Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Learning When the Concept Shifts: Confounding, Invariance, and Dimension Reduction

Kulunu Dharmakeerthi, YoonHaeng Hur & Tengyuan Liang

To cite this article: Kulunu Dharmakeerthi, YoonHaeng Hur & Tengyuan Liang (2026) Learning When the Concept Shifts: Confounding, Invariance, and Dimension Reduction, Journal of the American Statistical Association, 121:554, 1000-1012, DOI: 10.1080/01621459.2026.2656457

To link to this article: https://doi.org/10.1080/01621459.2026.2656457

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2026, VOL. 121, NO. 554, 1000–1012: Special Issue on Statistical Science in Artificial Intelligence
https://doi.org/10.1080/01621459.2026.2656457

Learning When the Concept Shifts: Confounding, Invariance, and Dimension Reduction

Kulunu Dharmakeerthi[a], YoonHaeng Hur[a], and Tengyuan Liang[b]

[a]Department of Statistics, The University of Chicago, Chicago, IL; [b]Booth School of Business, The University of Chicago, Chicago, IL

ABSTRACT

Practitioners often face the challenge of deploying prediction models in new environments with shifted distributions of covariates and responses. With observational data, such shifts are often driven by unobserved confounding, and can in fact alter the concept of which model is best. This article studies distribution shifts in the domain adaptation problem with unobserved confounding. We postulate a linear structural causal model to account for endogeneity and unobserved confounding, and we leverage exogenous invariant covariate representations to cure concept shifts and improve target prediction. We propose a data-driven representation learning method that optimizes for a lower-dimensional linear subspace and a prediction model confined to that subspace. This method operates on a non-convex objective—that interpolates between predictability and stability—constrained to the Stiefel manifold, using an analog of projected gradient descent. We analyze the optimization landscape and prove that, provided sufficient regularization, nearly all local optima align with an invariant linear subspace resilient to distribution shifts. This method achieves a nearly ideal gap between target and source risk. We validate the method and theory with real-world datasets to illustrate the tradeoffs between predictability and stability. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received August 2024
Accepted March 2026

KEYWORDS
Concept shift; Distribution shift; Invariance; Representation learning; Structural causal model; Unobserved confounding

1. Introduction

Practitioners often deploy predictive models in new environments where the covariate and response distributions have shifted. Consider two observational datasets collected from different environments: a source environment used for training and a target environment where the model is deployed. With observational datasets, a model maximizing prediction accuracy in the source environment may experience a substantial drop in accuracy in the target environment, often due to unobserved confounding factors lurking in the environments. Such confounding can fundamentally alter the data distribution across environments and, in turn, alter the underlying concept of which statistical model is best.

A predictive model that generalizes well to new datasets is the statistical ideal, yet theory and methodology weaken when the environment underpinning the dataset shifts. Recently, a growing literature in causal inference has shown the utility of comparing datasets from distinct environments to identify causal relations (Peters, Bühlmann, and Meinshausen 2016; Heinze-Deml, Peters, and Meinshausen 2017; Pfister, Bühlmann, and Peters 2019; Arjovsky et al. 2019). At the same time, a body of works in machine learning (ML) and artificial intelligence (AI) has been tackling prediction under distribution shifts, leading to a rich field called domain adaptation (Ben-David et al. 2006, 2010; Blitzer, McDonald, and Pereira 2006; Blitzer et al. 2007; Tzeng et al. 2014; Long et al. 2015; Ganin et al. 2016). While both lines of research aim to address the problem of distribution shifts, they use two distinct notions of stability under different problem settings, which we discuss below.

Causal Stability. Consider a covariate-response pair (X, Y) and its joint probability distribution $\mathcal{P}_\varepsilon(X, Y)$ that differs across environments ε. We consider two environments $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$ corresponding to source and target, respectively. Due to the presence of unobserved confounding factors, both the *covariate distribution* $\mathcal{P}_\varepsilon(X)$, the marginal distribution of X, and the *conditional concept* $\mathcal{P}_\varepsilon(Y|X)$, the conditional distribution of Y given X, could be shifted and altered by the change in the environment. The principle of independent mechanisms in causality (Peters, Janzing, and Schölkopf 2017) points to a plausible source of invariance: the mechanism producing the effect given the cause should remain unchanged across environments. This principle can be made practical. For instance, the invariant risk minimization work (Peters, Bühlmann, and Meinshausen 2016; Heinze-Deml, Peters, and Meinshausen 2017; Pfister, Bühlmann, and Peters 2019; Rothenhäusler et al. 2021) seeks to identify a causally "stable" representation, that is, a map $X \mapsto \phi(X)$ such that $\mathcal{P}(Y|\phi(X))$, the conditional concept given $\phi(X)$, stays invariant across environments (hence no subscript ε). This line of work requires labeled data across environments, specifically, it assumes access to samples from both $\mathcal{P}_\mathcal{S}(X, Y)$ and $\mathcal{P}_\mathcal{T}(X, Y)$.

CONTACT Kulunu Dharmakeerthi *kulunud@uchicago.edu* Department of Statistics, The University of Chicago, George Herbert Jones Laboratory, Suite 222, 5747 S. Ellis Avenue, Chicago, IL 60637.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Distributional Stability. Unlike the setting above, domain adaptation considers the case where practitioners observe labeled covariate-response pairs from the source distribution $\mathcal{P}_{\mathcal{S}}(X, Y)$, but only *unlabeled* covariate samples from a target distribution $\mathcal{P}_{\mathcal{T}}(X)$. The seminal work of Ben-David et al. (2006) derived an upper bound on target risk, a key to domain adaptation. Given a representation $\phi : X \mapsto \phi(X)$, the upper bound involves a balancing act between (i) the source risk of the model under the representation ϕ, and (ii) a distance between probability distributions $\mathcal{P}_{\mathcal{S}}(\phi(X))$ and $\mathcal{P}_{\mathcal{T}}(\phi(X))$. Here, a notion of distributional "stability" arises. One prefers a representation $X \mapsto \phi(X)$ where the difference between $\mathcal{P}_{\mathcal{S}}(\phi(X))$ and $\mathcal{P}_{\mathcal{T}}(\phi(X))$ is small, holding source risk fixed. This notion of invariance is also natural: if the distribution shifts significantly, the model selected based on source data may suffer severely when extrapolating to target data.

