variables. The reason for excluding latitude and longitude from the set of predictors is that when extrapolating the model to new regions, latitude and longitude do not help with the prediction unless assumptions are made about the function class $\mathcal{F}$ (see Assumption 2). The temperature data are integrated from PRISM (PRISM, 2020) and are used as a three-decade average. While data from 1990 are unavailable, we conjecture that this thirty-year range adequately reflects typical weather patterns correlated with housing prices.

To evaluate the ability of the boosted control function (BCF) estimator to handle distributional shifts, we create multiple training/testing geographical splits of the data. These splits, defined by the 35th parallel north and the 120th meridian west, partition the dataset into regions such as N/S, E/W, and SE/(N+SW). For instance, SE/(N+SW) means that the models are trained on the SE region and evaluated on the union of the N and SW regions. For each split: (i) we randomly select 80% of the training split to train the methods; (ii) we evaluate the methods on the held-out 20% of the training split and the entire testing split; (iii) we repeat (i) and (ii) 10 times with different random subsamples.

The BCF estimator is configured using extreme gradient boosting (Chen and Guestrin, 2016). The learning rate for estimating $\hat{f}$, $\hat{\gamma}$, and $\hat{\delta}$ is set to 0.05, and the number of passes in the ControlTwicing algorithm (see Algorithm 2) is fixed to $J = 10$. As competing methods,

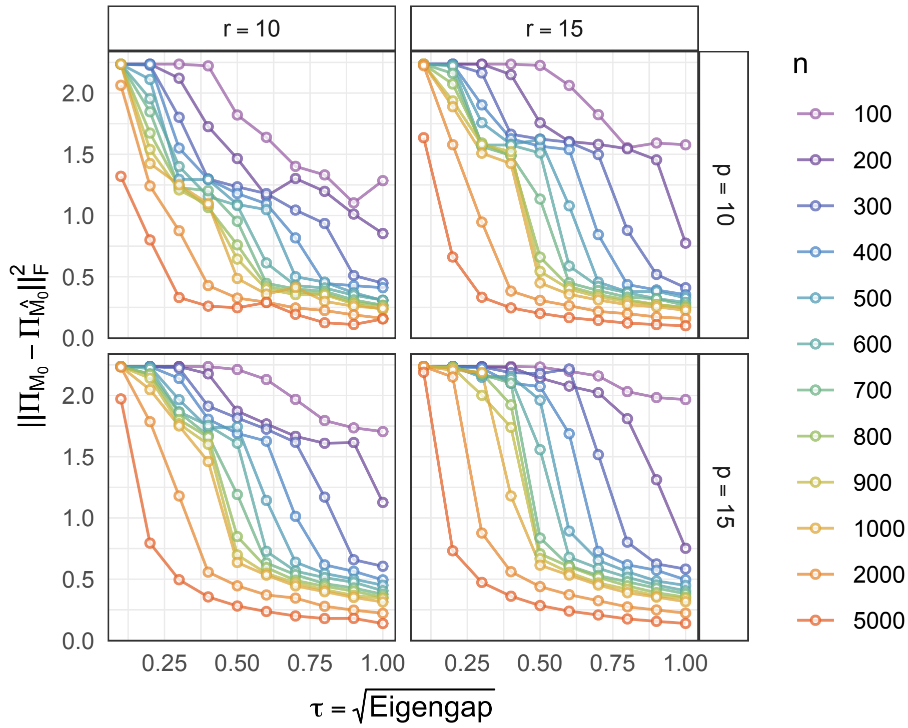

Figure 2: Average distance between the column space of the true matrix M_0 and estimated $\hat{M}_0$ for increasing eigengap τ and sample size n over different number of covariates p and exogenous variables r. We fix $q = \text{rank}(M_0) = 5$. Each point is an average over 50 repetitions. For large values of the eigengap τ, the distance between the linear subspaces converges to zero much faster, compared to smaller values of τ.

we consider GroupDRO (Sagawa et al., 2020), which estimates a predictive function that minimizes the worst-case risk across multiple training distributions, and Anchor Boosting (Londschien et al., 2025), which extends the anchor regression methodology to the nonlinear setting using gradient-boosted trees. For GroupDRO, we follow the original algorithm (Sagawa et al., 2020), but use a lightweight neural network with a single hidden layer of 64 nodes. We form the training groups by splitting each training dataset into four equally sized parts based on latitude and longitude. The model is trained using the Adam optimizer (Kingma and Ba, 2015) with a learning rate 0.001, mirror ascent step size 0.01, batch size 256, and 500 epochs. For Anchor Boosting, we consider the original implementation and follow the recommended default settings for the tuning parameters (Londschien et al., 2025, Section 4). We fit the method for several values of $\gamma \in (1, 5]$ and report the best-performing configuration ($\gamma = 1.5$).

As baselines, we include the classical control function (CF) estimator (corresponding to Lines 1–3 of Algorithm 1 and sharing the same settings as the BCF estimator), the least squares (LS) estimator (using extreme gradient boosting with a learning rate of 0.05) and a constant mean estimator as a reference.

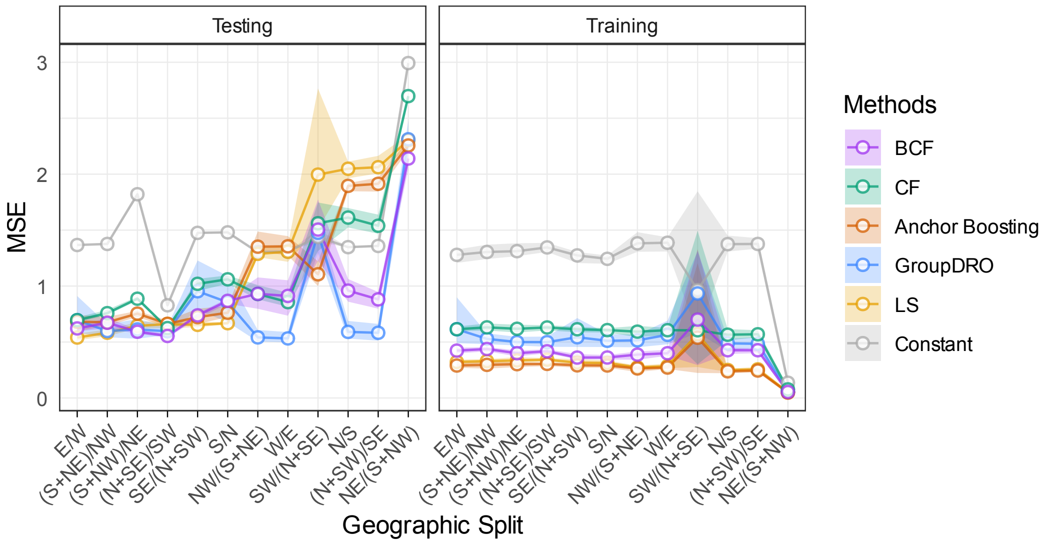

Figure 3: Mean squared error (MSE) for the BCF estimator and competing methods evaluated on the testing (left) and training (right) splits. The notation E/W means that region East was used for training, region West for testing. For each geographical split and each method, the points correspond to the average MSE over 10 repetitions, and the uncertainty bands indicate the range of MSE values over such repetitions.

Figure 3 presents the mean squared error (MSE) on the testing (left) and training (right) splits for the different methods. Each point represents the average MSE across the 10 repetitions, and the shaded bands indicate the range of MSE values (smallest to largest) observed across these repetitions. Compared to the LS and CF estimator, the BCF estimator appears more robust to distributional shifts on certain test splits, such as SW/(N+SE), (N+SW)/SE, and N/S. Compared to Anchor Boosting, the BCF generally achieves lower MSE across most test splits, except for the SW/(N+SE), where Anchor Boosting performs best. Finally, we observe that GroupDRO and BCF achieve comparable performance, despite being fundamentally different in their training procedures, how they model distributional shifts, and their generalization guarantees. On the training data, all methods perform similarly, except for the constant mean estimator.

6. Conclusion and Future Work

This work studies the challenge of distribution generalization in the presence of unobserved confounding when distributional shifts are induced by exogenous variables. We propose a strong notion of invariance that, unlike existing weaker notions, ensures distribution generalization when the structural function is nonlinear and possibly not identifiable. We define the Boosted Control Function (BCF), a novel target of inference that satisfies our strong notion of invariance and is identifiable from the training data. The BCF is provably worst-case optimal against distributional shifts induced by the exogenous variable and aligns with the invariant most predictive (IMP) function. Our theoretical analysis is built around the Simultaneous Equation Models for Distribution Generalization (SIMDGs), a novel framework that connects machine learning with econometrics by describing data-generating processes under distributional shifts. We demonstrate the practical effectiveness of our findings by introducing the ControlTwicing algorithm, which leverages nonparametric machine-learning techniques to learn the BCF. Empirical evaluations on synthetic and real-world datasets indicate that our methodology outperforms traditional machine learning methods based on empirical risk minimization when generalizing to unseen distributional shifts.

Looking forward, several routes for future research emerge. One direction is to extend our strong invariance notion to settings where the outcome variable Y influences some covariates X. Such scenarios, common in time-series and dynamic systems, have been explored in linear settings, but there is limited research on generalizing these findings to nonlinear models. Addressing this challenge could significantly broaden the applicability of invariance-based methods to more realistic and complex settings. Another direction involves understanding how non-additive effects of exogenous variables Z on (X, Y) affect our notion of invariance and distribution generalization. Previous research has focused on unconfounded settings or empirically analyzed worst-case optimality guarantees. However, formulating a notion of invariance for confounded non-additive models that leads to provable generalization guarantees remains an open question. From a theoretical perspective, it would be valuable to extend the analysis of the convergence rate of the BCF estimator to the case in which the first-step quantities are treated as random. This would require additional conditions on the nonparametric regression functions and could build on the literature of nonparametric regression with generated covariates (Mammen et al., 2012).

Acknowledgments

We thank Jeffrey Glenn Adams, Alexander M. Christgau, Anton R. Lundborg, Leonard Henckel, Manuel Hentschel, Cesare Miglioli, Sorawit Saengkyongam, and Stanislav Volgushev for helpful discussions. In particular, we thank Alexander M. Christgau for pointing out the trigonometric argument in the proof of Lemma 17. We are grateful to the editorial team and the three anonymous reviewers for the valuable comments. During parts of the project, JP was supported by a research grant (18968) from VILLUM FONDEN. SE was supported by a research grant (186858) from the Swiss National Science Foundation (SNSF). NP was supported by a research grant (0069071) from Novo Nordisk Fonden. NG was supported by a research grant (210976) from the Swiss National Science Foundation (SNSF).

Appendix A. Additional Results on SIMDGs

A.1 Generating Distributions with SIMDGs

In this section, we shortly discuss formally how a SIMDG generates the class of distributions $\mathcal{P}_0$ and provide the full technical details on the notation used throughout this work.

We assume there is a fixed background probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define $\mathcal{X} := \mathbb{R}^{1+p+r+p+1}$ and let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a fixed measurable sample space, where $\mathcal{B}(\mathcal{X})$ denotes the Borel sigma algebra on $\mathcal{X}$. Let $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$ be a SIMDG as defined in Definition 5. For all $Q \in \mathcal{Q}_0$, the model (f_0, M_0, Λ_0, Q) generates a random variable $(U_Q, V_Q, X_Q, Y_Q, Z_Q)$ with distribution P_Q on the sample space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ as follows:

(1) Let $((U_Q, V_Q), Z_Q) : \Omega \to \mathbb{R}^{p+1} \times \mathbb{R}^r$ be a random variable with distribution $\Lambda_0 \otimes Q$.

(2) Let $X_Q : \Omega \to \mathbb{R}^p$ be the random variable defined by $X_Q = M_0 Z_Q + V_Q$.

(3) Let $Y_Q : \Omega \to \mathbb{R}$ be the random variable defined by $Y = f_0(X_Q) + U_Q$.

(4) Denote the distribution of $(U_Q, V_Q, X_Q, Y_Q, Z_Q)$ by P_Q.

For each $Q \in \mathcal{Q}_0$, we therefore get a random vector $(U_Q, V_Q, X_Q, Y_Q, Z_Q)$ and a distribution P_Q on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. To ease notation, we suppress the Q subscript in the notation. Moreover, to make the dependence on the distribution clear we index probabilities and expectations with the corresponding distribution P, i.e., $\mathbb{P}_P$ and E_P.

The above procedure is similar to what has been suggested by Bongers et al. (2021) with the difference that Bongers et al. (2021) consider solutions of structural equations which then are assumed to hold almost surely, not for every $\omega \in \Omega$ (this is particularly important when introducing cycles).

A.2 SIMDGs as Triangular Simultaneous Equation Models

The SIMDG $(f_0, M_0, \Lambda_0, \mathcal{Q}_0)$ in Definition 5 consists of a set of SIMs that are closely related to the triangular simultaneous equation models considered by e.g., Newey et al. (1999). Formally, these are given by the simultaneous equations

$$\begin{aligned} Y &= g(X, Z) + U, \\ X &= h(Z) + V, \end{aligned} \quad \text{with} \quad \mathrm{E}[U \mid V, Z] = \mathrm{E}[U \mid V], \quad \mathrm{E}[V \mid Z] = \mathrm{E}[V] = 0, \tag{13}$$

where (U, V, Z) are exogenous and (X, Y) are endogenous. SIMs are a tool to model identifiable causal effects from a set of covariates (X, Z) to a response Y. The structural function g in (13) describes the average causal effects that interventions on (X, Z) have on Y. Generally, g is assumed to be identifiable, that is, it is unique for any distribution over (X, Y, Z) satisfying (13). Unlike SCMs (discussed in Appendix A.3), SIMs are generally interpreted as non-generative in the sense that they do not describe the full causal structure. Instead, they only implicitly model parts of the causal structure via the conditional mean assumptions on the residuals (U, V) and the separation of the covariates into exogenous Z and endogenous X variables. As such, SIMs naturally extend regression models to settings in which some covariates (the endogenous ones) are confounded with the response. Nevertheless, they can also be viewed – as done here for SIMDGs – as generative models: first generate the exogenous variables (Z, U, V), then generate X according to the reduced form equation and finally generate Y from the structural equation.

An attractive feature of these models is that they only describe the relationship of interest as causal (i.e., the structural equation of Y), while reducing the remaining causal relations into a reduced form conditional relationship between exogenous and endogenous variables that may be causally misspecified. Similarly, in the SIMs used in SIMDGs (3a)–(3c) we do not assume that the causal relations are correctly specified in the reduced form equation. Furthermore, as our goal is not to model causal effects but instead to model the invariant parts of the distributions in $\mathcal{P}_0$, there are several further differences between the SIMs used in SIMDGs and the commonly used ones in (13).

On the one hand, the SIMs used in SIMDGs relax (13) in two ways. First, we do not assume identifiability of the structural function f_0 (as it is often done for g in (13)) since we do not impose the necessary rank and order conditions of identifiability (Amemiya, 1985), i.e., $\text{rank}(M_0) \leq p$ and $p \leq r$. In particular, we do not assume that $\text{rank}(M_0) = \min\{p, r\}$ and allow the number of covariates to exceed the number of exogenous variables, i.e., $p > r$. Second, we allow the exogenous variable Z to directly affect the response variable Y (see Remark 4) and therefore f_0 does not need to coincide with the structural function from X to Y.

On the other hand, SIMDGs are more restrictive than common SIMs, as unlike model (13), we assume h is linear in Z, and we replace the assumptions on the conditional expectations of the exogenous variables (U, V, Z) with the assumptions in (3b). The independence assumption $Z \perp\!\!\!\perp (U, V)$ directly implies the two conditional expectation constraints in (13). We make this stronger assumption to allow for the factorization $(U, V, Z) \sim \Lambda_0 \otimes Q$. This ensures that the marginal distribution Q can be changed without affecting the distribution Λ_0.