This Article. Studying concept shifts induced by unobserved confounding—a central topic in causal inference with observational data—has received comparatively less attention in domain adaptation. We propose a structural causal model to address the challenge of domain adaptation in the presence of distribution shifts caused by unobserved confounding factors. This structural model unites notions of stability and the methods used to enforce them, developed independently in the ML/AI and statistics literatures.

Much like the role of instrumental variables in the presence of confounding, we find that leveraging an exogenous and invariant covariate representation can address both concept and covariate shifts lurking in the environments. Further, in the context of our structural model, notions of causal stability and distributional stability will align. With this understanding, we will motivate, from first principles, a stability-regularized risk minimization method that aims for invariance and stability when adapting to new observational domains.

1.1. A Structural Causal Model

Throughout the article, we postulate a Structural Causal Model (SCM) for unobserved confounding inspired by the idea of instrumental variables (Stock and Trebbi 2003; Pearl 2009). The crucial difference is that, in our model, these exogenous variables—resembling instruments—are latent and unobserved.

Definition 1 (Model). A pair of covariate and response, denoted as $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, are generated from mutually independent exogenous random variables $E \in \mathbb{R}^r, Z \in \mathbb{R}^k$ with $r, k \leq d$ and $U \in \mathbb{R}, W \in \mathbb{R}^d$ via the following structural causal model,

$$Y = \langle \beta^\star, X \rangle + \langle \gamma, E \rangle + U \,, \tag{1}$$

$$X = \Theta Z + \Delta E + W \,, \tag{2}$$

where $\beta^\star \in \mathbb{R}^d, \gamma \in \mathbb{R}^r, \Theta \in \mathbb{R}^{d \times k}$, and $\Delta \in \mathbb{R}^{d \times r}$ are unknown parameters; see Figure 1.

The *unobserved confounding* variable E models the environmental influence, whose distribution shifts across settings. Concretely, in domain adaptation, the distribution of E shifts from the source distribution $\rho_{\mathcal{S}}$ to the target distribution $\rho_{\mathcal{T}}$, where $\rho_{\mathcal{S}}, \rho_{\mathcal{T}}$ are probability distributions defined on $\mathbb{R}^r$. The *latent invariant* Z models a source of randomness that is not subject to environment shifts: a stable, exogenous structure shared across both source and target environments. It plays a similar role to an instrumental variable, with the notable difference that it is unobserved across environments. The *exogenous noise* U, W are mutually independent and mean-zero, and their distributions do not depend on the environment.

Observational Data and the Adaptation Problem. Consider the source and target environments $\mathcal{S}$ and $\mathcal{T}$. The joint distributions of (X, Y) under $E \sim \rho_{\mathcal{S}}$ and $E \sim \rho_{\mathcal{T}}$ are denoted as $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X, Y)$, with $\mathcal{P}_{\mathcal{S}}(X)$ and $\mathcal{P}_{\mathcal{T}}(X)$ denoting the covariate distributions. We use $\mathbb{E}_\varepsilon$ to denote the expectation w.r.t. the joint distribution under the environment $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$. For the expectation of random variables depending only on Z, U, W, we drop the subscript ε as it is invariant across environments.

The practitioner observes labeled covariate-response pairs jointly drawn from the source distribution $\mathcal{P}_{\mathcal{S}}(X, Y)$, but only unlabeled covariate data from a target distribution $\mathcal{P}_{\mathcal{T}}(X)$. Given the observational data, the practitioner aims to estimate a linear model based on $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$ that performs well in the target environment.

Confounding and Concept Shift. A few remarks follow for our model. First, suppose the dotted line connections in Figure 1 are removed. In that case, we arrive at the prototypical well-specified covariate shift setting: environment shifts only affect the covariate distribution, not the conditional distribution of Y given X. In this setting, without confounding, using (weighted) least squares to estimate a model from X to Y consistently is possible. Our model incorporating the dotted lines—capturing unobserved confounding—naturally extends the covariate shift setting. Second, since the unobserved confounding variable E lurking in the environment influences both X and Y, the least squares estimate is biased and unreliable when deploying to new environments. As we shall see in Proposition 1, the distribution shift in E under this structural model induces a *concept shift*, meaning that the best linear model depends on the environment. Finally, introducing the invariant, exogenous Z may remind readers of instrumental variables. However, in our domain adaptation problem, Z is unobserved; we merely postulate its existence. We develop a methodology that leverages invariance across environments to learn a stable, lower-dimensional linear subspace without directly observing Z.

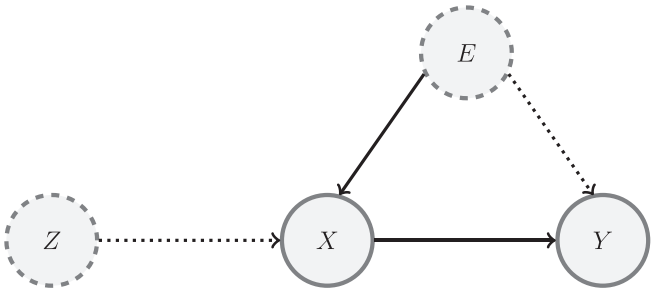


Figure 1. Diagram visualizing the model in Definition 1. The endogenous confounding variable E lurking in the environment and the exogenous invariant variable Z are both latent and unobserved.