A.3 SIMDGs as Structural Causal Models

The set of distributions $\mathcal{P}_0$ defined in (4) can also be written as a set of distributions induced by a class of SCMs. To see this, we consider the SCM over the variables $(X, Y, Z, H) \in \mathbb{R}^{p+1+r+q}$ given by

$$\begin{aligned} Z &:= \epsilon_Z \\ H &:= B_{HZ} Z + \epsilon_H \\ X &:= B_{XZ} Z + B_{XH} H + \epsilon_X \\ Y &:= g_0(X) + B_{YZ} Z + B_{YH} H + \epsilon_Y, \end{aligned} \tag{14}$$

where $\epsilon_Z, \epsilon_H, \epsilon_X, \epsilon_Y$ are jointly independent noise terms. Here, (X, Y, Z) are observed and H is unobserved. Moreover, $g_0 : \mathbb{R}^p \to \mathbb{R}$ denotes the causal function from X to Y and $B_{(\cdot,\cdot)}$ are linear maps of suitable sizes. The graph induced by (14) is given in Figure 4 (left). The SCM in (14) corresponds to the nonlinear anchor regression setup with a nonlinear causal function (Rothenhäusler et al., 2021; Bühlmann, 2020).

We now show how to rewrite the SCMs (14) as a model generated by a SIMDG $(f_0, M_0, \Lambda_0, \mathcal{Q}_Z)$, where $\mathcal{Q}_Z$ denotes the set of distributions induced by interventions on Z. To this end, define the matrix $M_0 := B_{XZ} + B_{XH} B_{HZ}$, the vector $\beta_0^\top := B_{YZ}$, and assume that the projectability condition (Rothenhäusler et al., 2021) holds, i.e., $B_{YZ}^\top \in \text{im}(M_0^\top)$. Define the two random variables

$$V := B_{XH} \epsilon_H + \epsilon_X \quad \text{and} \quad U := B_{YH} \epsilon_H + \epsilon_Y,$$

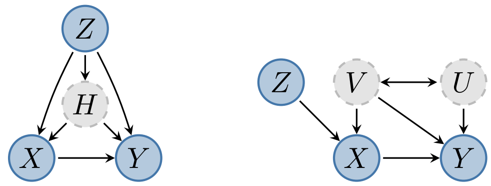

Figure 4: (left) Graph corresponding to the SCM defined in (14). (right) Graph corresponding to the SCM defined in (15). Both models induce the same observational distribution and the same interventional distribution for interventions on Z.

and note that $Z \perp\!\!\!\perp (U, V)$ since ϵ_Z is independent of $(\epsilon_H, \epsilon_X, \epsilon_Y)$. Moreover, substituting the structural equations (14) into each other yields the reduced SCM

$$\begin{aligned} Z &:= \epsilon_Z \\ X &:= M_0 Z + V \\ Y &:= \left(g_0(X) + \beta_0^\top M_0^\dagger X\right) + \left(U - \beta_0^\top M_0^\dagger V\right), \end{aligned} \tag{15}$$

where U and V are not necessarily independent of each other. Furthermore, denote by Λ_0 the distribution of the random vector $(U - \beta_0^\top M_0^\dagger V, V)$. The graph induced by (15) is given in Figure 4 (right), where bi-directed edges indicate dependence. By construction the induced observational distribution over (X, Y) is the same as in (14). Moreover, since ϵ_Z is independent of (U, V) also the interventional distributions over (X, Y) are the same as in (14) for any intervention on Z. Hence, we get that $(g_0 + \beta_0^\top M_0^\dagger, M_0, \Lambda_0, \mathcal{Q}_Z)$ is a SIMDG satisfying Definition 5. The reduced SCM (15) and the original SCM (14) only induce the same intervention distribution on (X, Y) for interventions on Z but generally not for interventions on other variables.

Appendix B. Proofs

B.1 Proof of Proposition 8

Proof Recall that $P_{\text{tr}} \in \mathcal{P}_0$ denotes the distribution of (U, V, X, Y, Z) induced by $\Lambda_0 \otimes Q_{\text{tr}}$, as detailed also in Appendix A.1. Due to this generative model we can express both X and Y as functions of U, V and Z. Moreover, for all measurable $A \subseteq \mathbb{R}$ it holds that

$$P_{\text{tr}}^{Y-f(X)}(A) := \int \mathbb{1}\left\{Y(u, v, z) - f\left(X(v, z)\right) \in A\right\} \mathrm{d}\Lambda_0(u, v) \mathrm{d}Q_{\text{tr}}(z).$$

Moreover, for all fixed $z_0 \in \mathbb{R}^r$ and all measurable sets $A \subseteq \mathbb{R}$, define

$$P_{z_0}^{Y-f(X)}(A) := \int \mathbb{1}\left\{Y(u, v, z_0) - f\left(X(v, z_0)\right) \in A\right\} \mathrm{d}\Lambda_0(u, v). \tag{16}$$

We first show that $f \in \mathcal{I}_0$ implies $Y - f(X) \perp\!\!\!\perp Z$ under P_{tr}. Let $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}^r$ be measurable sets. Then,

$$\begin{aligned}
P_{\text{tr}}^{Y-f(X),Z}(A \times B) &= \int \mathbb{1}\left\{Y(u,v,z) - f(X(v,z)) \in A, z \in B\right\} \mathrm{d}\Lambda_0(u,v)\, \mathrm{d}Q_{\text{tr}}(z) \\
&= \int_B \left[\int \mathbb{1}\left\{Y(u,v,z) - f(X(v,z)) \in A\right\} \mathrm{d}\Lambda_0(u,v)\right] \mathrm{d}Q_{\text{tr}}(z) \\
&\overset{(i)}{=} \int_B P_z^{Y-f(X)}(A)\, \mathrm{d}Q_{\text{tr}}(z) \overset{(ii)}{=} \int_B P_{\text{tr}}^{Y-f(X)}(A)\, \mathrm{d}Q_{\text{tr}}(z) \\
&= P_{\text{tr}}^{Y-f(X)}(A) \int_B \mathrm{d}Q_{\text{tr}}(z) = P_{\text{tr}}^{Y-f(X)}(A)\, Q_{\text{tr}}(B),
\end{aligned}$$

where (i) holds by equation (16), and (ii) holds by invariance and $\delta_z \in \mathcal{Q}_0$.

We now show that $Y - f(X) \perp\!\!\!\perp Z$ under P_{tr} implies $f \in \mathcal{I}_0$ if Assumption 1 holds. Fix measurable sets $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}^r$. Then, with the same argument as above, we get

$$P_{\text{tr}}^{Y-f(X),Z}(A \times B) = \int_B P_z^{Y-f(X)}(A)\, \mathrm{d}Q_{\text{tr}}(z). \tag{17}$$

Moreover, since $Y - f(X) \perp\!\!\!\perp Z$ under P_{tr}, it holds that

$$P_{\text{tr}}^{Y-f(X),Z}(A \times B) = P_{\text{tr}}^{Y-f(X)}(A)\, Q_{\text{tr}}(B). \tag{18}$$

From (17) and (18) it holds that

$$\int_B P_z^{Y-f(X)}(A) - P_{\text{tr}}^{Y-f(X)}(A)\, \mathrm{d}Q_{\text{tr}}(z) = 0.$$

Since B is arbitrary, it holds Q_{tr}-a.s. that

$$P_z^{Y-f(X)}(A) = P_{\text{tr}}^{Y-f(X)}(A). \tag{19}$$

Now, fix a distribution $Q_* \in \mathcal{Q}_0$ and consider $P_* \in \mathcal{P}_0$ induced by $\Lambda_0 \otimes Q_*$. From Assumption 1, it holds that $P_* \ll P_{\text{tr}}$, which implies that $Q_* \ll Q_{\text{tr}}$: For any measurable set $C \subseteq \mathbb{R}^r$, it holds that $Q_{\text{tr}}(C) = 0 \implies P_{\text{tr}}(\mathbb{R}^{2p+2} \times C) = 0 \implies P_*(\mathbb{R}^{2p+2} \times C) = 0 \implies Q_*(C) = 0$. Thus, equation (19) holds Q_*-a.s., too. Therefore,

$$P_*^{Y-f(X)}(A) = \int P_z^{Y-f(X)}(A)\, \mathrm{d}Q_*(z) = \int P_{\text{tr}}^{Y-f(X)}(A)\, \mathrm{d}Q_*(z) = P_{\text{tr}}^{Y-f(X)}(A).$$

Since $A \subseteq \mathbb{R}$ was arbitrary, it follows that $P_{\text{tr}}^{Y-f(X)} = P_*^{Y-f(X)}$. Since $P_* \in \mathcal{P}_0$ was arbitrary, it follows that $f \in \mathcal{I}$.

If Assumption 1 is not satisfied, Example 1 shows that $\mathcal{H}_1 \not\subseteq \mathcal{I}_0$ even under Assumption 2. ■

B.2 Proof of Proposition 11

Proof All equalities in this proof are meant to hold P_{tr}-a.s.

(“*only if*”) Suppose that $f_\star(X)$ is identifiable and that $h(X)+g(V)=0$. Fix $f(X) \coloneqq f_0(X)+h(X)$ and $\gamma(V) \coloneqq \gamma_0(V)+g(V)$. Clearly $f(X)+\gamma(V) = f_0(X)+\gamma_0(V)$, and so, see (8),

$$\mathrm{E}_{P_{\text{tr}}}[Y \mid X, V] = f(X) + \gamma(V). \tag{20}$$

We first consider the case $q = p$. Then, by Definition 9, the BCF is $f_\star = f_0$ and since $f_\star$ is identifiable (see Definition 10) (20) implies that $f_\star(X) = f(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma(V)]$. Hence, $h(X) = f(X) - f_0(X) = f_\star(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma(V)] - f_\star(X) = \mathrm{E}_{P_{\text{tr}}}[\gamma(V)]$ and so implication (11) holds.

Next, consider the case $q < p$. Then, again by Definition 9, the BCF is defined for P_{tr} all $x \in \mathbb{R}^p$ by $f_\star(x) \coloneqq f_0(x) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top X = R^\top x]$ and since $f_\star$ is identifiable (20) implies that $f_\star(X) = f(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma(V) \mid R^\top X]$. Therefore

$$h(X) = f(X) - f_0(X) = \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) - \gamma(V) \mid R^\top X] =: \delta(R^\top X),$$

and so implication (11) holds.

(“*if*”) Suppose that implication (11) holds and that $\mathrm{E}_{P_{\text{tr}}}[Y \mid X, V] = f(X) + \gamma(V)$ for arbitrary fixed functions $f, \gamma \in \mathcal{F}$. Since $\mathrm{E}_{P_{\text{tr}}}[Y \mid X, V] = f_0(V) + \gamma_0(V)$ and since the conditional expectation is unique P_{tr}-a.s., it follows that

$$f(X) + \gamma(V) = f_0(X) + \gamma_0(V). \tag{21}$$

Define $h(X) \coloneqq f(X) - f_0(X)$ and $g(V) \coloneqq \gamma(V) - \gamma_0(V)$. From (21) it follows that $h(X) + g(V) = 0$.

We first consider the case $q = p$. In this case, implication (11) implies that $h(X) = c$. Hence, $g(V) = -c$ and

$$f_\star(X) = f_0(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V)] = f(X) - c + \mathrm{E}_{P_{\text{tr}}}[\gamma(V)] + c,$$

which implies that $f_\star$ is identifiable.

Next, we consider the case $q < p$. In this case, the implication 11 implies that there exists a function $\delta : \mathbb{R}^{p-q} \to \mathbb{R}$ such that $h(X) = \delta(R^\top X)$. Since $h(X) + g(V) = 0$ and $h(X) = \delta(R^\top X)$, it follows that $g(V) = -\delta(R^\top X)$. Thus, by the way we defined h and g and the fact that $\delta(R^\top X)$ is $R^\top X$-measurable it follows

$$\begin{aligned} f(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma(V) \mid R^\top X] &= f_0(X) + h(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) + g(V) \mid R^\top X] \\ &= f_0(X) + \delta(R^\top X) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) - \delta(R^\top X) \mid R^\top X] \\ &= f_0(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top X] = f_\star(X). \end{aligned}$$

Thus, $f_\star$ identifiable. ∎

B.3 Proof of Proposition 12

Proof Unless otherwise stated, all equalities involving random variables hold P_{tr}-a.s.

First, Setting 1 ensures that M_0 is identified by

$$M_0 = \mathrm{E}_{P_{\text{tr}}}[XZ^\top]\{\mathrm{E}_{P_{\text{tr}}}[ZZ^\top]\}^{-1},$$

since $\mathrm{E}_{P_{\text{tr}}}[ZZ^\top] \succ 0$. Since X and Z are observed and M_0 is identified, the control variables $V = X - M_0 Z$ are also identified.

Suppose now that Assumption 3 holds and let $f : \mathbb{R}^p \to \mathbb{R}$ be measurable and $\gamma \in \mathbb{R}^p$ such that $\mathrm{E}_{P_{\text{tr}}}[Y \mid X, V] = f(X) + \gamma^\top V$. Since the conditional expectation is unique P_{tr}-a.s., it holds that $f(X) + \gamma^\top V = f_0(X) + \gamma_0^\top V$. Define $h : x \mapsto f(x) - f_0(x)$ so that

$$h(X) = (\gamma_0 - \gamma)^\top V = (\gamma_0 - \gamma)^\top (X - M_0 Z).$$

Define $\tilde{\mathcal{Z}} \coloneqq \{z_1, \tilde{z}_1, \dots, z_q, \tilde{z}_q\} \subseteq \mathcal{Z}$, $A \coloneqq \{(v, z) \in \mathbb{R}^{p+r} \mid h(v + M_0 z) - (\gamma_0 - \gamma)^\top v = 0\}$ and $A_z \coloneqq \{x \in \mathbb{R}^p \mid h(x) - (\gamma_0 - \gamma)^\top (x - M_0 z) = 0\}$ for all $z \in \mathcal{Z}$. First, note that $P_{\text{tr}}^{V,Z}(A) = 1$. Moreover, since

$$\sum_{z \in \mathcal{Z}} P_{\text{tr}}^{V + M_0 z}(A_z) P_{\text{tr}}^Z(\{z\}) = \sum_{z \in \mathcal{Z}} P_{\text{tr}}^V(A_z - M_0 z) P_{\text{tr}}^Z(\{z\}) = P_{\text{tr}}^{V,Z}(A) = 1,$$

and since $P^Z(\{z\}) > 0$ for all $z \in \mathcal{Z}$, it follows that $P^{V + M_0 z}(A_z) = 1$ for all $z \in \mathcal{Z}$.

Using Assumption 3, fix an arbitrary $j \in \{1, \dots, q\}$ and let $z_j, \tilde{z}_j \in \tilde{\mathcal{Z}}$. Denote by S_j and $\tilde{S}_j$ the support of $V + M_0 z_j$ and $V + M_0 \tilde{z}_j$, respectively and denote by S_V the support of V. Since the random variables $V + M_0 z_j$ and $V + M_0 \tilde{z}_j$ are not mutually singular, it holds for all Borel sets $E \subseteq \mathbb{R}^p$ that

$$\left(P_{\text{tr}}^{V + M_0 z_j}(E) = 1 \implies P_{\text{tr}}^{V + M_0 \tilde{z}_j}(E) > 0\right). \tag{22}$$

We now show that the set

$$A_j^* \coloneqq \left\{x \in \mathbb{R}^p \mid \exists\, v, \tilde{v} \in S_V \text{ s.t. } x = M_0 z_j + v = M_0 \tilde{z}_j + \tilde{v} \text{ and } x \in A_{z_j} \cap A_{\tilde{z}_j}\right\} \neq \varnothing. \tag{23}$$

Since $P_{\text{tr}}^{V + M_0 z_j}(A_{z_j}) = 1$ and $P_{\text{tr}}^{V + M_0 z_j}(S_j) = 1$ it follows that $P_{\text{tr}}^{V + M_0 z_j}(A_{z_j} \cap S_j) = 1$. Therefore, from (22), $P_{\text{tr}}^{V + M_0 \tilde{z}_j}(A_{z_j} \cap S_j) > 0$. By symmetry, it holds that $P_{\text{tr}}^{V + M_0 \tilde{z}_j}(A_{\tilde{z}_j} \cap \tilde{S}_j) = 1$ and $P_{\text{tr}}^{V + M_0 z_j}(A_{\tilde{z}_j} \cap \tilde{S}_j) > 0$. Fix the set $B_j \coloneqq S_j \cap \tilde{S}_j \cap A_{z_j} \cap A_{\tilde{z}_j}$ and notice that $P_{\text{tr}}^{V + M_0 z_j}(B_j) > 0$ and $P_{\text{tr}}^{V + M_0 \tilde{z}_j}(B_j) > 0$, which implies that $B_j \neq \varnothing$. We now show that $B_j \subseteq A_j^*$.

Fix $x \in B_j$ and note that $x \in S_j \cap \tilde{S}_j \cap A_{z_j} \cap A_{\tilde{z}_j}$. We want to show that there exist $v, \tilde{v} \in S_V$ such that $x = M_0 z_j + v$ and $x = M_0 \tilde{z}_j + \tilde{v}$. By definition of the support, $v \in S_V$ if and only if all open neighborhoods of v, $N_v \subseteq \mathbb{R}^p$, have positive probability, i.e., $P^V(N_v) > 0$. Fix $v = x - M_0 z_j$, and let N_v be a neighborhood of v. $P^V(N_v) = P^{V + M_0 z_j}(N_v + M_0 z_j) = P^{V + M_0 z_j}(N_x)$, where $N_x \coloneqq N_v + M_0 z_j \subseteq \mathbb{R}^p$ is an open neighborhood of x. Since $x \in S_j$, by definition of the support, it holds that $0 < P^{V + M_0 z_j}(N_x) = P^V(N_v)$. Hence, since N_v was arbitrary, we have shown that $v = x - M_0 z_j \in S_V$. By the same argument, there exists

$\tilde{v} \in S_V$ such that $x = M_0\tilde{z}_j + \tilde{v}$. Therefore, by the definition of A_j^*, it follows that $x \in A_j^*$, and thus $B_j \subseteq A_j^*$ and therefore (23) holds.

For all $j \in \{1, \dots, q\}$ fix $x \in A_j^*$, then by the definition of A_j^* it holds that

$$(\gamma_0 - \gamma)^\top (x - M_0 z_j) = h(x) = (\gamma_0 - \gamma)^\top (x - M_0 \tilde{z}_j),$$

which implies $(\gamma_0 - \gamma)^\top M_0(z_j - \tilde{z}_j) = 0$. Furthermore, from Assumption 3, we have $\text{span}(\{M_0(z_j - \tilde{z}_j) \mid j \in \{1, ..., q\}\}) = \text{im}(M_0)$, which implies that $\gamma_0 - \gamma \in \ker(M_0^\top)$.

First consider the case $q = p$. Since $\gamma_0 - \gamma \in \ker(M_0^\top)$ it holds that $\gamma_0 - \gamma = 0$, thus $h(X) = 0$ and hence $f(X) = f_0(X) = f_\star(X)$ which implies that $f_\star$ is identifiable from P_{tr}. Next, consider the case $q < p$. Then, $\gamma_0 - \gamma \in \ker(M_0^\top)$ implies that

$$h(X) = (\gamma_0 - \gamma)^\top V = (\gamma_0 - \gamma)^\top X.$$

Since $(\gamma_0 - \gamma)^\top V = (\gamma_0 - \gamma)^\top X$ is $R^\top X$-measurable, it follows that

$$\begin{aligned} f(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma^\top V \mid R^\top X] &= f_0(X) + h(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma^\top V \mid R^\top X] \\ &= f_0(X) + (\gamma_0 - \gamma)^\top X + \mathrm{E}_{P_{\text{tr}}}[\gamma^\top V \mid R^\top X] \\ &= f_0(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma^\top V + (\gamma_0 - \gamma)^\top V \mid R^\top X] \\ &= f_\star(X), \end{aligned}$$

and therefore $f_\star$ is identifiable from P_{tr}.

Suppose now that Assumption 4 holds and $f, \gamma : \mathbb{R}^p \to \mathbb{R}$ are differentiable functions such that $\mathrm{E}_{P_{\text{tr}}}[Y \mid X, V] = f(X) + \gamma(V)$. We adapt the proof of (Newey et al., 1999, Theorem 2.3). Denote the support of (V, Z) and its interior by $\text{supp}(P_{\text{tr}}^{V,Z})$ and $\text{supp}(P_{\text{tr}}^{V,Z})^\circ$, respectively. Since the conditional expectation is unique P_{tr}-a.s., it holds that $f(X)+\gamma(V) = f_0(X) + \gamma_0(V)$. Define $h : x \mapsto f(x) - f_0(x)$ and $g : v \mapsto \gamma(v) - \gamma_0(v)$ so that

$$h(X) + g(V) = h(M_0 Z + V) + g(V) = 0 \tag{24}$$

where we used the fact that $X = M_0 Z + V$. We now argue that (24) holds for all $(v, z) \in \text{supp}(P_{\text{tr}}^{V,Z})^\circ$. Define the set $A \coloneqq \{(v, z) \mid h(M_0 z + v) + g(v) = 0\}$, and note that $P_{\text{tr}}^{V,Z}(A) = 1$. By Assumption 4, the boundary of $\text{supp}(P_{\text{tr}}^{V,Z})$ has probability zero, and therefore $P_{\text{tr}}^{V,Z}(\text{supp}(P_{\text{tr}}^{V,Z})^\circ) = 1$. Therefore, $P_{\text{tr}}^{V,Z}(\text{supp}(P_{\text{tr}}^{V,Z})^\circ \cap A) = 1$, i.e., equality (24) holds for all $(v, z) \in \text{supp}(P_{\text{tr}}^{V,Z})^\circ$.

Since h and g are differentiable, differentiating (24) with respect to z for all $(v, z) \in \text{supp}(P_{\text{tr}}^{V,Z})^\circ$ yields

$$M_0^\top \nabla h(M_0 z + v) = 0, \text{ for all } (v, z) \in \text{supp}(P_{\text{tr}}^{V,Z})^\circ, \tag{25}$$

where $\nabla h : \mathbb{R}^p \to \mathbb{R}^p$ is the gradient of h.

Define $S_X \coloneqq \text{supp}(P_{\text{tr}}^X)$, $S_V \coloneqq \text{supp}(P_{\text{tr}}^V)$ and $S_Z \coloneqq \text{supp}(P_{\text{tr}}^Z)$ and note that S_V° and S_Z° are convex since (V, Z) are independent under P_{tr} and since $P_{\text{tr}}^{V,Z}$ has convex support by assumption (to see this, use $\text{supp}(P_{\text{tr}}^{V,Z})^\circ = S_V^\circ \times S_Z^\circ$ and that the projection of a convex set is convex). Furthermore, it holds that

$$S_X^\circ \subseteq M_0 S_Z^\circ + S_V^\circ. \tag{26}$$

To see this, assume that $S_X^\circ \not\subseteq M_0 S_Z^\circ + S_V^\circ$. Therefore, there exists a set $A \subseteq (M_0 S_Z^\circ + S_V^\circ)^c$ such that $A \neq \varnothing$ and A is open (since S_X° is open). Fix $x \in A$. Since A is open and since $A \subseteq S_X^\circ$, using the definitions of open set and support, there exists an open neighborhood $N_x \subseteq A$ such that $x \in N_x$ and $P_{\text{tr}}^X(N_x) > 0$. Moreover, notice that

$$P_{\text{tr}}^X(M_0 S_Z^\circ + S_V^\circ) = \int_{z \in S_Z^\circ} P_{\text{tr}}^{V+M_0 z}(S_V^\circ + M_0 z) \mathrm{d}P_{\text{tr}}^Z(z) = \int_{z \in S_Z^\circ} P_{\text{tr}}^V(S_V^\circ) \mathrm{d}P_{\text{tr}}^Z(z) = 1.$$

Therefore, since $N_x \subseteq A \subseteq (M_0 S_Z^\circ + S_V^\circ)^c$ it holds that $P_{\text{tr}}^X(N_x) \leq P_{\text{tr}}^X((M_0 S_Z^\circ + S_V^\circ)^c) = 0$, which is a contradiction.

We now show (by contradiction) that for all $v \in S_V^\circ$ the function $z \mapsto h(M_0 z + v) \in \mathbb{R}$ is constant on S_Z°. Fix $v_0 \in S_V^\circ$ and suppose there exist $\bar{z}, \tilde{z} \in S_Z^\circ$ such that $h(M_0\bar{z} + v_0) \neq h(M_0\tilde{z} + v_0)$. Define the function $\ell : [0,1] \to \mathbb{R}$ for all $t \in [0,1]$ by

$$\ell(t) := h\left(M_0 \left\{\bar{z} + t(\tilde{z} - \bar{z})\right\} + v_0\right).$$

Notice that $\ell(0) = h(M_0\bar{z} + v_0) \neq h(M_0\tilde{z} + v_0) = \ell(1)$. Therefore, using convexity of the support S_Z°, by the mean value theorem there exists a $c \in (0,1)$ such that $\ell^{'}(c) = \ell(1) - \ell(0) \neq 0$. By the definition of the function ℓ, the chain rule, and (25), it follows that

$$\ell^{'}(c) = (\bar{z} - \tilde{z})^\top M_0^\top \nabla h\left(M_0 \left\{\bar{z} + c(\tilde{z} - \bar{z})\right\} + v_0\right) = 0,$$

which is a contradiction. Therefore, for all $v \in S_V^\circ$ the function $z \mapsto h(M_0 z + v) \in \mathbb{R}$ is constant on S_Z, that is, for all $v \in S_V$ and for all $z_1, z_2 \in S_Z$ it holds that

$$h(M_0 z_1 + v) = h(M_0 z_2 + v). \tag{27}$$

Using this we will now show that for all $x, \tilde{x} \in S_X^\circ$ the following implication holds

$$\text{there exists } u \in \mathbb{R}^r \text{ such that } x - \tilde{x} = M_0 u \implies h(x) = h(\tilde{x}). \tag{28}$$

To see this, fix arbitrary $x, \tilde{x} \in S_X^\circ$ such that there exists $u \in \mathbb{R}^r$ satisfying $x - \tilde{x} = M_0 u$. Using (26), let $v, \tilde{v} \in S_V^\circ$ and $z, \tilde{z} \in S_Z^\circ$ be such that $x = M_0 z + v$ and $\tilde{x} = M_0\tilde{z} + \tilde{v}$. Next, define the path $\gamma : [0,1] \to \text{Im}(M_0)$ for all $t \in [0,1]$ by

$$\gamma(t) := t(v - \tilde{v}).$$

This is well-defined since $x - \tilde{x} = M_0(z - \tilde{z}) + (v - \tilde{v}) \in \text{Im}(M_0)$. Next, fix $\eta > 0$ such that

$$\{\bar{z} \in \mathbb{R}^r \mid \|M_0\tilde{z} - M_0\bar{z}\|_2 < \eta\} \subseteq S_Z^\circ. \tag{29}$$

Then, by uniform continuity of γ there exists $K \in \mathbb{N}$ such that for all $t \in [0, 1 - \frac{1}{K}]$ it holds that $\|\gamma(t + \frac{1}{K}) - \gamma(t)\|_2 < \eta$. Now for all $k \in \{0, \ldots, K\}$ define $v_k := v - \gamma(\frac{k}{K})$, which is in S_V° since S_V° is convex. Then, for all $k \in \{0, \ldots, K-1\}$ it holds by construction that $\|M_0\tilde{z} - (M_0\tilde{z} + v_k - v_{k+1})\|_2 = \|v_{k+1} - v_k\|_2 < \eta$. Since $v_k - v_{k+1} = \gamma(\frac{1}{K}) \in \text{im}(M_0)$ it holds that $M_0\tilde{z} + v_k - v_{k+1} \in \text{im}(M_0)$, i.e., there exists $z_k \in \mathbb{R}^r$ such that

$$M_0\tilde{z} + v_k - v_{k+1} = M_0 z_k. \tag{30}$$

Hence, by (29) $z_k \in S_Z^\circ$. Finally, we can use (27) (indicated by $\star$) and (30) (indicated by $\dagger$) to get that

$$\begin{aligned} h(x) &\overset{\star}{=} h(M_0\tilde{z} + v_0) \overset{\dagger}{=} h(M_0 z_0 + v_1) \\ &\overset{\star}{=} h(M_0\tilde{z} + v_1) \overset{\dagger}{=} h(M_0 z_1 + v_2) \\ &\dots \\ &\overset{\star}{=} h(M_0\tilde{z} + v_k) \overset{\dagger}{=} h(M_0 z_k + v_{k+1}) \\ &\dots \\ &\overset{\star}{=} h(M_0\tilde{z} + v_{K-1}) \overset{\dagger}{=} h(M_0 z_{K-1} + v_K) \\ &\overset{\star}{=} h(M_0\tilde{z} + v_K) = h(\tilde{x}), \end{aligned}$$

which proves (28).

We first consider the case $q = p$. In this case, (28) implies that h is constant on all of S_X°. Hence, by Proposition 11, $f_\star$ is identifiable from P_{tr}. Next, consider the case $q < p$. Define the function $\delta : \mathbb{R}^{p-q} \to \mathbb{R}$ such that for all $x \in S_X^\circ$ it holds that

$$\delta(R^\top x) := h(x).$$

The function $\delta : \mathbb{R}^{p-q} \to \mathbb{R}$ is well-defined: Take $x, \tilde{x} \in S_X^\circ$ such that $R^\top x = R^\top \tilde{x}$. Then, $R^\top(x - \tilde{x}) = 0$ which implies that $x - \tilde{x} = M_0 b$ for some $b \in \mathbb{R}^r$. Therefore, from (28) it follows that

$$\delta(R^\top x) = h(x) = h(\tilde{x}) = \delta(R^\top \tilde{x}).$$

Hence, by Proposition 11, $f_\star$ is identifiable from P_{tr}. ■

B.4 Proof of Proposition 13

Proof Recall that $P_{\text{tr}} \in \mathcal{P}_0$ denotes the distribution over (U, V, X, Y, Z) induced by $\Lambda_0 \otimes Q_{\text{tr}}$. From (3a) and the definition of the BCF we get for $q = p$ that

$$Y - f_\star(X) = f_0(X) + U - f_0(X) = U,$$

and similarly for $q < p$ that

$$Y - f_\star(X) = f_0(X) + U - f_0(X) - \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top X] = U - \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top V],$$

where in the last step we used that $R^\top X = R^\top M_0 Z + R^\top V = R^\top V$. In both cases, the residuals $Y - f_\star(X)$ only depend on (U, V) which have the same marginal distribution for all $P \in \mathcal{P}_0$ (see Appendix A.1 for details on the generative model). Therefore, it holds that for all $P \in \mathcal{P}_0$

$$P_{\text{tr}}^{Y - f_\star(X)} = P^{Y - f_\star(X)}$$

and hence $f_\star$ is invariant, i.e., $f_\star \in \mathcal{I}_0$.