1.2. A Motivating Example

We peek at data from an influential economic study by Tabellini (2010). Tabellini uses this data to explore the impact of historic social, economic and political development on the European economy in the late 90s. We will treat this data as observational, and consider the covariate-response pair:

$$X = \{\textit{Institutions}, \textit{Literacy}, \textit{Enrollment}, \textit{Culture}, \textit{Urbanization}\},$$
$$Y = \{\textit{Economic Output}\}\ .$$

Due to circumstances, suppose we only have access to data from rural regions (the source environment) and wish to construct a linear predictor for Y in historically urban areas (the target environment). How do we find an accurate predictor for *Economic Output* in the target environment? Considering our feature set, X, we would certainly expect distribution shifts in at least one coordinate—"Urbanization". Furthermore, Tabellini establishes "Institutions" as an instrumental variable that is independent of "Urbanization", and so, we would also expect at least one feature to be stable amid this environment shift. We can capture this setting in our data model; "Institutions" and "Urbanization" are natural candidates as components in Z and E, respectively. However, we do not assume this structure a priori, and hope to leverage the stability of certain features, like "Institutions", in a data-driven way. We will revisit this example in Section 5, after introducing the stability-regularized risk minimization method.

Notations. Let $\|v\|$ denote the standard Euclidean norm of a vector $v \in \mathbb{R}^d$. For any symmetric matrix $M \in \mathbb{R}^{d\times d}$, let $\lambda_{\max}(M), \lambda_{\min}(M)$ denote the largest, smallest eigenvalues of M, respectively; if M is positive semidefinite, let $M^{1/2}$ denote its matrix root and $\|v\|_M := \|M^{1/2}v\|$ for $v \in \mathbb{R}^d$. Also, $\|\cdot\|_{\mathrm{op}}, \|\cdot\|_{\mathrm{F}}$ are matrix norms (spectral, Frobenius, respectively). The Stiefel manifold is defined as $\mathrm{St}(d,\ell) := \{A \in \mathbb{R}^{d\times\ell} : A^\top A = I_\ell\}$. Finally, R_ε denotes the squared risk depending on environment $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$, namely, $R_\varepsilon(\beta) = \mathbb{E}_\varepsilon[(Y - \langle X, \beta\rangle)^2]$, where the expectation is w.r.t. the joint distribution $\mathcal{P}_\varepsilon(X, Y)$.

1.3. Organization and Contribution

The article studies domain adaptation in the presence of unobserved confounding. Unobserved confounding is a likely issue for a practitioner attempting to extrapolate with observational data. As seen in Figure 1, we extend the standard covariate shift setting to incorporate confounding factors lurking in the environment. These latent variables simultaneously influence covariate X and response Y.

In Section 2, we explore the shortcomings of vanilla risk minimization when both covariate distribution and "concept" are expected to shift under the linear SCM. As shown in Proposition 2, an invariant subspace in this model will unify concept stability and distributional stability, two distinct ideas in the literature a priori. Proposition 3 further establishes the necessity and benefit of leveraging a lower-dimensional, invariant subspace for predictive power.

Guided by insights into the invariant subspace, we solve the domain adaptation task in Section 3. Given distributional access to the unlabeled target, $\mathcal{P}_\mathcal{T}(X)$, and labeled data from the source, $\mathcal{P}_\mathcal{S}(X, Y)$, we seek a methodology gauging toward the inaccessible target risk by learning a subspace that balances predictability and stability/invariance. Concretely, we seek an estimator composed of a linear subspace representation, $V \in \mathbb{R}^{d\times\ell}$, and a low-dimensional ridge regressor, $\alpha \in \mathbb{R}^\ell$, jointly optimized according to the objective

$$V,\ \alpha := \arg\min_{V,\alpha} \mathbb{E}_\mathcal{S}[(Y - \langle X, V\alpha\rangle)^2] + \upsilon\|\alpha\|^2 + \tfrac{\eta}{2}\underbrace{\|V^\top(\mathbb{E}_\mathcal{T}[XX^\top] - \mathbb{E}_\mathcal{S}[XX^\top])V\|_{\mathrm{F}}^2}_{(*)}\ . \tag{3}$$

We define the composed estimator $\beta^{\upsilon,\eta} = V\alpha$.

Proposition 5 derives (3) as an upper bound on the target risk. Therefore, the $(*)$ term is a natural notion of invariance for domain adaptation under the linear SCM. Importantly, (3) is an actionable proxy for optimization based on the information available, and in Section 3.2, a practical first-order manifold optimization method is devised to navigate its non-convex landscape. By optimizing over the choice of regularization parameters η, υ in (3), the target risk can effectively be optimized. This empirical fact is demonstrated using real-world data examples in Section 5.

Moving to Section 4, we provide a theoretical characterization of the landscape of the non-convex manifold optimization. When the regularization parameter η is sufficiently large, we show that almost all local optima align well with an exogenous, invariant linear subspace and are resilient to distribution shifts driven by endogenous confounding factors. Indeed, denoting $\mathrm{Endo} \subset \mathbb{R}^d$ as the endogenous subspace corresponding to Δ in (2), we find the first-order stationary point $V \in \mathrm{St}(d,\ell)$ of (3) satisfies:

$$\max_{\mathbf{v}\in V, \mathbf{w}\in\mathrm{Endo}} \big(\cos\angle(\mathbf{v},\mathbf{w})\big)^6 \le O\left(\frac{1}{\upsilon\eta^2}\right)\ .$$

This alignment result leads to a stability bound between the target and source risks that directly addresses the domain adaptation problem. For any stationary point of (3),

$$R_\mathcal{T}(\beta^{\upsilon,\eta}) - R_\mathcal{S}(\beta^{\upsilon,\eta}) \le \inf_{\beta\perp\mathrm{Endo}} \underbrace{\{R_\mathcal{T}(\beta) - R_\mathcal{S}(\beta)\}}_{\text{oracle term}} + O\left(\frac{1}{\upsilon^{4/3}\eta^{2/3}}\right).$$

Namely, a predictive model using the learned lower-dimensional subspace could incur a nearly ideal gap between target and source risk, quantified by the oracle term. The theoretical results on invariance alignment and risk stability corroborate empirical observations in real-world datasets. Despite the difficult nature of non-convex manifold optimization, the practical first-order method often identifies good local optima for domain adaptation. Lastly, we also provide a finite sample error analysis to quantify the gap between the optimization and its plug-in estimation.

We provide a detailed literature review and discussion, deferred to Appendix A in the supplementary material, due to the space limit. All technical proofs are collected in Appendix C for the same reason.

2. Undesired Properties of Source Risk Minimization

In this section, we study the linear SCM in Definition 1, identify the undesired theoretical properties of vanilla source risk minimization, and further characterize the benefits of leveraging an invariant subspace to enforce stable learning and improve the target risk.

2.1. Concept Shift, Confounding, and Subspace Invariance

We first state the assumptions used in this section.

Assumption 1. In Definition 1, the exogenous variables Z, U, W are mean-zero, namely, $\mathbb{E}[Z] = 0, \mathbb{E}[U] = 0, \mathbb{E}[W] = 0$, and satisfy $\mathbb{E}[WW^\top] = \tau^2 I_d$ with $\tau > 0$.

For each environment $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$, classical risk minimization takes the following form:

$$\arg\min_{\beta \in \mathbb{R}^d} R_\varepsilon(\beta) = \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}_\varepsilon[(Y - \langle X, \beta \rangle)^2] \,. \tag{4}$$

In the SCM in Definition 1, when the endogeneity parameter $\gamma \neq 0$, the unobserved variable E lurking in the environment plays a confounding role in the relationship between X and Y. A natural question is: will there be a linear model simultaneously optimal for source and target risk minimization? The answer is no, thus, implying that the best model will be a moving target. We shall see that a shift in the distribution of the confounding variable $E \sim \rho_\mathcal{S}$ to $E \sim \rho_\mathcal{T}$ can lead to a stark difference in the best linear model across distributions. The following proposition formally defines the above phenomenon as a *concept shift*.

Proposition 1 (Concept Shift). Under Assumption 1, the risk minimization (4) admits a unique minimizer, denoted as β_ε for $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$. Suppose $[\Theta, \Delta] \in \mathrm{St}(d, k + r)$. Then, if $\mathbb{E}_\mathcal{S}[EE^\top] \neq \mathbb{E}_\mathcal{T}[EE^\top]$, there exists $\gamma \in \mathbb{R}^r$ such that $\beta_\mathcal{S} \neq \beta_\mathcal{T}$, namely, the best linear predictor (concept) shifts across the two environments for some endogeneity parameter γ.

Remark. It is immediate to show that under Assumption 1, the best linear model is well-defined and given as $\beta_\varepsilon = \beta^\star + (\mathbb{E}_\varepsilon[XX^\top])^{-1}\Delta\mathbb{E}_\varepsilon[EE^\top]\gamma$. The above result shows that the best linear model shifts as long as the environment changes. The curious reader may wonder whether the Bayes optimal model $\mathbb{E}_\varepsilon[Y|X]$ also changes. If we assume in addition that (E, Z, W) is drawn from a multivariate Gaussian and E is mean-zero, we can show $\mathbb{E}_\varepsilon[Y|X] = \langle \beta^\star + (\mathbb{E}_\varepsilon[XX^\top])^{-1}\Delta\mathbb{E}_\varepsilon[EE^\top]\gamma, X\rangle$, which matches the best linear model. Therefore, the Bayes optimal concept is also environment dependent. The simple derivation above also shows that the re-weighting method of Shimodaira (2000), based on the likelihood ratio of the covariate distributions, will not fix the concept shift issue. Due to endogeneity, a new methodology is needed.

In plain language, source risk minimization fails to recover a model that stays invariant across environments. The concept shift induced by the movement of the second moments of E reflects a bias that cannot be reconciled through traditional least squares. The root cause of the above concept shift is endogeneity. Inspired by the instrumental variable literature, we instead resort to exogenous linear subspaces of covariates invariant to environment shifts. The following proposition shows two desiderata—dimension reduction and invariance—can be achieved simultaneously.

Proposition 2 (Subspace Invariance). Under Assumption 1, let β_ε^Θ be the best linear predictor restricted to the linear subspace $\Theta \in \mathrm{St}(d, k)$ for each environment $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$:

$$\beta_\varepsilon^\Theta := \Theta\alpha_\varepsilon^\Theta, \text{ where } \alpha_\varepsilon^\Theta = \arg\min_{\alpha \in \mathbb{R}^k} \mathbb{E}_\varepsilon[(Y - \langle X, \Theta\alpha \rangle)^2] \,.$$

Suppose $[\Theta, \Delta] \in \mathrm{St}(d, k + r)$. Then, for any $\gamma \in \mathbb{R}^r$, we have $\beta_\mathcal{S}^\Theta = \beta_\mathcal{T}^\Theta$.

Remark. The above result shows that deconfounding is possible with an invariant subspace. In our model, the effect of the confounding variable, E, is restricted to the linear subspace Δ. Therefore, the covariate distribution projected to a lower-dimensional linear subspace, Θ, remains stable despite the distribution shifts from $\mathcal{P}_\mathcal{S}(X, Y)$ to $\mathcal{P}_\mathcal{T}(X, Y)$. Proposition 2 should be read in contrast to Proposition 1; risk minimization constrained to this "invariant" linear subspace resists arbitrary confounding effects parameterized by γ. And so, restricting to the invariant, exogenous space Θ yields stability in learned concepts, resilient to environment shifts.