We now show how the risk of $f_\star$ relates to the risk of the least squares predictor f_{LS}. First, for the case $q = p$, we get that

$$
\begin{aligned}
\mathcal{R}(P_{\mathrm{tr}}, f_\star) &= \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - f_\star(X)\right)^2\right] \\
&= \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - f_0(X)\right)^2\right] \\
&= \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - (f_0(X) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X]) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X] - \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V)]\right)^2\right] \\
&= \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - f_{\mathrm{LS}}(X)\right)^2\right] + \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(\mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X] - \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V)]\right)^2\right] \\
&= \mathcal{R}(P_{\mathrm{tr}}, f_{\mathrm{LS}}) + \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(\mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X] - \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V)]\right)^2\right],
\end{aligned}
$$

where in the third equality we used that $\mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V)] = 0$ and in the fourth equality that the cross terms are zero. Similarly, in the case $q < p$, we get

$$
\begin{aligned}
\mathcal{R}(P_{\mathrm{tr}}, f_\star) &= \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - f_\star(X)\right)^2\right] \\
&= \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - f_0(X) - \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid R^\top X]\right)^2\right] \\
&= \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - \left\{f_0(X) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X]\right\} + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X] - \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid R^\top X]\right)^2\right] \\
&\overset{\blacktriangle}{=} \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(Y - f_{\mathrm{LS}}(X)\right)^2\right] + \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(\mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X] - \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid R^\top X]\right)^2\right] \\
&= \mathcal{R}(P_{\mathrm{tr}}, f_{\mathrm{LS}}) + \mathrm{E}_{P_{\mathrm{tr}}}\left[\left(\mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid X] - \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid R^\top X]\right)^2\right].
\end{aligned}
$$

The cross term in $\blacktriangle$ vanishes by the properties of the conditional expectation $f_{\mathrm{LS}}(X) = \mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X]$ (specifically, the property that for all measurable functions g it holds that $\mathrm{E}_{P_{\mathrm{tr}}}[g(X)(Y - \mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X])] = 0$). ■

B.5 Proof of Proposition 14

Proof We want to show that

$$
\inf_{P \in \mathcal{P}_0} \mathrm{E}_P\left[\left(\mathrm{E}_P[\gamma_0(V) \mid R^\top X] - \mathrm{E}_P[\gamma_0(V) \mid X]\right)^2\right] = 0. \tag{31}
$$

By definition of the infimum, it is equivalent to show that for any $\varepsilon > 0$ there exists a $P^* \in \mathcal{P}_0$ such that

$$
\mathrm{E}_{P^*}\left[\left(\mathrm{E}_{P^*}[\gamma_0(V) \mid R^\top X] - \mathrm{E}_{P^*}[\gamma_0(V) \mid X]\right)^2\right] < \varepsilon.
$$

Let $Q_k := N(0, k^2 I_r) \in \mathcal{Q}_0$ and denote by $(U_k, V_k, X_k, Y_k, Z_k)$ the random vector and by $P_k \in \mathcal{P}_0$ the distribution induced by $(f_0, M_0, \Lambda_0, Q_k)$. Moreover, denote by (U, V, X, Y, Z) the random vector generated by P_{tr} (see Setting 1). Throughout, this proof we will always add subscripts to the random variables to increase clarity (see Appendix A.1 for details on this). We prove the statement for each of the Conditions (a) and (b) separately.

(a) Suppose Λ_0 satisfies that for $(U,V) \sim \Lambda_0$ it holds that V has a density w.r.t. Lebesgue and $\gamma_0(V) = \mathrm{E}_{P_{\mathrm{tr}}}[U \mid V]$ is almost surely bounded.

Since $\mathrm{im}(R) = \ker(M_0^\top)$, it holds for all $k \in \mathbb{N}$ that $R^\top X_k = R^\top V_k$ and hence it holds P_k-a.s. that

$$\mathrm{E}_{P_k}[\gamma_0(V_k) \mid R^\top X_k] = \mathrm{E}_{P_k}[\gamma_0(V_k) \mid R^\top V_k].$$

Therefore, we get that

$$\begin{aligned}
&\mathrm{E}_{P_k}\left[\left(\mathrm{E}_{P_k}[\gamma_0(V_k) \mid R^\top X_k] - \mathrm{E}_{P_k}[\gamma_0(V_k) \mid X_k]\right)^2\right] \\
&= \mathrm{E}_{P_k}\left[\left(\mathrm{E}_{P_k}[\gamma_0(V_k) \mid R^\top V_k] - \mathrm{E}_{P_k}[\gamma_0(V_k) \mid X_k]\right)^2\right].
\end{aligned}$$

By assumption, the control function $\gamma_0(V_k)$ is bounded, so there exists $c > 0$ such that $|\gamma_0(V_k)| < c$ P_k-a.s., and so it holds P_k-a.s. for all $k \in \mathbb{N}$ that $\left|\mathrm{E}_{P_k}[\gamma_0(V_k) \mid X_k]\right| \leq c$. Therefore, by Lemma 17, it holds that

$$\lim_{k\to\infty} \mathrm{E}_{P_k}\left[\left(\mathrm{E}_{P_k}[\gamma_0(V_k) \mid R^\top V_k] - \mathrm{E}_{P_k}[\gamma_0(V_k) \mid X_k]\right)^2\right] = 0. \tag{32}$$

Fix $\varepsilon > 0$. By (32), there exists a $k^* > 0$ such that

$$\mathrm{E}_{P_{k^*}}\left[\left(\mathrm{E}_{P_{k^*}}[\gamma_0(V_{k^*}) \mid R^\top V_{k^*}] - \mathrm{E}_{P_{k^*}}[\gamma_0(V_{k^*}) \mid X_{k^*}]\right)^2\right] < \varepsilon,$$

which proves the claim.

(b) Suppose Λ_0 is a multivariate centered (non-degenerate) Gaussian distribution and denote by Σ the covariance matrix corresponding to the marginal distribution of V. Since $(U,V) \sim \Lambda_0$ are jointly Gaussian, the control function $\gamma_0(V) = \mathrm{E}_{P_{\mathrm{tr}}}[U \mid V]$ is linear in V so that $\gamma_0(V) = \gamma_0^\top V$, for some fixed vector $\gamma_0 \in \mathbb{R}^p$. For all $k \in \mathbb{N}$ define the matrix $B_k := k^2 M_0 M_0^\top + \Sigma$. By construction of P_k and $(U_k, V_k, X_k, Y_k, Z_k)$ it holds that $Z_k \sim N(0, k^2 I_r)$, $X_k = M_0 Z_k + V_k \sim N(0, B_k)$ and $\mathrm{E}_{P_k}[X_k V_k^\top] = \Sigma$. Using the joint Gaussianity of (X_k, V_k), we now show how to rewrite the conditional expectations in (31) for P_k. Since $R^\top X_k = R^\top V_k$, it holds P_k-a.s.

$$\begin{aligned}
\mathrm{E}_{P_k}[V_k \mid R^\top X_k] = \mathrm{E}_{P_k}[V_k \mid R^\top V_k] &= \mathrm{E}_{P_k}[V_k V_k^\top R]\mathrm{E}_{P_k}[R^\top V_k V_k^\top R]^{-1} R^\top V_k \\
&= \Sigma R (R^\top \Sigma R)^{-1} R^\top V_k = \Sigma \Pi V_k,
\end{aligned} \tag{33}$$

where we defined $\Pi := R(R^\top \Sigma R)^{-1} R^\top$. Furthermore, it holds P_k-a.s.

$$\mathrm{E}_{P_k}[V_k \mid X_k] = \mathrm{E}_{P_k}[V_k X_k^\top]\mathrm{E}_{P_k}[X_k X_k^\top]^{-1} X_k = \Sigma B_k^{-1} X_k. \tag{34}$$

Therefore, using (33) and (34), it holds

$$\begin{aligned}
&\lim_{k\to\infty} \mathrm{E}_{P_k}\left[\left(\mathrm{E}_{P_k}[\gamma_0^\top V_k \mid R^\top X_k] - \mathrm{E}_{P_k}[\gamma_0^\top V_k \mid X_k]\right)^2\right] \\
&= \lim_{k\to\infty} \gamma_0^\top \mathrm{E}_{P_k}\left[\left(\Sigma\Pi V_k - \Sigma B_k^{-1} X_k\right)\left(V_k^\top \Pi\Sigma - X_k^\top B_k^{-1}\Sigma\right)\right]\gamma_0 \\
&\overset{\diamondsuit}{=} \lim_{k\to\infty} \gamma_0^\top \left[\Sigma\Pi\Sigma + \Sigma B_k^{-1}\Sigma - \Sigma\Pi\Sigma B_k^{-1}\Sigma - \Sigma B_k^{-1}\Sigma\Pi\Sigma\right]\gamma_0 \\
&\overset{\clubsuit}{=} \gamma_0^\top \left[\Sigma\left(\Pi + \Pi - \Pi\Sigma\Pi - \Pi\Sigma\Pi\right)\Sigma\right]\gamma_0 \overset{\diamondsuit}{=} 0,
\end{aligned} \tag{35}$$

where in $\diamondsuit$ we used that $\Pi\Sigma\Pi = \Pi$ and in $\clubsuit$ we used Lemma 18 which states that $\lim_{k\to\infty} B_k^{-1} = \Pi$. Fix $\varepsilon > 0$. By (35), there exists a $k^* > 0$ such that

$$\mathrm{E}_{P_{k^*}}\left[\left(\mathrm{E}_{P_{k^*}}[\gamma_0^\top V_{k^*} \mid R^\top X_{k^*}] - \mathrm{E}_{P_{k^*}}[\gamma_0^\top V_{k^*} \mid X_{k^*}]\right)^2\right] < \varepsilon,$$

which proves the claim.

■

Lemma 17 *Assume Setting 1 and suppose that*

$$\{N(0, k^2 I_r) \mid k \in \mathbb{N}\} \subseteq \mathcal{Q}_0.$$

Define $\gamma_0(V) := \mathrm{E}_{P_{\mathrm{tr}}}[U \mid V]$ *and assume it is almost surely bounded, and define* R *as in* (9)*. For all* $k \in \mathbb{N}$ *let* $Q_k := N(0, k^2 I_r) \in \mathcal{Q}_0$ *and* $P_k \in \mathcal{P}_0$ *the distribution induced by* $(f_0, M_0, \Lambda_0, Q_k)$*. Furthermore, make the following additional assumptions:*

(i) for $(U, V) \sim \Lambda_0$*, the random variable* V *has a density w.r.t. the Lebesgue measure,*

(ii) there exists a constant $c > 0$ *such that* P_1*-a.s. for all* $k \in \mathbb{N}$ *it holds that*

$$\left|\mathrm{E}_{P_1}[\gamma_0(V) \mid kM_0 Z + V]\right| \leq c.$$

Then, it holds that

$$\lim_{k\to\infty} \mathrm{E}_{P_k}\left[\left(\mathrm{E}_{P_k}[\gamma_0(V) \mid R^\top V] - \mathrm{E}_{P_k}[\gamma_0(V) \mid X_k]\right)^2\right] = 0.$$

Proof Let (U, V, X, Y, Z) be the random vector generated by P_1 (details on how the random variables are generated are given in Appendix A.1). Moreover, fix $k \in \mathbb{N}$ and note that by construction it holds that $(U, V, X_k, Y_k, Z) \sim P_k$.

Let $S = (s_1, \ldots, s_q) \in \mathbb{R}^{p\times q}$ with $q = \mathrm{rank}(M_0)$ denote the left singular vectors of M_0 associated to non-zero singular values – this implies that S is an orthonormal basis for $\mathrm{im}(M_0)$. Moreover, define the transformed random variables $\tilde{X}_k := (S, R)^\top X_k$, $\tilde{Z} := S^\top M_0 Z$ and $\tilde{V} = (\tilde{V}_S, \tilde{V}_R) := (S^\top V, R^\top V)$. It then holds that,

$$\tilde{X}_k = k\begin{bmatrix} S^\top M_0 Z \\ R^\top M_0 Z \end{bmatrix} + \begin{bmatrix} S^\top V \\ R^\top V \end{bmatrix} = k\begin{bmatrix} \tilde{Z} \\ 0 \end{bmatrix} + \begin{bmatrix} \tilde{V}_S \\ \tilde{V}_R \end{bmatrix}.$$

Let $\tilde{\gamma}_0 := \gamma_0 \circ (S, R)$ and rewrite the conditional expectations, using $\gamma_0(V) = \tilde{\gamma}_0(\tilde{V})$ and that (S, R) are an orthonormal basis, as

$$\mathrm{E}_{P_k}[\gamma_0(V) \mid X_k] = \mathrm{E}_{P_k}[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{X}_k] = \mathrm{E}_{P_1}[\tilde{\gamma}_0(\tilde{V}) \mid k\tilde{Z} + \tilde{V}_S, \tilde{V}_R], \tag{36}$$

$$\mathrm{E}_{P_k}[\gamma_0(V) \mid R^\top X_k] = \mathrm{E}_{P_1}[\tilde{\gamma}_0(\tilde{V}) \mid R^\top V] = \mathrm{E}_{P_1}[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{V}_R]. \tag{37}$$

Using the trigonometric identities $\sin(\arctan(k)) = \frac{k}{\sqrt{1+k^2}}$ and $\cos(\arctan(k)) = \frac{1}{\sqrt{1+k^2}}$, we have the following equalities of sigma-algebras,

$$\sigma(k\tilde{Z} + \tilde{V}_S) = \sigma\left(\frac{k\tilde{Z} + \tilde{V}_S}{\sqrt{1+k^2}}\right) = \sigma\left(\tilde{Z}\sin\theta_k + \tilde{V}_S\cos\theta_k\right),$$

where $\theta_k = \arctan(k) \in (0, \pi/2)$. So we can rewrite the r.h.s. of (36) as

$$\mathrm{E}_{P_1}[\tilde{\gamma}_0(\tilde{V}) \mid k\tilde{Z} + \tilde{V}_S, \tilde{V}_R] = \mathrm{E}_{P_1}[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{Z}\sin\theta_k + \tilde{V}_S\cos\theta_k, \tilde{V}_R]. \tag{38}$$

Since $k \in \mathbb{N}$ was fixed arbitrarily, (38) holds for all $k \in \mathbb{N}$. Next, fix an arbitrary $\theta \in (0, \pi/2)$ and define the random variables $W^1 = (W_S^1, W_R^1) \in \mathbb{R}^p$ and $W^2 = (W_S^2, W_R^2) \in \mathbb{R}^p$ by

$$\begin{pmatrix} W_S^1 & W_S^2 \\ W_R^1 & W_R^2 \end{pmatrix} := \begin{pmatrix} \tilde{Z} & \tilde{V}_S \\ 0 & \tilde{V}_R \end{pmatrix} \begin{pmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{pmatrix}.$$

This implies that

$$\begin{pmatrix} \tilde{Z} & \tilde{V}_S \\ 0 & \tilde{V}_R \end{pmatrix} = \begin{pmatrix} W_S^1 & W_S^2 \\ W_R^1 & W_R^2 \end{pmatrix} \begin{pmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{pmatrix}.$$

By construction and the fact that $\mathrm{supp}(P_1^{M_0 Z}) = \mathrm{Im}(M_0)$ (since P_1^Z has full support) the support of $W_S^1 = \tilde{Z}\cos(\theta) + \tilde{V}_S\sin(\theta)$ satisfies $\mathrm{supp}(P_1^{W_S^1}) = \mathrm{Im}(S^\top)$ and hence does not depend on θ. By assumption (i), we know that V has a density w.r.t. Lebesgue which implies that also the transformed variable $\tilde{V}$ has a density w.r.t. Lebesgue, which we denote by $f_{\tilde{V}} : \mathbb{R}^p \to \mathbb{R}$. Moreover, $\tilde{Z} \sim N(0, M_0^\top M_0)$, hence it has a density w.r.t. Lebesgue $f_{\tilde{Z}}$ which satisfies $\sup_{z \in \mathbb{R}^r} f_{\tilde{Z}}(z) < \infty$. By density transformation, and by the independence of $\tilde{Z}$ and $\tilde{V}$, (W^1, W^2) has density

$$\begin{aligned} f_{W^1,W^2}(w_S^1, w_R^1, w_S^2, w_R^2) &= f_{\tilde{Z},\tilde{V}}(w_S^1\sin\theta - w_S^2\cos\theta, w^1\cos\theta + w^2\sin\theta) \\ &= f_{\tilde{Z}}(w_S^1\sin\theta - w_S^2\cos\theta)\ f_{\tilde{V}}(w^1\cos\theta + w^2\sin\theta), \end{aligned}$$

where $w^1 = (w_S^1, w_R^1)$ and $w^2 = (w_S^2, w_R^2)$. For all $w_S^1 \in \mathrm{supp}(P_1^{W_S^1})$ and all $v_R \in \mathrm{supp}(P_1^{V_R})$ it holds that

$$\begin{aligned} &\mathrm{E}_{P_1}\left[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{Z}\sin\theta + \tilde{V}_S\cos\theta = w_S^1, \tilde{V}_R = v_R\right] \\ &= \mathrm{E}_{P_1}\left[\tilde{\gamma}_0(W^1\cos\theta + W^2\sin\theta) \mid W_S^1 = w_S^1, W_R^1\cos\theta + W_R^2\sin\theta = v_R\right] \\ &= \frac{\int \tilde{\gamma}_0(w_S^1\cos\theta + w_S^2\sin\theta, v_R)\ f_{\tilde{Z}}(w_S^1\sin\theta - w_S^2\cos\theta)\ f_{\tilde{V}}(w_S^1\cos\theta + w_S^2\sin\theta, v_R)\ \mathrm{d}w_S^2}{\int f_{\tilde{Z}}(w_S^1\sin\theta - w_S^2\cos\theta)\ f_{\tilde{V}}(w_S^1\cos\theta + w_S^2\sin\theta, v_R)\ \mathrm{d}w_S^2} \\ &= \frac{\int \tilde{\gamma}_0(t, v_R)\ f_{\tilde{Z}}(w_S^1\sin\theta - h(t,\theta)\cos\theta)\ f_{\tilde{V}}(t, v_R)\ \mathrm{d}t}{\int f_{\tilde{Z}}(w_S^1\sin\theta - h(t,\theta)\cos\theta)\ f_{\tilde{V}}(t, v_R)\ \mathrm{d}t}, \end{aligned} \tag{39}$$

where in the last equality we substituted $t := w_S^1\cos\theta + w_S^2\sin\theta$ and $h(t,\theta) := (t - w_S^1\cos\theta)/\sin\theta$. Notice that $\lim_{\theta\uparrow\pi/2} h(t,\theta) = t$ and $\lim_{\theta\uparrow\pi/2} f_{\tilde{Z}}(w_S^1\sin\theta - h(t,\theta)\cos\theta) =$

$f_{\tilde{Z}}(w_S^1)$ by continuity of $f_{\tilde{Z}}$. Since $\theta \in (0, \pi/2)$ was arbitrary, (39) implies for all $w_S^1 \in \operatorname{supp}(P_1^{W_S^1})$ and all $v_R \in \operatorname{supp}(P_1^{V_R})$ (both $\operatorname{supp}(P_1^{W_S^1})$ and $\operatorname{supp}(P_1^{V_R})$ do not depend on θ) that

$$\begin{aligned}
&\lim_{\theta \uparrow \pi/2} \mathrm{E}_{P_1}\left[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{Z}\sin\theta + \tilde{V}_S \cos\theta = w_S^1, \tilde{V}_R = v_R\right] \\
&= \lim_{\theta \uparrow \pi/2} \frac{\int \tilde{\gamma}_0(t, v_R)\ f_{\tilde{Z}}(w_S^1 \sin\theta - h(t,\theta)\cos\theta)\ f_{\tilde{V}}(t, v_R)\ \mathrm{d}t}{\int f_{\tilde{Z}}(w_S^1 \sin\theta - h(t,\theta)\cos\theta)\ f_{\tilde{V}}(t, v_R)\ \mathrm{d}t} \\
&= \frac{f_{\tilde{Z}}(w_S^1) \int \tilde{\gamma}_0(t, v_R)\ f_{\tilde{V}}(t, v_R)\ \mathrm{d}t}{f_{\tilde{Z}}(w_S^1) \int f_{\tilde{V}}(t, v_R)\ \mathrm{d}t} \\
&= \mathrm{E}_{P_1}[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{V}_R = v_R],
\end{aligned} \tag{40}$$

where in the second equality we invoked the dominated convergence theorem since $f_{\tilde{Z}}$ is bounded and $\tilde{\gamma}_0$ is bounded.

Now, fix $\varepsilon > 0$. By (40), there exists $k^* \in \mathbb{N}$ such that $\theta^* := \arctan(k^*) \in (0, \pi/2)$ such that for all $\theta \in [\theta^*, \pi/2)$ it holds P_1-a.s. that

$$\left| \mathrm{E}_{P_1}[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{V}_R] - \mathrm{E}_{P_1}\left[\tilde{\gamma}_0(\tilde{V}) \mid \tilde{Z}\sin\theta + \tilde{V}_S\cos\theta, \tilde{V}_R\right] \right| < \varepsilon.$$

Moreover, by (38), for all $k > k^*$ this implies P_1-a.s. that

$$\left| \mathrm{E}_{P_1}[\gamma_0(V) \mid R^\top V] - \mathrm{E}_{P_1}\left[\gamma_0(V) \mid kM_0 Z + V\right] \right| < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, it follows P_1-a.s.

$$\lim_{k \to \infty} \mathrm{E}_{P_1}\left[\gamma_0(V) \mid kM_0 Z + V\right] = \mathrm{E}_{P_1}[\gamma_0(V) \mid R^\top V]. \tag{41}$$

Finally, by assumption (ii), there exists a constant $c > 0$ such that P_1-a.s. for all $k \in \mathbb{N}$ it holds that

$$\left| \mathrm{E}_{P_1}[\gamma_0(V) \mid kM_0 Z + V] \right| \leq c. \tag{42}$$

Therefore, by (41) and (42), using (Chung, 2001, Theorem 4.1.4) it holds that

$$\lim_{k \to \infty} \mathrm{E}_{P_1}\left[\left(\mathrm{E}_{P_1}[\gamma_0(V) \mid R^\top V] - \mathrm{E}_{P_1}[\gamma_0(V) \mid kM_0 Z + V]\right)^2\right] = 0.$$

Using that $(U, V, kZ, X_k, Y_k) \sim P_k$ and $X_k = kM_0 Z + V$, this in particular implies that

$$\lim_{k \to \infty} \mathrm{E}_{P_k}\left[\left(\mathrm{E}_{P_k}[\gamma_0(V) \mid R^\top V] - \mathrm{E}_{P_k}[\gamma_0(V) \mid X_k]\right)^2\right] = 0.$$

■

Lemma 18 *For each $k \in \mathbb{N}$, define $B_k := k^2 M_0 M_0^\top + \Sigma$, where $M_0 \in \mathbb{R}^{p \times r}$ with $\operatorname{rank}(M_0) = q$ and $\Sigma \in \mathbb{R}^{p \times p}$ is positive definite. Define $R \in \mathbb{R}^{p \times (p-q)}$ as in (9). Then, as $k \to \infty$, it holds that $B_k^{-1} \to R(R^\top \Sigma R)^{-1} R^\top$.*

Proof Let $S \in \mathbb{R}^{p\times q}$ denote the left singular vectors of M_0 associated to non-zero singular values – this implies that S is an orthonormal basis for $\text{im}(M_0)$ and $S^\top M_0 M_0^\top S = L \in \mathbb{R}^{q\times q}$ is a positive definite diagonal matrix. Define $Q = [S, R] \in \mathbb{R}^{p\times p}$ so that $QQ^\top = I_p$. Then, $B_k^{-1} = (QQ^\top B_k QQ^\top)^{-1} = Q(Q^\top B_k Q)^{-1}Q^\top$, and so

$$Q^\top B_k Q = \begin{bmatrix} k^2 L + S^\top \Sigma S & S^\top \Sigma R \\ (S^\top \Sigma R)^\top & R^\top \Sigma R \end{bmatrix} =: \begin{bmatrix} A & B \\ B^\top & D \end{bmatrix} =: K.$$

Using the block matrix inversion formula and the Schur complement $(K/D) := A - BD^{-1}B^\top$, we get

$$\begin{aligned}(Q^\top B_k Q)^{-1} &= \begin{bmatrix} (K/D)^{-1} & -(K/D)^{-1}BD^{-1} \\ -D^{-1}B^\top (K/D)^{-1} & D^{-1} + D^{-1}B^\top (K/D)^{-1} BD^{-1} \end{bmatrix}, \text{ and} \qquad (43)\\ (K/D)^{-1} &= \left(k^2 L + S^\top \Sigma S - S^\top \Sigma R (R^\top \Sigma R)^{-1} R^\top \Sigma S\right)^{-1} \\ &= \frac{1}{k^2}\left(L + E/k^2\right)^{-1},\end{aligned}$$

where $E := S^\top \Sigma S - S^\top \Sigma R (R^\top \Sigma R)^{-1} R^\top \Sigma S$. Since E is symmetric and it is constant with respect to k, it has real valued eigenvalues. Therefore, there exists $k_0 \in \mathbb{N}$ such that for all $k > k_0$ the matrix $L + E/k^2$ is positive definite. By using the continuity of matrix inversion for non-singular matrices we get

$$\lim_{k\to\infty} \left(L + E/k^2\right)^{-1} = \left(L + \lim_{k\to\infty} E/k^2\right)^{-1} = L^{-1},$$

and thus $\lim_{k\to\infty}(K/D)^{-1} = 0$. Since B and D are constant with respect to k, all factors in (43) containing $(K/D)^{-1}$ vanish as $k \to \infty$. Therefore, we obtain

$$\lim_{k\to\infty} (Q^\top B_k Q)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & (R^\top \Sigma R)^{-1} \end{bmatrix},$$

which implies that

$$\lim_{k\to\infty} B_k^{-1} = Q \lim_{k\to\infty} (Q^\top B_k Q)^{-1} Q^\top = R(R^\top \Sigma R)^{-1} R^\top.$$

■

B.6 Proof of Theorem 15

Proof We first show that

$$\inf_{P\in\mathcal{P}_0} \mathrm{E}_P\left[\left(f_\star(X) - \mathrm{E}_P[Y \mid X]\right)^2\right] = 0.$$

Recall that if $q < p$ we have P_{tr}-a.s. $f_\star(X) = f_0(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top X]$. Therefore, by Assumption 2 it holds for $P \in \mathcal{P}_0$ that P-a.s. $f_\star(X) = f_0(X) + \mathrm{E}_{P_{\text{tr}}}[\gamma_0(V) \mid R^\top X]$.

Moreover, by (3a) and Assumption 2 it also holds for all $P \in \mathcal{P}_0$ that P-a.s. that $\mathrm{E}_P[Y \mid X] = f_0(X) + \mathrm{E}_P[\gamma_0(V) \mid X]$. Hence, by Assumption 5 it holds that

$$\inf_{P \in \mathcal{P}_0} \mathrm{E}_P \left[\left(f_\star(X) - \mathrm{E}_P[Y \mid X] \right)^2 \right] = \inf_{P \in \mathcal{P}_0} \mathrm{E}_P \left[\left(\mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V) \mid R^\top X] - \mathrm{E}_P[\gamma_0(V) \mid X] \right)^2 \right] = 0. \tag{44}$$

And similarly, for $q = p$ we get that

$$\inf_{P \in \mathcal{P}_0} \mathrm{E}_P \left[\left(f_\star(X) - \mathrm{E}_P[Y \mid X] \right)^2 \right] = \inf_{P \in \mathcal{P}_0} \mathrm{E}_P \left[\left(\mathrm{E}_P[\gamma_0(V) \mid X] \right)^2 \right] = 0. \tag{45}$$

Now, recall that the BCF has constant risk across $\mathcal{P}_0$ since it is invariant by Proposition 13, that is

$$\mathcal{R}(P_{\mathrm{tr}}, f_\star) = \mathrm{E}_{P_{\mathrm{tr}}} \left[\left(Y - f_\star(X) \right)^2 \right] = \mathrm{E}_P \left[\left(Y - f_\star(X) \right)^2 \right], \quad \text{for all } P \in \mathcal{P}_0.$$

We now want to show that for all $f \in \mathcal{F}$ there exists a $P \in \mathcal{P}_0$ such that

$$\mathrm{E}_P \left[(Y - f_\star(X))^2 \right] \leq \mathrm{E}_P \left[(Y - f(X))^2 \right].$$

Let $\varepsilon > 0$ and $f \in \mathcal{F}$, and fix $\delta := \min(\frac{\varepsilon}{2}, \frac{\varepsilon^2}{16\mathcal{R}(P_{\mathrm{tr}}, f_\star)}) > 0$. Thus, by (44) and (45), there exists a $P^* \in \mathcal{P}_0$ such that

$$\mathrm{E}_{P^*} \left[\left(f_\star(X) - \mathrm{E}_{P^*}[Y \mid X] \right)^2 \right] < \delta.$$

Then, using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathrm{E}_{P^*} \left[\left(Y - f_\star(X) \right)^2 \right] &= \mathrm{E}_{P^*} \left[\left(Y - \mathrm{E}_{P^*}[Y \mid X] + \mathrm{E}_{P^*}[Y \mid X] - f_\star(X) \right)^2 \right] \\
&\leq \mathrm{E}_{P^*} \left[\left(Y - \mathrm{E}_{P^*}[Y \mid X] \right)^2 \right] + \mathrm{E}_{P^*} \left[\left(\mathrm{E}_{P^*}[Y \mid X] - f_\star(X) \right)^2 \right] \\
&\quad + 2 \left(\mathrm{E}_{P^*} \left[\left(\mathrm{E}_{P^*}[Y \mid X] - f_\star(X) \right)^2 \right] \mathrm{E}_{P^*} \left[\left(Y - \mathrm{E}_{P^*}[Y \mid X] \right)^2 \right] \right)^{1/2} \\
&< \mathrm{E}_{P^*} \left[\left(Y - \mathrm{E}_{P^*}[Y \mid X] \right)^2 \right] + \delta + 2 \left(\delta \ \mathcal{R}(P_{\mathrm{tr}}, f_\star) \right)^{1/2} \\
&\leq \mathrm{E}_{P^*} \left[\left(Y - \mathrm{E}_{P^*}[Y \mid X] \right)^2 \right] + \varepsilon \\
&\leq \mathrm{E}_{P^*} \left[\left(Y - f(X) \right)^2 \right] + \varepsilon.
\end{aligned}$$

This completes the proof. ■

B.7 Proof of Corollary 16

Proof
Using Theorem 15 and $\mathcal{I}_0 \subseteq \mathcal{F}$ (which holds by definition), we obtain

$$\begin{aligned}\mathcal{R}(P_{\mathrm{tr}}, f_\star) &= \sup_{P \in \mathcal{P}_0} \mathcal{R}(P, f_\star) = \inf_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}_0} \mathcal{R}(P, f) \\ &\leq \inf_{f \in \mathcal{I}_0} \sup_{P \in \mathcal{P}_0} \mathcal{R}(P, f) = \inf_{f \in \mathcal{I}_0} \mathcal{R}(P_{\mathrm{tr}}, f).\end{aligned}$$

Also, since $f_\star \in \mathcal{I}_0$ it holds that

$$\mathcal{R}(P_{\mathrm{tr}}, f_\star) \geq \inf_{f \in \mathcal{I}_0} \mathcal{R}(P_{\mathrm{tr}}, f).$$

■

Lemma 19 *Let $V \sim N(0, \sigma^2)$. Then*

$$\mathrm{E}[(V - \sin(V))^2] = \sigma^2 + \frac{(1 - e^{-2\sigma^2})}{2} - 2\sigma^2 e^{-\sigma^2/2}.$$

Proof First note that $\mathrm{E}[\sin(V)] = 0$ since the sine is an odd function. Also, recall that $\sin(x)^2 = (1 - \cos(2x))/2$, for all $x \in \mathbb{R}$. Using Euler's formula and the characteristic function for the Gaussian distribution, it holds for all $t \in \mathbb{R}$ that

$$\mathrm{E}[\cos(tV)] = \frac{1}{2}\mathrm{E}[e^{itV} + e^{-itV}] = e^{-\sigma^2 t^2/2}.$$

Using Stein's lemma, we have that

$$\mathrm{E}[V \sin(V)] = \sigma^2 \mathrm{E}[\cos(V)] = \sigma^2 e^{-\sigma^2/2}.$$

Putting everything together, we have that

$$\begin{aligned}\mathrm{E}[(V - \sin(V))^2] &= \mathrm{E}[V^2] + \mathrm{E}[\sin(V)^2] - 2\mathrm{E}[V \sin(V)] \\ &= \sigma^2 + \frac{(1 - e^{-2\sigma^2})}{2} - 2\sigma^2 e^{-\sigma^2/2}.\end{aligned}$$

■

B.8 Convergence rate of the BCF estimator with oracle first step

Recall the definition of the BCF $f_\star(x) = f_0(x) + \mathrm{E}[\gamma_0(V) \mid R^\top X = R^\top x]$, for $x \in \mathbb{R}^p$. We have a dataset $\mathcal{D}$ consisting of iid copies $(X_i, Y_i, Z_i, U_i, V_i) \sim P$, with $i \in \{1, \ldots, 2n\}$. Split the dataset into equally sized disjoint parts $\mathcal{D}_1$ and $\mathcal{D}_2$. Estimate the BCF using sample splitting as follows.