In view of the decomposition, $Y = \langle \Theta^\top\beta^\star, Z\rangle + \langle \Delta^\top\beta^\star + \gamma, E\rangle$ + noise, the subspace model can be shown as $\beta_\mathcal{E}^\Theta \equiv \Theta\Theta^\top\beta^\star$. This decomposition separates variation in Y into invariant, exogenous component $\langle \Theta^\top\beta^\star, Z\rangle$ and environment dependent, endogenous component $\langle \Delta^\top\beta^\star + \gamma, E\rangle$. In particular, when $(I - \Theta\Theta^\top)\beta^\star = 0$, learning in the exogenous space consistently recovers the true signal; $\beta_\mathcal{E}^\Theta = \beta^\star$.

Proposition 2 only demonstrates the existence of an invariant linear subspace Θ. However, since the exogenous variable Z is unobserved, a data-driven approach to learning such a lower-dimensional linear subspace is still largely unclear. Motivated by invariance and stability, in Section 3, we will develop a methodology that inherits a manifold optimization problem to learn a lower-dimensional linear subspace. Later in Section 4, we will derive a theoretical characterization of the subspace alignment between the invariant Θ and a first-order stationary point of the manifold optimization. A target risk bound exploiting this alignment will also be established therein.

2.2. Target Risk Improvement via Invariance

For domain adaptation, the lingering question is whether leveraging the invariant subspace Θ can improve the target risk compared to $\beta_\mathcal{S}$, the vanilla source risk minimizer. This section provides a sufficient and necessary condition for the above question. The following proposition delineates when one can simultaneously achieve three desiderata: target risk improvement, invariance, and dimension reduction.

Proposition 3 (Target Risk Improvement). Denote $\Lambda_\varepsilon := \mathbb{E}_\varepsilon[EE^\top]$ for $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$. Under Assumption 1, and the assumption that $[\Theta, \Delta] \in \mathrm{St}(d, k + r)$ with $k + r = d$. Then, the linear

subspace predictor indexed by Θ based on source risk, that is, $\beta_{\mathcal{S}}^{\Theta}$, has a smaller target risk than the usual source risk minimizer $\beta_{\mathcal{S}}$, namely, $R_{\mathcal{T}}(\beta_{\mathcal{S}}^{\Theta}) < R_{\mathcal{T}}(\beta_{\mathcal{S}})$ if and only if

$$\|\Delta^\top \beta^\star + (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma\|^2_{\tau^2 I_r + \Lambda_{\mathcal{T}}} < \|(\tau^2 I_r + \Lambda_{\mathcal{S}})^{-1}\Lambda_{\mathcal{S}}\gamma - (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma\|^2_{\tau^2 I_r + \Lambda_{\mathcal{T}}} .$$

The right-hand side can be interpreted as the amount of concept shift, and the left-hand side can be interpreted as the sub-optimality of target risk due to the (invariant) subspace model. A special case helps to unpack the result. Consider $(I - \Theta\Theta^\top)\beta^\star = 0$. Then, the above expression reads

$$\|\Lambda_{\mathcal{T}}\gamma\|_{(\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}} < \|\tau^2 \underbrace{(\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}})}_{\text{environment shift}} (\tau^2 I_r + \Lambda_{\mathcal{S}})^{-1}\gamma\|_{(\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}} .$$

Conceptually, when the environment shift—parameterized by the second moment of the confounding variable E—is large enough, the subspace model strictly improves upon the vanilla source risk minimizer. When such a condition holds, target risk improvement, invariance, and dimension reduction can be achieved simultaneously.

Before concluding this section, we use the following simple example to elucidate the condition in Proposition 3.

Example. Consider the simple setting where $\Lambda_{\mathcal{S}} = \sigma_s^2 \cdot I_r$, $\Lambda_{\mathcal{T}} = \sigma_t^2 \cdot I_r$, and $\Delta^\top \beta^\star = c \cdot \gamma$ with a scalar $c \in \mathbb{R}$. Proposition 3 boils down to: $R_{\mathcal{T}}(\beta_{\mathcal{S}}^{\Theta}) < R_{\mathcal{T}}(\beta_{\mathcal{S}})$ if and only if

$$\left(c + \frac{\sigma_t^2}{\sigma_t^2 + \tau^2}\right)^2 < \left(\frac{\sigma_s^2}{\sigma_s^2 + \tau^2} - \frac{\sigma_t^2}{\sigma_t^2 + \tau^2}\right)^2 .$$

Fixing other parameters σ_t, σ_s, τ, this reduces to inequalities for the scalar parameter c; (i) *Target Rich Regime* with $\sigma_t > \sigma_s$: $\frac{-\sigma_s^2}{\sigma_s^2+\tau^2} + \frac{-2\tau^2(\sigma_t^2-\sigma_s^2)}{(\sigma_t^2+\tau^2)(\sigma_s^2+\tau^2)} < c < \frac{-\sigma_s^2}{\sigma_s^2+\tau^2}$; (ii) *Source Rich Regime* with $\sigma_s > \sigma_t$: $\frac{-\sigma_s^2}{\sigma_s^2+\tau^2} < c < \frac{-\sigma_s^2}{\sigma_s^2+\tau^2} + \frac{-2\tau^2(\sigma_t^2-\sigma_s^2)}{(\sigma_t^2+\tau^2)(\sigma_s^2+\tau^2)}$. In summary, the magnitude of confounding and the amount of environment shift altogether determine when invariance renders risk improvement.

So far, we have shown that leveraging subspace invariance allows one to surpass the limitations of source risk minimization in situations where the concept and the covariate shift. In the next section, we will propose a new domain adaptation method that learns a linear subspace resilient to shifting environments, making the insights obtained in this section actionable. The method trades off predictability and stability, using a lower-dimensional representation as a vehicle.

3. Methodology: Domain Adaptation via Manifold Optimization

The previous section briefly discusses the fundamental tradeoff between invariance and predictive performance. In general, an invariant subspace can improve upon the source risk minimizer but may not be optimal for the target risk in domain adaptation. This section contributes to delineating this tradeoff in the SCM defined in Definition 1. Concretely, we design a representation learning method for domain adaptation, parameterized by subspace projections, and navigate the tradeoff between representation invariance and risk minimization. Though the proposed manifold optimization is non-convex, we demonstrate that a standard first-order method works both empirically in Section 5 with real datasets, and later provably in Section 4, where we develop theory matching the empirics.

We motivate our main methodology in two ways: (i) as a surrogate for an upper bound on the target risk directly, and (ii) as a tradeoff to balance robustness/invariance and predictive performance.

3.1. A Surrogate for the Target Risk

In typical domain adaptation, labels from the target environment $Y \sim \mathcal{P}_{\mathcal{T}}(Y)$ are unavailable, and thus $R_{\mathcal{T}}(\beta)$ cannot be directly minimized. To circumvent this issue, we design a simple surrogate objective that does not require labels from the target environment, requiring only information about $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$. In the sequel we will denote $\Sigma_\varepsilon := \mathbb{E}_\varepsilon[XX^\top]$ for $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$. We start with a simple algebraic relation between the source and the target.

Proposition 4. Under Assumption 1 and $\Delta^\top\Delta = I_r$, we have

$$R_{\mathcal{T}}(\beta) = R_{\mathcal{S}}(\beta) + \big\langle \beta - (\beta^\star + \Delta\gamma), (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})(\beta - (\beta^\star + \Delta\gamma)) \big\rangle \quad \forall \beta \in \mathbb{R}^d . \tag{5}$$

The geometry of the target risk landscape involves a balance of two terms with clear distance interpretations. The first term $R_{\mathcal{S}}(\beta)$ is the source risk. The second term encodes a distance between β and the endogenous component, $\beta^\star + \Delta\gamma$, under a geometry determined by the covariance shift matrix $\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}$. The second term is not actionable based on data because $\beta^\star + \Delta\gamma$ is unknown. In the following proposition, we define an actionable objective, which serves as a surrogate upper bound for the target risk, for all β and $\beta^\star + \Delta\gamma$.

To make the presentation simple, we impose the following additional assumption.

Assumption 2 (Target Richer than Source). Let $\Lambda_\varepsilon := \mathbb{E}_\varepsilon[EE^\top]$ for $\varepsilon \in \{\mathcal{S}, \mathcal{T}\}$. Assume that $D := \Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}} \succ 0$.

Note that the assumption is for simplicity of presentation: we can state more general results by separating the positive and negative parts of $\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}}$. More importantly, we use Assumption 2 to focus on the interesting regime of distribution shift where the target distribution strictly dominates the source. This is when out-of-domain extrapolation can happen. Informally speaking, the Loewner order, $\Lambda_{\mathcal{T}} \succ \Lambda_{\mathcal{S}}$, indicates more variability present in the target domain.

With Assumption 2 in hand, an upper bound on target risk can be derived, which suggests a practical proxy for $R_{\mathcal{T}}(\beta)$. To elucidate the low-dimensional subspace dependence, we explicitly construct the estimator β to be a composite map of a lower-dimensional representation, parameterized by $V \in \mathbb{R}^{d\times\ell}$, and a linear model restricted to the representation, parameterized by $\alpha \in \mathbb{R}^\ell$.

Proposition 5 (Surrogate for Target). If Assumptions 1 and 2 hold with $\Delta^\top \Delta = I_r$, then for any $(V, \alpha) \in \mathbb{R}^{d\times\ell} \times \mathbb{R}^\ell$ and $\xi, \zeta > 0$, we have $R_{\mathcal{S}}(V\alpha) \le R_{\mathcal{T}}(V\alpha)$ and

$$R_{\mathcal{T}}(V\alpha) \le R_{\mathcal{S}}(V\alpha) + \tfrac{(1+\xi)\zeta}{2}\|\alpha\|^4 + \tfrac{(1+\xi)}{2\zeta}\|V^\top DV\|_{\mathrm{F}}^2 + (1+\tfrac{1}{\xi})\langle \beta^\star + \Delta\gamma, D(\beta^\star + \Delta\gamma)\rangle. \quad (6)$$

Proposition 5 highlights that the gap between target and source risks of a composite estimator $\beta = V\alpha$ can be controlled by two complexities: $\|\alpha\|$ and $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$. The term $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$ quantifies the alignment of the subspace V with the directions of covariance shift; in essence, a term describing stability for extrapolation. The term may also be interpreted as a quadratic maximum mean discrepancy (MMD) between covariate distributions in the linear subspace V. Indeed, under the map $\psi_V(X) = V^\top XX^\top V$,

$$\left\| V^\top(\mathbb{E}_{\mathcal{T}}[XX^\top] - \mathbb{E}_{\mathcal{S}}[XX^\top])V \right\|_{\mathrm{F}} = \sup_{M:\|M\|_{\mathrm{F}}\le 1} \mathbb{E}_{\mathcal{T}}[\langle M, \psi_V(X)\rangle] - \mathbb{E}_{\mathcal{S}}[\langle M, \psi_V(X)\rangle]\,. \quad (7)$$

Finally, note that the right-hand side of (6) is a regularized source risk minimization, regularized by ridge and stability penalties. More importantly, it is an actionable surrogate objective for domain adaptation, solely depending on $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$.