1. (*First split* $\mathcal{D}_1$). On $\mathcal{D}_1$, fit the additive model $\hat{f}_n(x) + \hat{\gamma}_n(v)$ by regressing Y on (X, V). The function $\hat{f}_n + \hat{\gamma}_n$ estimates the conditional expectation $\mathrm{E}[Y \mid X = x, V = v] = f_0(x) + \gamma_0(v)$.

2. (*Second split* $\mathcal{D}_2$). On $\mathcal{D}_2$ evaluate the pseudo-outcomes $\hat{\gamma}_n(V)$ and fit the regression function $\hat{m}_n : \mathbb{R}^{p-q} \to \mathbb{R}$ by regressing $\hat{\gamma}_n(V)$ on $R^\top X$. The function $\hat{m}_n$ estimates the conditional expectation $m_n(w) \coloneqq \mathrm{E}[\hat{\gamma}_n(V) \mid R^\top X = w]$.

The resulting BCF estimator satisfies $\hat{f}_\star(x) = \hat{f}_n(x) + \hat{m}_n(R^\top x)$. Before stating the proposition we report the definition of the angle between two subspaces (see Deutsch, 2001, Definition 9.4).

Definition 20 (Angle between two subspaces) *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{M}, \mathcal{N} \subseteq \mathcal{H}$ two closed subspaces such that $\mathcal{M} \cap \mathcal{N} = \{0\}$. Following (Deutsch, 2001, Definition 9.4), define the angle $\theta \in [0, \pi/2]$ between $\mathcal{M}$ and $\mathcal{N}$ to be the angle whose cosine $\cos(\theta)$ is defined by*

$$\cos(\theta) \coloneqq \sup\left\{|\langle u, v\rangle| \colon u \in \mathcal{M}, v \in \mathcal{N}, \|u\| \le 1, \|v\| \le 1\right\}.$$

Definition 20 is used in condition (S3) of Proposition 21, which is equivalent to require that $\mathcal{M} + \mathcal{N}$ is closed (see Deutsch, 2001, Theorem 9.35). This assumption is often used in the additive models' literature (Buja et al., 1989, e.g.,). In the following, all norms are $L_2(P)$ unless specified.

Proposition 21 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space where the random vector $(X, Y, Z, U, V) \sim P$ and the independent datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ are defined. Define $\mathcal{H} \coloneqq L_2(\mathcal{F})$ with inner product $\langle A, B\rangle = \mathrm{E}[AB]$ and define the closed subspaces $\mathcal{H}_X \coloneqq L_2(\sigma(X))$, $\mathcal{H}_V \coloneqq L_2(\sigma(V))$, and $\mathcal{H}_W \coloneqq L_2(\sigma(W))$, where $W \coloneqq R^\top X$. Consider the BCF and its estimator*

$$f_\star(x) = f_0(x) + \mathrm{E}[\gamma_0(V) \mid W = R^\top x], \qquad \hat{f}_\star(x) = \hat{f}_n(x) + \hat{m}_n(R^\top x),$$

and assume that $f_\star$ is identifiable according to Definition 10. Furthermore, assume the following conditions.

(S1) Let r_n be a sequence of strictly positive real numbers. The additive function $\hat{f}_n(x) + \hat{\gamma}_n(v)$ fitted on $\mathcal{D}_1$ by regressing Y on (X, V) satisfies

$$\|\hat{f}_n(X) + \hat{\gamma}_n(V) - f_0(X) - \gamma_0(V)\| = O_p(r_n),$$

that is, for every $\varepsilon > 0$ there exist $M_1 > 0$ and $N_1 > 0$ such that for all $n > N_1$ it holds

$$\mathbb{P}\left(\|\hat{f}_n(X) + \hat{\gamma}_n(V) - f_0(X) - \gamma_0(V)\| > M_1 r_n\right) < \varepsilon.$$

(S2) Let s_n be a sequence of strictly positive real numbers. Conditional on $\mathcal{F}_1 \coloneqq \sigma(\mathcal{D}_1)$, the regression function $\hat{m}_n \colon \mathbb{R}^{p-q} \to \mathbb{R}$ fitted on $\mathcal{D}_2$ satisfies

$$\|\hat{m}_n - m_n\| = O_p(s_n),$$

that is, for every $\varepsilon > 0$ there exist $M_2 > 0$ and $N_2 > 0$ not depending on $\mathcal{F}_1$ such that for all $n > N_2$ it holds

$$\mathbb{P}\left(\|\hat{m}_n - m_n\| > M_2 s_n \mid \mathcal{F}_1\right) < \varepsilon.$$

*(S3) (*Angle condition*). The angle $\theta \in [0, \pi/2]$ between $\mathcal{H}_X$ and $\mathcal{H}_V \cap \mathcal{H}_W^{\perp}$ as in Definition 20 is such that $\theta > 0$.*

Then

$$\|\hat{f}_\star - f_\star\| = O_p\big(r_n + s_n\big).$$

Proof Since the BCF $f_\star$ is identifiable, by Lemma 22 this implies $\mathcal{H}_X \cap (\mathcal{H}_V \cap \mathcal{H}_W^{\perp}) = \{0\}$. The trivial intersection between two subspaces is called "no-concurvity" in the additive models' literature (Buja et al., 1989) and plays the same role as the absence of multicollinearity in linear regression. Together with the angle condition (S3), it will help us to upper bound the rate of the BCF estimator with the bounds of the additive model and the regression estimator $\hat{m}_n$.

We have the following quantities.

(Q1) $\hat{f}_n, \hat{\gamma}_n$ are fitted on $\mathcal{D}_1$,

(Q2) $\hat{m}_n$ is fitted on $\mathcal{D}_2$,

(Q3) $m_n(W) = \mathrm{E}[\hat{\gamma}_n(V) \mid W]$,

(Q4) $\tilde{\gamma}_n(V) \coloneqq \hat{\gamma}_n(V) - \hat{m}_n(W)$,

(Q5) $\gamma^{\perp}(V) \coloneqq \tilde{\gamma}_n(V) - \mathrm{E}[\tilde{\gamma}_n(V) \mid W]$.

Condition on $\mathcal{F}_{12} \coloneqq \sigma(\mathcal{D}_1, \mathcal{D}_2)$. We have that $\mathrm{E}[\tilde{\gamma}_n(V) \mid W] = m_n(W) - \hat{m}_n(W)$, and therefore,

$$\gamma^{\perp}(V) - \tilde{\gamma}_n(V) = -\mathrm{E}[\tilde{\gamma}_n(V) \mid W] = \hat{m}_n(W) - m_n(W). \tag{46}$$

Define $\gamma_\star(V) \coloneqq \gamma_0(V) - \mathrm{E}[\gamma_0(V) \mid W]$ so that $f_0(X) + \gamma_0(V) = f_\star(X) + \gamma_\star(V)$. We want to bound

$$\hat{f}_\star(X) - f_\star(X) = \hat{f}_n(X) + \hat{m}_n(W) - f_\star(X). \tag{47}$$

Note that

$$\begin{aligned}
S_n &\coloneqq \hat{f}_n(X) + \hat{\gamma}_n(V) - f_0(X) - \gamma_0(V) \\
&= \hat{f}_n(X) + \hat{\gamma}_n(V) - f_\star(X) - \gamma_\star(V) \\
&\overset{(Q4)}{=} \hat{f}_n(X) + \hat{m}_n(W) + \tilde{\gamma}_n(V) - f_\star(X) - \gamma_\star(V) \\
&= \hat{f}_n(X) + m_n(W) - f_\star(X) + \gamma^{\perp}(V) - \gamma_\star(V) \\
&\quad + \hat{m}_n(W) - m_n(W) + \tilde{\gamma}_n(V) - \gamma^{\perp}(V) \\
&\overset{\clubsuit}{=} \hat{f}_n(X) + m_n(W) - f_\star(X) + \gamma^{\perp}(V) - \gamma_\star(V),
\end{aligned} \tag{48}$$

where in ♣ we used (46). Now, define $Q := \hat{f}_n(X)+m_n(W)-f_\star(X)$ and $T := \gamma^\perp(V)-\gamma_\star(V)$, so that $S_n = Q+T$. Conditional on $\mathcal{F}_{12}$, Q is a measurable function of X and so $Q \in \mathcal{H}_X$. Also, T is a measurable function of V with $\mathrm{E}[T \mid W] = 0$ so $T \in \mathcal{H}_V \cap \mathcal{H}_W^\perp$. Since $\mathcal{H}_X \cap (\mathcal{H}_V \cap \mathcal{H}_W^\perp) = \{0\}$, using Lemma 23 and the fact that $\theta > 0$, by (S3) we get that

$$\|Q\| \leq \frac{\|S_n\|}{\sin(\theta)}. \tag{49}$$

Also, from (47), it holds

$$\|\hat{f}_\star(X) - f_\star(X)\| \leq \|Q\| + \|\hat{m}_n(W) - m_n(W)\|. \tag{50}$$

Combining (49) and (50), conditional on $\mathcal{F}_{12}$, it holds that

$$\|\hat{f}_\star(X) - f_\star(X)\| \leq \frac{\|S_n\|}{\sin(\theta)} + \|\hat{m}_n(W) - m_n(W)\|. \tag{51}$$

Fix $\varepsilon > 0$. For $\tilde{\varepsilon} := \varepsilon/2$, there exist $M_1, N_1 > 0$ and $M_2, N_2 > 0$ such that (S1) and (S2) hold with $\tilde{\varepsilon}$, respectively. Define $c_\theta := 1/\sin(\theta) > 0$ and fix $M \geq \max\{c_\theta M_1, M_2\}$. Using (51), it holds that

$$\begin{aligned}
&\Big\{\|\hat{f}_\star(X) - f_\star(X)\| > M(r_n + s_n)\Big\}\\
&\quad\subseteq \big\{\|S_n\|c_\theta + \|\hat{m}_n(W) - m_n(W)\| > M(r_n + s_n)\big\}\\
&\quad\subseteq \big\{\|S_n\|c_\theta > Mr_n\big\} \cup \big\{\|\hat{m}_n(W) - m_n(W)\| > Ms_n\big\}\\
&\quad\subseteq \big\{\|S_n\| > M_1 r_n\big\} \cup \big\{\|\hat{m}_n(W) - m_n(W)\| > M_2 s_n\big\}.
\end{aligned} \tag{52}$$

Conditional on $\mathcal{F}_{12}$, $\|S_n\|$ and $\|\hat{m}_n - m_n\|$ are deterministic quantities, and therefore using (52) we get

$$\begin{aligned}
&\mathbb{P}\Big(\|\hat{f}_\star(X) - f_\star(X)\| > M(r_n + s_n) \mid \mathcal{F}_{12}\Big)\\
&\quad\leq \mathbf{1}\big\{\|S_n\| > M_1 r_n\big\} + \mathbf{1}\big\{\|\hat{m}_n - m_n\| > M_2 s_n\big\}.
\end{aligned}$$

Using the law of iterated expectations twice we get, for all $n > \max\{N_1, N_2\}$,

$$\begin{aligned}
&\mathbb{P}\Big(\|\hat{f}_\star(X) - f_\star(X)\| > M(r_n + s_n)\Big)\\
&\quad\leq \mathbb{P}\big(\|S_n\| > M_1 r_n\big) + \mathrm{E}\Big[\mathbb{P}\big(\|\hat{m}_n - m_n\| > M_2 s_n \mid \mathcal{F}_1\big)\Big] \leq \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this yields

$$\|\hat{f}_\star(X) - f_\star(X)\| = O_p(r_n + s_n).$$

■

Lemma 22 *Let $\mathcal{H} = L_2(\mathcal{F})$ with inner product $\langle A, B\rangle = \mathrm{E}[AB]$, and consider the closed subspaces $\mathcal{H}_X = L_2(\sigma(X))$, $\mathcal{H}_V = L_2(\sigma(V))$, and $\mathcal{H}_W = L_2(\sigma(W))$. Suppose $f_\star$ is identifiable according to Definition 10. Then, it holds that $\mathcal{H}_X \cap (\mathcal{H}_V \cap \mathcal{H}_W^\perp) = \{0\}$.*

Proof Fix $\ell \in \mathcal{H}_X \cap (\mathcal{H}_V \cap \mathcal{H}_W^\perp)$. Then, there exist $f(X) \in \mathcal{H}_X$ and $\gamma(V) \in \mathcal{H}_V \cap \mathcal{H}_W^\perp$ such that $\ell = f(X) = \gamma(V)$, P_{tr}-a.s., and therefore $f(X) - \gamma(V) = 0$, P_{tr}-a.s.. From Proposition 11, there exists a δ such that $f(X) = \delta(R^\top X) = \delta(W)$, P_{tr}-a.s.. Conditioning on W and using the fact that $f(X) - \gamma(V) = 0$, P_{tr}-a.s., we get

$$0 = \mathrm{E}[f(X) - \gamma(V) \mid W] = \delta(W) - \mathrm{E}[\gamma(V) \mid W] = \delta(W) - 0 = \delta(W), \quad P_{\text{tr}}\text{-a.s.},$$

because $\mathrm{E}[\gamma(V) \mid W] = 0$ since $\gamma(V) \in \mathcal{H}_W^\perp$. Therefore, $\ell = f(X) = \gamma(V) = 0$, P_{tr}-a.s. Since $\ell \in \mathcal{H}_X \cap (\mathcal{H}_V \cap \mathcal{H}_W^\perp)$ was arbitrary, we conclude that $\mathcal{H}_X \cap (\mathcal{H}_V \cap \mathcal{H}_W^\perp) = \{0\}$. ■

Lemma 23 *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{M}, \mathcal{N} \subseteq \mathcal{H}$ two closed subspaces such that $\mathcal{M} \cap \mathcal{N} = \{0\}$. Let θ denote the angle between $\mathcal{M}$ and $\mathcal{N}$ as in Definition 20. Then, for all $u \in \mathcal{M}$ and $v \in \mathcal{N}$ it holds that*

$$\|u\| \sin(\theta) \leq \|u + v\|.$$

Proof Fix $u \in \mathcal{M}$ and $v \in \mathcal{N}$ and note that by Definition 20 it holds

$$|\langle u, v\rangle| \leq \cos(\theta)\|u\|\|v\|. \tag{53}$$

Also note that

$$\begin{aligned} \|u + v\|^2 &= \|u\|^2 + \|v\|^2 + 2\langle u, v\rangle \\ &\geq \|u\|^2 + \|v\|^2 - 2|\langle u, v\rangle| \\ &\overset{\clubsuit}{\geq} \|u\|^2 + \|v\|^2 - 2\cos(\theta)\|u\|\|v\|, \end{aligned} \tag{54}$$

where in ♣ we used (53). The right-hand side of (54) is a quadratic function in $\|v\|$, which is minimized at $t^* = \cos(\theta)\|u\|$, therefore,

$$\|u + v\|^2 \geq \|u\|^2 + \cos^2(\theta)\|u\|^2 - 2\cos^2(\theta)\|u\|^2 \tag{55}$$

$$= \|u\|^2 \sin^2(\theta), \tag{56}$$

from which it follows that $\|u\| \sin(\theta) \leq \|u + v\|$. ■

Appendix C. Comparison with Existing OOD Generalization Frameworks

This section provides a discussion of existing families of methods for out-of-distribution (OOD) generalization, comparing their assumptions to those of BCF, and why they may fail under the types of distributional shifts considered in this work.

A first family of approaches addresses the problem of OOD generalization by optimizing the worst-case risk over a neighborhood of the training distribution. In distributionally robust optimization (DRO) the robustness set is typically defined as a ball around the empirical training distribution with respect to a probability distance, such as the Wasserstein metric (Sinha et al., 2018) or an f-divergence (Bagnell, 2005; Hu et al., 2018). Although such methods provide formal guarantees, they can be overly conservative. To address this, Sagawa et al. (2020) proposed GroupDRO, which defines the robustness set as the convex hull of finitely many training distributions and minimizes the worst-case empirical risk across them. As discussed in Shen et al. (2025), such models cannot guarantee robustness on perturbations outside the training support and do not provide a clear geometric characterization of the robustness set. For this reason, their connection to our setting remains unclear.