3.2. An Optimization Procedure on the Stiefel Manifold

Motivated by the upper bound on the target risk shown in Proposition 5, we define the following penalized objective to search for a linear subspace V, and a model restricted to that subspace parameterized by α,

$$F_{\upsilon,\eta}(V, \alpha) := \tfrac{1}{2}\left\{ R_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2 + \tfrac{\eta}{2}\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2 \right\}. \quad (8)$$

By optimizing over a linear subspace $V \in \mathbb{R}^{d\times\ell}$ that balances predictive power and stability, the above objective provides an explicit tradeoff. Note that the two-level optimization is convex in α for fixed V, where the inner optimum $\min_{\alpha\in\mathbb{R}^\ell} F_{\upsilon,\eta}(V, \alpha)$ is attained at

$$\alpha_V := (V^\top \mathbb{E}_{\mathcal{S}}[XX^\top]V + \upsilon I_\ell)^{-1} V^\top \mathbb{E}_{\mathcal{S}}[XY]\,. \quad (9)$$

We search for a representation confined to the Stiefel manifold $V \in \mathrm{St}(d, \ell)$. Putting things together, we arrive at the following optimization on the Stiefel manifold,

$$\min_{V\in\mathrm{St}(d,\ell)} \Phi_{\upsilon,\eta}(V) := \min_{V\in\mathrm{St}(d,\ell)} F_{\upsilon,\eta}(V, \alpha_V) \quad (10)$$

where α_V is defined in (9). The problem (10) is non-convex in V; however, we will derive provable results for any first-order stationary point of the landscape. Informally speaking, as we shall see in Section 4, under certain conditions, almost all local optima demonstrate good alignment with the invariant subspace.

Let us first introduce a skeletal implementation of a first-order manifold optimization method to find local optima of $\Phi_{\upsilon,\eta}(V)$. Then, we move to fill in the essential background on the geometry of the Stiefel manifold to rationalize the algorithm.

Algorithm 1 Optimization on Stiefel Manifold

Require: Initial point: $V_0 \in \mathrm{St}(d, \ell)$.
for $k = 0, 1, 2, \ldots$ **do**
Gradient: $G_k \leftarrow (I - V_kV_k^\top)((\Sigma_{\mathcal{S}}V_k\alpha_{V_k} - \mathbb{E}_{\mathcal{S}}[XY])\alpha_{V_k}^\top + \eta DV_kV_k^\top DV_k)$ as per (11).
Retraction: $V_{k+1}(t) := (V_k + t\cdot G_k)(I_\ell + t^2\cdot G_k^\top G_k)^{-1/2}$ for $t \in \mathbb{R}_+$ as per (12).
Line-Search: choose step-size $t = t_k$ via line-search, and set $V_{k+1} \leftarrow V_{k+1}(t_k)$.
end for

Geometry on the Stiefel Manifold. A few essential geometric items must be defined to understand the first-order method optimizing (10) detailed in Algorithm 1. We treat $\mathrm{St}(d, \ell)$ as an embedded sub-manifold in the vector space $\mathbb{R}^{d\times\ell}$ and introduce appropriate notions of tangent space, projection, and gradient. We provide a cursory overview of the Stiefel geometry for a self-contained treatment. A familiar reader may skip this part.

First, the tangent space at a point on a sub-manifold is intuitively the plane tangent to the sub-manifold at that point. The normal space is the corresponding orthogonal complement. The tangent space and the normal space to $\mathrm{St}(d, \ell)$ at V are given as $T_V\mathrm{St}(d, \ell) = \{Z \in \mathbb{R}^{d\times\ell} : V^\top Z + Z^\top V = 0\}$ and $(T_V\mathrm{St}(d, \ell))^\perp = \{VS : S = S^\top \in \mathbb{R}^{\ell\times\ell}\}$, respectively. We view the Stiefel manifold as a sub-manifold of $\mathbb{R}^{d\times\ell}$, with its tangent spaces inheriting the Euclidean metric (Frobenius norm) $\|\cdot\|_{\mathrm{F}}$ and inner product $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$. Then, projections of $\xi \in \mathbb{R}^{d\times\ell}$ on to the tangent and the normal spaces at V are given by $P_V(\xi) = (I_d - VV^\top)\xi + V\mathrm{skew}(V^\top\xi)$ and $P_V^\perp(\xi) = V\mathrm{sym}(V^\top\xi)$, respectively, where $\mathrm{skew}(A) = (A - A^\top)/2$ and $\mathrm{sym}(A) = (A + A^\top)/2$.

Finally, if F is a smooth function defined on $\mathbb{R}^{d\times\ell}$, and $\bar{F}$ is its restriction to the Riemannian sub-manifold $\mathrm{St}(d, \ell)$, the gradient of $\bar{F}$ is equal to the projection of the gradient of F onto $T_V\mathrm{St}(d, \ell)$, namely, $\mathrm{grad}\,\bar{F}(V) = P_V(\mathrm{grad}\,F(V))$. A rigorous treatment of the above can be found in (Absil, Mahony, and Sepulchre 2008, chap. 3 and 4).

Gradient Formula and Retraction. With the geometry understood, the Riemannian gradient of the objective (8) can be readily calculated.

Proposition 6 (Riemannian Gradient and Retraction). Consider $\bar{\Phi}_{\upsilon,\eta} : \mathrm{St}(d, \ell) \to \mathbb{R}$, the restriction of $\Phi_{\upsilon,\eta}$ to the Riemannian sub-manifold $\mathrm{St}(d, \ell)$. The Riemannian gradient of the objective (10) is

$$\mathrm{grad}\,\bar{\Phi}_{\upsilon,\eta}(V) = (I_d - VV^\top)\big((\Sigma_{\mathcal{S}}V\alpha_V - \mathbb{E}_{\mathcal{S}}[XY])\alpha_V^\top + \eta DVV^\top DV\big) \in T_V\mathrm{St}(d, \ell). \quad (11)$$

Moreover, the gradient formula for the objective (8) is the same for either the Stiefel or the Grassmann manifold.