A second family of methods seeks predictors that satisfy a notion of invariance across different environments Z. An established approach is to find a representation $x \mapsto \Phi(x)$ such that $\mathbb{P}[Y \mid \Phi(X) = \Phi(x)]$ remains invariant across perturbations in Z (Magliacane et al., 2018; Rojas-Carulla et al., 2018; Krueger et al., 2021; Arjovsky et al., 2020). In our framework, these approaches may fail in the presence of hidden confounding between X and Y, as we show in the next example. For simplicity, we restrict our attention to linear representations $x \mapsto \Phi(x) \coloneqq a + bx$, for $a, b \in \mathbb{R}$. Consider the linear Gaussian structural causal model $X \coloneqq Z + V$, $Y \coloneqq X + U$, where $(U, V) \sim N(0, \Sigma)$, and $Z \perp\!\!\!\perp (U, V)$, and assume that the shifts are generated by arbitrary perturbations of the marginal distribution of Z. We now show that for all $a, b \in \mathbb{R}$, there exists no invariant Φ. When $b = 0$, the representation $\Phi(x) \coloneqq a$, for $a \in \mathbb{R}$, is not invariant since interventions on Z change the marginal distribution of Y. When $b \neq 0$,

$$\begin{aligned} \mathrm{E}[Y \mid \Phi(X) = x] &= \mathrm{E}\left[Y \mid X = \frac{x-a}{b}\right] = \frac{x-a}{b} + \mathrm{E}\left[U \mid X = \frac{x-a}{b}\right] \\ &= \left(1 + \frac{\mathrm{Cov}(U, X)}{\mathrm{Var}(X)}\right) \frac{x-a}{b} \end{aligned}$$

is not invariant either, since it depends on $\mathrm{Var}(X) = \mathrm{Var}(Z) + \mathrm{Var}(V)$.

A distinct notion of invariance is the counterfactual invariance of Veitch et al. (2021). In their definition, a predictor f is counterfactually invariant to Z if $f(X(z)) = f(X(z'))$ for all z, z'. This definition expressed in the potential outcome notation, applies to both the causal and anticausal setting, including cases where Z and Y are confounded. Within our SIMDG defined in (3a)–(3c), a counterfactually invariant predictor corresponds to a function of the form $f(X) = \delta(R^\top X, X^\top e)$, where $e_j = 1$ if X_j is not affected by Z and $e_j = 0$ otherwise. If the structural function f_0 depends on components of X affected by Z, then (i) our BCF predictor is not counterfactually invariant in the sense of Veitch et al. (2021), and (ii) any counterfactually invariant predictor does not satisfy our invariance notion (see Definition 6).

Several above approaches, including (Rojas-Carulla et al., 2018; Arjovsky et al., 2020; Krueger et al., 2021; Puli et al., 2021; Veitch et al., 2021) can handle anti-causal settings, where the target Y causes some of the observed covariates X. In general, our BCF method does not handle anti-causal setting of the form

$$\begin{aligned} X &= M_0 Z + V \\ Y &= f_0(X) + U \\ \tilde{X} &= g_0(Y) + \tilde{M}_0 Z + \tilde{V}, \end{aligned} \tag{57}$$

because Z influences $\tilde{X}$ via the composition $g_0(f_0(M_0 \cdot + v) + u)$, which may be nonlinear. Indeed, when g_0 and f_0 are nonlinear, Christiansen et al. (2022) shows that generalization is impossible under arbitrary interventions on Z. However, if the composition $g_0(f_0(M_0 \cdot +v) + u)$ happens to be linear, then our framework still applies since we can rewrite (57) as a SIMDG as defined in Definition 5.

Appendix D. Simulated Trees for Experiment 1 (Section 5.1)

The function $f_0 : \mathbb{R}^p \to \mathbb{R}$ is a decision tree depending on the first $p_{\text{eff}} < p$ predictors. For any $x \in \mathbb{R}^p$ it is defined by

$$f_0(x) := \sum_{h=1}^{2^d} \theta_h \mathbf{1}\{x \in t_h\},$$

where θ_h denote the constant values over the rectangular regions $t_h \subseteq \mathbb{R}^p$. We sample the constant values independently according to $\theta \sim N(0, 1.5^2)$. We build the rectangular regions t_h recursively; for each t_h, we uniformly sample a predictor $j \in \{1, \dots, p_{\text{eff}}\}$, where $p_{\text{eff}} < p$ denotes the number of effective predictors, and randomly choose the split point $s_j \sim \text{U}([-2, 2])$ to obtain two children regions $t_{h,0}$ and $t_{h,1}$. We set the number of effective sample predictors to $p_{\text{eff}} = 3$ and the tree-depth to $d = 3$.

Appendix E. Fitting BCF with Neural Networks

In Section 4.3, we described how to estimate the BCF with nonparametric estimators. In particular, we introduced the ControlTwicing algorithm to estimate the conditional expectation

$$\mathrm{E}_{P_{\text{tr}}}[Y \mid X = x, V = v] = f_0(x) + \gamma_0(v)$$

via nonparametric regression methods for both f_0 and γ_0. Here, we show how to estimate the BCF using neural networks. The first step–which estimates the matrix $M_0 \in \mathbb{R}^{p \times q}$, its left null space matrix $R \in \mathbb{R}^{p \times (p-q)}$, and the control variables $V \in \mathbb{R}^p$–remains unchanged. The second and third steps are modified as follows. Given a sample of n observations $(X_1, Y_1, V_1), \dots, (X_n, Y_n, V_n)$:

1. Estimate the conditional expectation $\mathrm{E}_{P_{\text{tr}}}[Y \mid X, V] = f_0(X) + \gamma_0(V)$ by solving

$$(\theta_1^*, \theta_2^*) \in \arg\min_{\theta_1, \theta_2} \frac{1}{n} \sum_{i=1}^n (Y_i - f_{\theta_1}(X_i) - \gamma_{\theta_2}(V_i))^2, \tag{58}$$

   where f_{θ_1} and γ_{θ_2} are neural networks parametrized by θ_1 and θ_2, respectively.

2. Estimate the conditional expectation $\mathrm{E}_{P_{\text{tr}}}[Y - f_{\theta_1^*}(X) \mid R^\top X]$ by solving

$$\theta_3^* \in \arg\min_{\theta_3} \frac{1}{n} \sum_{i=1}^n (Y_i - f_{\theta_1}(X_i) - \delta_{\theta_3}(V_i))^2,$$

   where δ_{θ_3} is a neural network, parametrized by θ_3.

The two optimization problems can be solved with classical stochastic gradient descent algorithms (Cauchy et al., 1847; Robbins and Monro, 1951) or adaptive optimizers, e.g., Kingma and Ba (2015); Loshchilov and Hutter (2019). A ready-to-use implementation of BCF with neural networks is provided in our repository `https://github.com/nicolagnecco/bcf-numerical-experiments` via the class `BCFMLP`.

Appendix F. Additional Numerical Experiments

F.1 Data-Generating Process

We consider a data-generating process similar to that in Saengkyongam et al. (2022), defined as

$$\begin{aligned} X_1 &= Z_k + V_1 + \varepsilon_1, \\ X_2 &= V_2 + \varepsilon_2, \\ Y &= f_0(X_1, X_2) + \gamma_0(V_1, V_2) + \varepsilon_Y, \end{aligned} \tag{59}$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_Y \overset{iid}{\sim} N(0, 0.1)$, $V_1, V_2 \overset{iid}{\sim} N(0,1)$, and $(Z_k, V_1, V_2, \varepsilon_1, \varepsilon_2, \varepsilon_Y)$ are mutually independent for $k \in (0, 4)$. We define two versions of the exogenous variable Z_k, namely

$$Z_k := \begin{cases} B_k U_{k,4} + (1 - B_k) U_{0,k}, & \text{if continuous,} \\ k B_2, & \text{if discrete,} \end{cases} \tag{60}$$

where $B_k \sim Bern(k/4)$ and $U_{a,b} \sim Unif(a, b)$ for $a, b \in \mathbb{R}$. Define the mixture of radial basis functions

$$\varphi(x) := \sum_{i=1}^{10} w_j \exp\left(\frac{-\left\|x - c_j\right\|}{3}\right)^2, \; x \in \mathbb{R}^2, \tag{61}$$

where $c_j \sim Unif[-5, 5]$ and $w_j \sim N(0, 4)$. The structural function f_0 is defined by $f_0(x) := \varphi(x)$, and the control function is defined by

$$\gamma_0(v_1, v_2) := \begin{cases} v_1 + v_2, & \text{if linear,} \\ \alpha(\varphi(v_1, v_2) - \beta), & \text{if nonlinear,} \end{cases} \tag{62}$$

where $\alpha > 0$ and $\beta \in \mathbb{R}$ are such that $\mathrm{E}_{P_{\text{tr}}}[\gamma_0(V_1, V_2)] = 0$ and $\mathrm{E}_{P_{\text{tr}}}[\gamma_0(V_1, V_2)^2] = 2$.

The parameter k controls the strength of the perturbation on Z_k. The direction of the perturbations is spanned by $M_0 = (1, 0)^\top$ and the invariant space by $R = (0, 1)^\top$. Each $k \in (0, 4)$ induces a distribution Q_k over Z_k which in turns induces a distribution $P_k^{(X,Y)}$ over $(X, Y) \in \mathbb{R}^{(2+1)}$, via (59).

F.2 Method Configurations

In the experiments of Sections F.3, F.4, and F.5 we consider two implementations of the BCF estimator, BCF-MLP and BCF-XGB. Both share the same logic as described in Algorithm 1 and differ only in the choice of base regressors to estimate f_0, γ_0, and $\delta_0 := \mathrm{E}_{P_{\text{tr}}}(\gamma_0(V) \mid R^\top X)$.

BCF-MLP. Each component function is modelled by a one-hidden layer neural network with 64 hidden units and a sigmoid activation function. The parameters of f_{θ_1} and γ_{θ_2} are optimized for 1000 epochs with learning rate 10^{-3} and weight decay 2.5×10^{-3}. The parameters of δ_{θ_3} are optimized for 1500 epochs with a smaller learning rate 10^{-4} and no weight decay. All optimizations use the AdamW optimizer (Loshchilov and Hutter, 2019), which decouples the learning rate from the weight decay. The larger number of epochs to train δ_{θ_3} helps to capture as much of the invariant signal from Step 4 of the BCF algorithm (Algorithm 1). A reference implementation is provided in our repository via the class `BCFMLP`.

BCF-XGB. Each component function is modelled by an XGBoost regressor (Chen and Guestrin, 2016) with learning rate of 0.05 for estimating $\hat{f}$ and $\hat{\gamma}$ and of 0.25 for estimating $\hat{\delta}$. The larger learning rate for $\hat{\delta}$ helps to capture as much of the invariant signal from Step 4 of the BCF algorithm (Algorithm 1). The number of passes in the ControlTwicing algorithm (see Algorithm 2) is set to $J = 75$ with an early stopping rule satisfied if

$$\frac{\left\|\hat{f}_{j+1}(\mathbf{X}) - \hat{f}_j(\mathbf{X})\right\|_2^2}{\left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_2^2} < 5 \times 10^{-3},$$

where the $\hat{f}_j$ denotes the gradient boosted tree at iteration j. The large number of passes, the early stopping rule, and the relatively low learning rate for $\hat{f}$ and $\hat{\gamma}$ help the ControlTwicing algorithm to learn the highly nonlinear f_0 defined in (61) while reducing the risk of overfitting. A reference implementation is provided in our repository via the class `BCF`.

In addition to the BCF implementations, we consider the following baselines.

CF-MLP. This baseline implements the standard control function approach, where the component functions f_0 and γ_0 are modelled by a one-hidden layer neural network with 64 hidden units and sigmoid activation function. The method returns the learned function f_{θ_1}. The configuration is identical to BCF-MLP, except that the step to estimate δ_{θ_3} is omitted. A reference implementation is provided in our repository via the class `BCFMLP` with parameter `predict_imp=False`.

CF-XGB. This baseline implements the standard control function approach, where the component functions f_0 and γ_0 are modelled by XGBoost regressors (Chen and Guestrin, 2016). The method returns the learned function $\hat{f}$. The configuration is identical to BCF-XGB, except that the step to estimate $\hat{\delta}$ is omitted. A reference implementation is provided in our repository via the class `BCF` with parameter `predict_imp=False`.

LS-MLP. Implementation of the least squares regression estimator using a one-hidden layer neural network with 64 hidden units and sigmoid activation function. The parameters of the network are optimized for 1000 epochs with a learning rate 10^{-3} and a weight decay 2.5×10^{-3} using the AdamW optimizer (Loshchilov and Hutter, 2019). A reference implementation is provided in our repository via the class `OLS-MLP`.

LS-XGB. Implementation of the least squares regression estimator using an XGBoost regressor (Chen and Guestrin, 2016) with learning rate of 0.05. A reference implementation is provided in our repository via the class `OLS`.

F.3 Experiment on the Efficiency and Robustness of BCF

This experiment studies how the performance of the BCF estimator evolves with the training sample size n and with perturbations k in the exogenous variable Z_k. Varying n tests the estimator's sample efficiency, namely its ability to achieve good predictive accuracy with limited data. Increasing k measures the estimator's robustness, that is, its stability across shifts in the distribution of Z.

We use the data-generating process described in Section F.1 with continuous Z_k and linear control function $\gamma_0(v_1, v_2) = v_1 + v_2$. For each training sample size $n \in \{500, 1000, 2500\}$, we repeat the following procedure ten times.

1. Draw a random realization of the structural function $f_0 = \varphi(x)$ as in (61).
2. Generate n training observations $\{(X_i, Y_i, Z_{ki})\}_{i=1}^n$ according to (59) with $k = 0.5$.
3. Estimate the BCF using the BCF-XGB and the BCF-MLP configurations described in Section F.2.
4. As oracle baselines, consider (i) the structural function f_0, and (ii) the IMP function $f_\star(x) \coloneqq f_0(x) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V1, V2) \mid R^\top X = x]$, where $R = (0, 1)^\top$.
5. For each perturbation strength $k \in \{0.5, 1, 1.5, \ldots, 3.5, 3.99\}$, generate $n_{\mathrm{te}} = 1000$ test samples $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{te}}}$ from (59) with the same f_0 and γ_0 and denote by $P_k^{(X,Y)}$ the resulting distribution of (X, Y).
6. Evaluate each method by its mean squared error (MSE) under $P_k^{(X,Y)}$.

Figure 5 displays the test MSE of each method as a function of the perturbation strength k, for training sample sizes $n \in \{500, 1000, 2500\}$. The solid lines show the average MSE across ten repetitions, while the shaded ribbons indicate the pointwise minimum and maximum MSE attained across runs. The dotted and dashed lines correspond to the oracle risks of the structural f_0 and the IMP $f_\star$, respectively.

We assess sample efficiency by looking at how the mean and variability of the MSE evolve with n for fixed k. As n increases, the shaded ribbons shrink and the mean curves move closer to the population baseline, indicating reduced variance and estimation error. In this setting, the BCF-MLP converges slightly faster toward the oracle than BCF-XGB. We assess robustness by looking at how the average MSE evolves with k for fixed n. For smaller sample sizes ($n = 500$ and $n = 1000$), BCF-MLP shows greater robustness than BCF-XGB. For $n = 2500$, both estimators show nearly constant MSE across perturbation strengths, approaching the invariant risk of $f_\star$.

F.4 Experiment on the Effect of Regularization on Control Function Method

Compared to nonparametric regression methods, such as random forests or boosted trees, neural networks offer a direct way to regularize the weights θ of a predictive function f_θ through a weight decay factor $\lambda > 0$. In this experiment, we regularize the neural networks' weights with the adaptive AdamW optimizer (Loshchilov and Hutter, 2019), which decouples the weight decay from the learning rate.

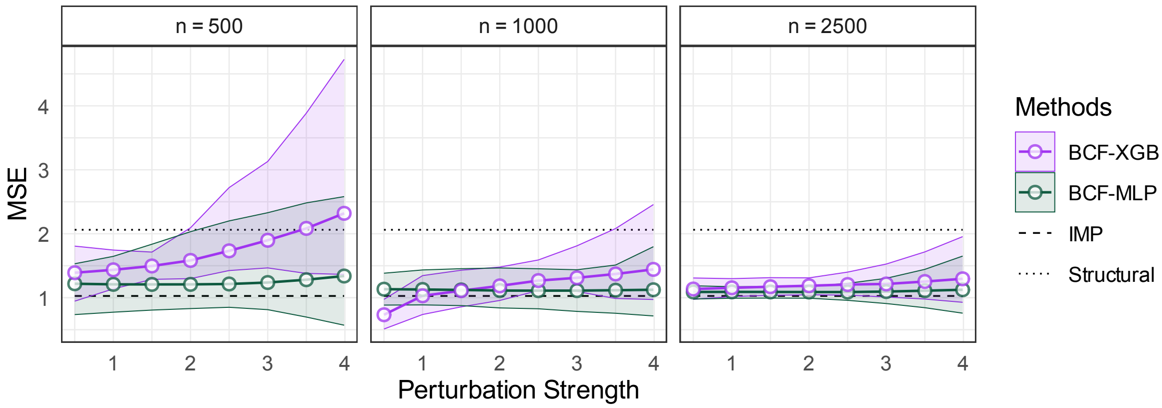

Figure 5: Test mean squared error (MSE) of different estimators as a function of the perturbation strength k. Solid lines show the average MSE over ten repetitions; shaded ribbons indicate the pointwise minimum and maximum MSE across runs. Dotted and dashed lines correspond to the oracle predictors f_0 and $f_\star$. As n increases, both methods show a reduced variability in the MSE and their average MSEs move closer to the population baseline. As k increases, BCF-MLP shows better robustness than BCF-XGB for small to medium sample sizes. For large n, both estimators show nearly constant average MSE across perturbation strengths.

Recall Step 3 of the BCF algorithm, where we estimate the conditional expectation

$$\mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X, V] = f_0(X) + \gamma_0(V).$$

As discussed in Example 2, f_0 and γ_0 may not be identifiable from P_{tr}; there can exist infinitely many pairs (f, γ) such that $\mathrm{E}_{P_{\mathrm{tr}}}[Y \mid X, V] = f_0(X) + \gamma_0(V)$, $P_{\mathrm{tr}}^{X,V}$-almost surely. When estimating this conditional expectation by $f_{\theta_1} + \gamma_{\theta_2}$ via neural networks, the particular solution (θ_1^*, θ_2^*) obtained during training depends on the capacity of the two estimators. In this experiment, we study how regularizing (θ_1, θ_2) influences the predictive performance and invariance of $f_{\theta_1^*}$.

We consider the data-generating process described in Section F.1 with continuous Z_k and linear control function $\gamma_0(v_1, v_2) = v_1 + v_2$. We repeat the following procedure ten times.

1. Draw a random realization of the structural function $f_0 = \varphi(x)$ as in (61).

2. Generate $n = 1000$ training observations $\{(X_i, Y_i, Z_{ki})\}_{i=1}^n$ according to (59) with $k = 0.5$.

3. Estimate the BCF using the BCF-MLP configurations described in Section F.2.

4. As baselines consider (i) the control function estimator CF-MLP described in Section F.2, with weight decay parameters $\lambda \in \{0.0025, 0.025, 0.25, 2.5\}$ for both f_{θ_1} and γ_{θ_2}; (ii) the least squares estimator LS-MLP configured as in Section F.2; (iii)

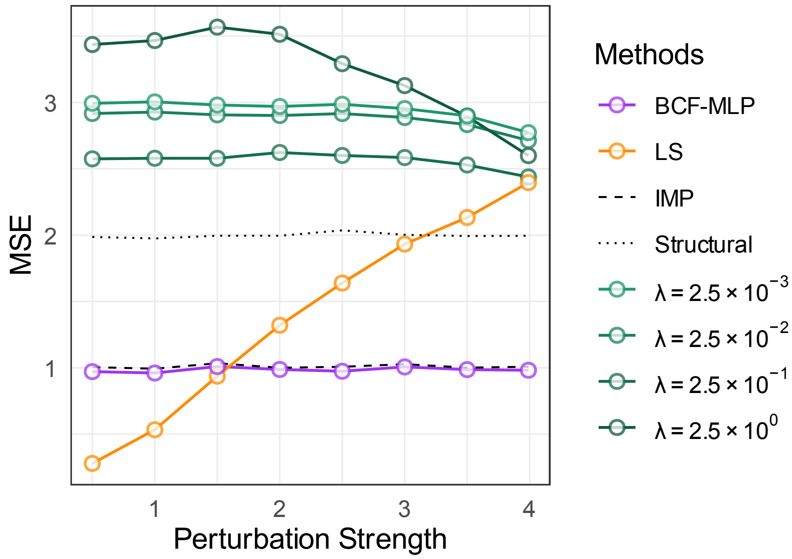

Figure 6: Test mean squared error (MSE) as a function of the perturbation strength k for different methods. Solid lines show the average MSE across ten repetitions. Dotted and dashed lines indicate the oracle predictors f_0 and $f_\star$. The green lines correspond to the standard control function estimator (CF-MLP) with weight decay parameter λ applied to both f_{θ_1} and γ_{θ_2}. Moderate regularization stabilizes CF-MLP, while strong regularization ($\lambda = 2.5 \times 10^0$) leads to non-invariant MSE. BCF-MLP achieves a low and nearly invariant MSE across k, close to the oracle IMP.

the structural function f_0; (iv) the IMP function $f_\star(x) := f_0(x) + \mathrm{E}_{P_{\mathrm{tr}}}[\gamma_0(V1, V2) \mid R^\top X = x]$, where $R = (0, 1)^\top$.

5. For each perturbation strength $k \in \{0.5, 1, 1.5, \ldots, 3.5, 3.99\}$, generate $n_{\mathrm{te}} = 1000$ test samples $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{te}}}$ from (59) with the same f_0 and γ_0 and denote by $P_k^{(X,Y)}$ the resulting distribution of (X, Y).

6. Evaluate each method by its mean squared error (MSE) under $P_k^{(X,Y)}$.

Figure 6 shows the test MSE of each method as a function of the perturbation strength k. The solid lines show the average MSE across ten repetitions, while the dotted and dashed lines correspond to the oracle risks of the structural f_0 and the IMP $f_\star$, respectively.

In this specific setting, moderate regularization, $\lambda \in \{0.0025, 0.025, 0.25\}$, stabilizes the control function estimator, yielding an approximately invariant MSE across different perturbation strengths. However, for strong regularization, $\lambda = 2.5$, the invariance property degrades due to underfitting of $f_{\theta_1^*}$. In the same setting, BCF-MLP trained with $\lambda = 0.0025$ achieves both low and invariant MSE across k, approaching the oracle risk of the IMP $f_\star$.

F.5 Experiment on Identifiability Assumptions

In this experiment, we study the performance of the BCF estimator in finite samples under the two identifiability assumptions discussed in Section 4 (Assumptions 3 and 4). Assumption 3 requires the control function γ_0 to be linear and the exogenous variable Z_k to be discrete at training time, while Assumption 4 requires f_0 and γ_0 to be differentiable and Z_k to be continuous. For each training sample size $n \in \{1000, 5000\}$ and each of the two assumptions, we repeat the following experiment ten times.

1. Draw a random realization of the structural function $f_0 = \phi(x)$ as in (61). Under Assumption 3, fix $\gamma_0(v_1, v_2) = v_1 + v_2$; otherwise draw a random realization $\gamma_0 = \alpha(\phi(v_1, v_2) - \beta)$ as in (62).

2. Generate $n = 1000$ training observations $\{(X_i, Y_i, Z_{ki})\}_{i=1}^n$ according to (59) with $k = 0.5$. Under Assumption 3, Z_k is discrete; otherwise it is continuous as described in (60).

3. Estimate the BCF using the BCF-XGB configurations described in Section F.2.

4. As baselines, consider CF-XGB and LS-XGB as configured in Section F.2. As oracle baselines, we consider (i) the structural function f_0, and (ii) the oracle-BCF function $\hat{f}_\star(x) \coloneqq f_0(x) + \hat{\mathrm{E}}_{P_{\mathrm{tr}}}[\gamma_0(V1, V2) \mid R^\top X = x]$, where the second term is estimated with an XGBoost regressor, and $R = (0, 1)^\top$.

5. For each perturbation strength $k \in \{0.5, 1, 1.5, \dots, 3.5, 3.99\}$, generate $n_{\mathrm{te}} = 1000$ test samples $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{te}}}$ from (59) with the same f_0 and γ_0 and Z_k continuous as in (60). Denote by $P_k^{(X,Y)}$ the resulting distribution of (X, Y).

6. Evaluate each method by its mean squared error (MSE) under $P_k^{(X,Y)}$.

Figure 7 shows the test MSE of each estimator as a function of the perturbation strength k for both assumptions and sample sizes $n \in \{1000, 5000\}$. The solid lines show the average MSE across ten repetitions. In this numerical experiment, across all settings, BCF-XGB outperforms the standard control function estimator CF-XGB and achieves lower MSE than LS-XGB for moderate to large perturbation strengths $k > 2$. Under both identifiability assumptions, the MSE curve of BCF-XGB approaches its oracle counterpart as the training sample size increases.

Appendix G. Guidelines on When BCF Can Be Helpful in Practice

In practice, using BCF can be helpful when (1) unobserved variables H influence both X and Y, and (2) the practitioner has access to variables Z that are (i) exogenous, and (ii) unavailable at test time. The choice of an exogenous Z typically relies on domain knowledge. The unavailability of Z at test time can arise for practical reasons. For example, this can occur when Z is categorical and new categories appear at test time, making Z unusable as a covariate. In other cases, such as the housing datasets described in Section 5.3, Z is continuous and available at test time, but the conditional distribution of $Y \mid X, Z$ may not extrapolate to unseen values of Z.

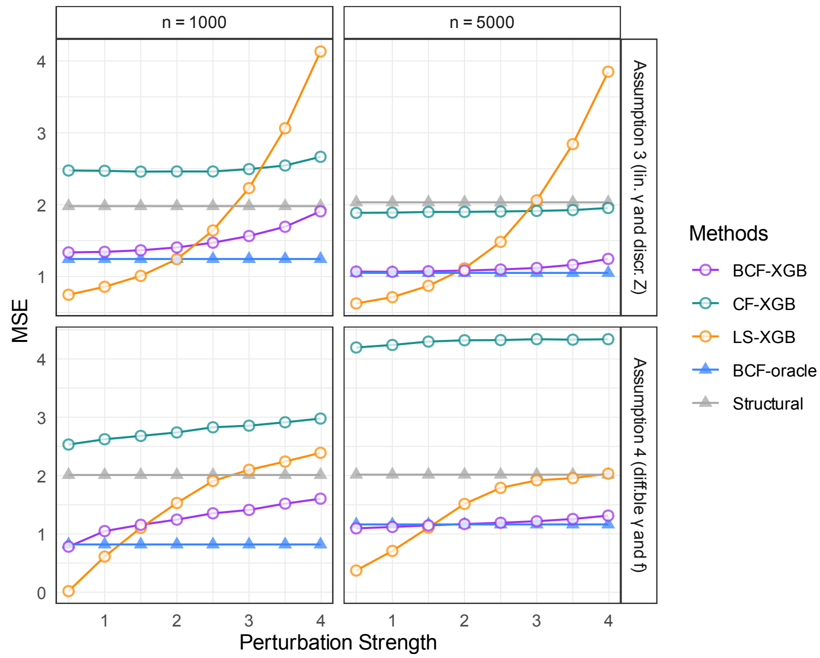

Figure 7: Test mean squared error (MSE) as a function of the perturbation strength k for different methods. Each panel shows different sample sizes and assumptions. Solid lines show the average MSE across ten repetitions. BCF-XGB outperforms CF-XGB under all settings and LS-XGB for moderate to large perturbation sizes. Under both assumptions, as n increases, BCF-XGB approaches the performance of its oracle counterpart BCF-oracle.

Once such a variable Z has been identified, a simple pipeline to motivate the use of BCF is as follows. First, split the training data according to the values of Z (for categorical Z, use distinct categories; for continuous Z, create bins or buckets). As a baseline, fit a least squares model, and evaluate whether its MSE remains stable across held-out splits. If it does not, this suggests that Z might induce distribution shifts in (X, Y), and applying BCF might be beneficial. Finally, to heuristically assess BCF's robustness, verify that its MSE remains stable across different test splits.

References

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

J. D. Angrist and V. Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2020.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

J. A. Bagnell. Robust supervised learning. In *Proceedings of 20th National Conference on Artificial Intelligence (AAAI)*, pages 714–719. American Association for Artifical Intelligence, 2005.

P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

P. Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

F. Bunea, Y. She, and M. H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

A. Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, 2016.

C. Cheng, Y. Wei, and Y. Chen. Tackling small eigen-gaps: fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Transactions on Information Theory*, 67(11):7380–7419, 2021.

A. M. Christgau and N. R. Hansen. Efficient adjustment for complex covariates: Gaining efficiency with dope. *arXiv preprint arXiv:2402.12980*, 2024.

R. Christiansen, N. Pfister, M. Jakobsen, N. Gnecco, and J. Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(10):6614–6630, 2022.

K. L. Chung. *A course in probability theory.* Academic Press, 2001.

F. Deutsch. *Best approximation in inner product spaces*, volume 7. Springer, 2001.

F. Dunker. Adaptive estimation for some nonparametric instrumental variable models with full independence. *Electronic Journal of Statistics*, 15(2):6151–6190, 2021.

F. Dunker, J.-P. Florens, T. Hohage, J. Johannes, and E. Mammen. Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics*, 178:444–455, 2014.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:iii–115, 1944.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, pages 2029–2037. PMLR, 2018.

M. E. Jakobsen and J. Peters. Distributional robustness of K-class estimators and the PULSE. *The Econometrics Journal*, 25(2):404–432, 2022.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, pages 5815–5826, 2021.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

I. Loh. Nonparametric identification and estimation with discrete instruments and regressors. *Journal of Econometrics*, 235(2):1257–1279, 2023.

M. Londschien, M. Burger, G. Rätsch, and P. Bühlmann. Domain generalization and adaptation in intensive care with anchor regression. *arXiv preprint arXiv:2507.21783*, 2025.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

S. Magliacane, T. Van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems (NeurIPS)*, 31, 2018.

E. Mammen, C. Rothe, and M. Schienle. Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, 40(2):1132 – 1170, 2012.

N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

W. K. Newey, J. L. Powell, and F. Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

S. Ng and J. Pinkse. Nonparametric-two-step estimation of unknown regression functions when the regressors and the regression error are not independent. *Cahier de recherche*, 9551, 1995.

R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

J. Pearl. *Causality*. Cambridge University Press, New York, USA, 2nd edition, 2009.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

N. Pfister and J. Peters. Identifiability of sparse causal effects using instrumental variables. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 1613–1622, 2022.

PRISM. Prism climate data. `https://prism.oregonstate.edu`, 2020. Data created Dec 2022, accessed Sep 2023.

A. Puli, L. H. Zhang, E. K. Oermann, and R. Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2021.

J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

G. C. Reinsel and R. P. Velu. *Multivariate reduced-rank regression: Theory and Applications*, volume 136 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

S. Saengkyongam, L. Henckel, N. Pfister, and J. Peters. Exploiting independent instruments: Identification and distribution generalization. In *International Conference on Machine Learning (ICML)*, 2022.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

X. Shen, P. Bühlmann, and A. Taeb. Causality-oriented robustness: exploiting general noise interventions in linear structural causal models. *Journal of the American Statistical Association*, 0(online):1–20, 2025.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244, 2000.

A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018.

J. H. Stock and M. W. Watson. Chapter 10: Forecasting with Many Predictors. Handbook of Economic Forecasting, pages 515–554. Elsevier, 2006.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

K. Tsai, S. R Pfohl, O. Salaudeen, N. Chiou, M. Kusner, A. D'Amour, S. Koyejo, and A. Gretton. Proxy methods for domain adaptation. *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 238, 2024.

J. W. Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems (NeurIPS)*, 34, 2021.

M. J. Wainwright. *High-dimensional statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.