For any $\xi \in T_V\mathrm{St}(d, \ell)$, the retraction map via matrix polar factorization is defined as

$$\mathrm{R}_V(\xi) = (V + \xi)(I_\ell + \xi^\top\xi)^{-1/2}\,. \quad (12)$$

Proposition 6 allows us to define first-order descent methods to optimize (10). For example, a simple line-search method on the Stiefel manifold updates the iterate as $V_{k+1} = \mathrm{R}_{V_k}(t_k \cdot G_k)$ for $G_k = \mathrm{grad}\, \bar{\Phi}_{\upsilon,\eta}(V_k)$, where R_{V_k} is the retraction map from (12) and t_k is a step-size. Now, all ingredients of Algorithm 1 readily unfold.

3.3. Stability, Predictability, and Connections to Modern ML

It was suggested in Section 3.2 that the penalized objective (10) is an instantiation of a robustness and accuracy tradeoff. Observe that under the linear SCM in Definition 1, two notions of stability/invariance coincide: optimizing over a *stable representation* amidst covariate shifts will align with searching for the *causal, invariant relationship*, stable across environments. We make the stability versus predictability tradeoff explicit now and, by doing so, relate the method to a class of procedures recently popularized in ML. Our procedure seeks a two-part solution: a linear projection V, enforcing a similarity between $\mathcal{P}_\mathcal{S}$ and $\mathcal{P}_\mathcal{T}$, and a linear estimator, α, built atop. Selecting penalization then translates to a balancing of stability and predictability. This is transparent once we decouple $F_{\upsilon,\eta}$ into loss and regularization,

$$F_{\upsilon,\eta}(V,\alpha) := \overbrace{\tfrac{1}{2}\mathbb{E}_\mathcal{S}[(Y - \langle X, V\alpha\rangle)^2]}^{\text{Predictivity}} + \overbrace{\tfrac{\upsilon}{2}\|\alpha\|^2 + \tfrac{\eta}{4}\left\|V^\top(\Sigma_\mathcal{T} - \Sigma_\mathcal{S})V\right\|_\mathrm{F}^2}^{\text{Stability}}.$$

The source *Predictivity* term corresponds to prediction under squared loss for the regression $Y \sim V^\top X$ confined to a subspace V; it extracts a linear relationship α between the label Y and a predictive subspace of the data $V^\top X$.

The *Stability* term quantifies two things. First, a notion of distributional stability determined by the penalty $\eta\left\|V^\top(\Sigma_\mathcal{T} - \Sigma_\mathcal{S})V\right\|_\mathrm{F}^2$ for the representation subspace V. The penalty can be interpreted as a distribution distance; see (7). Second, a standard ridge penalization quantifies the stability of the linear relationship α given the subspace V.

Conceptually, as $\eta \to \infty$, V captures the linear subspace over which the covariate second moments stay invariant. Moreover, such an invariant linear subspace corresponds to the contribution from the exogenous Z that identifies the invariant causal relationship. Indeed, $\Theta^\top\left(\mathbb{E}_\mathcal{T}[XX^\top] - \mathbb{E}_\mathcal{S}[XX^\top]\right)\Theta = 0$, therefore, provided $l > k$, $\Theta \in \mathbb{R}^{d\times k}$ is an element of the non-empty solution set. In this sense, the case when $\eta \to \infty$ coincides with the invariant estimator explored in Proposition 3. More generally, as we will see in Section 5, a choice of η in a certain interval may balance stability and source predictability to harness improvements in target risk, the primary object in domain adaptation.

It is easy to see how the framework can be generalized to deal with different notions of distributional discrepancy, loss function, and model class. Specifically, a family of methods is based on finding a stable and predictive representation $\phi\colon \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a suitable feature space. Let $\phi_\#\mu$ and $\phi_\#\nu$ denote the source and target covariate distributions pushed forward to the mapped space $\mathcal{Z}$. The goal is to find a representation $\phi\colon \mathcal{X} \to \mathcal{Z}$ and a composite predictor $g\colon \mathcal{Z} \to \mathcal{Y}$ simultaneously by solving $\min_{g,\phi} R_\mathcal{S}(g\circ\phi) + \eta \cdot d(\phi_\#\mu, \phi_\#\nu)$, where $d(\cdot,\cdot)$ is a suitable discrepancy between probability distributions on the mapped space $\mathcal{Z}$. Several choices for d have been proposed in the ML literature: Tzeng et al. (2014) and Long et al. (2015) use the kernel MMD with a suitable kernel on $\mathcal{Z}$, Ganin et al. (2016) uses a suitable hypothesis-based discrepancy introduced in Ben-David et al. (2006, 2010), and Shen et al. (2018) uses the Wasserstein-1 distance, to name a few.

4. Theory: Invariance Alignment and Risk Stability

We now study theoretical properties of the methodology proposed in Section 3: alignment with the invariant subspace and risk across environments. The first result in Theorem 1 characterizes the landscape of the non-convex manifold optimization. We show almost all local optima exhibit favorable alignment with the invariant linear subspace as long as the regularization parameter is sufficiently large. The second result speaks to domain adaptation. Building on the landscape result, we derive a stability bound between the source and target risk in Theorem 2, which provides theoretical underpinnings for the empirical phenomena observed in Section 5. On top of these, Theorem 3 analyzes a finite-sample estimation error by quantifying the gap between the optimization procedure (10) and its plug-in estimation.

4.1. Alignment with the Invariant Subspace

In Propositions 2 and 3, the invariant subspace Θ is shown to be advantageous. The following theorem proves that the data-driven manifold optimization method given in Section 3 can identify linear subspaces that are approximately orthogonal to the contribution of the endogenous, confounding variable E. We employ the canonical correlation (or angle) between linear subspaces to quantify the notion of approximate orthogonality. By dodging the endogenous subspace where E influences X, the learned subspace V approximately aligns with the invariant subspace—crucial for concept invariance and distributional stability.

Theorem 1 (Alignment with the Invariant Subspace). Consider the model in Definition 1 that satisfies Assumptions 1 and 2, and $\Delta^\top\Delta = I_r$. Recall that $D := \Sigma_\mathcal{T} - \Sigma_\mathcal{S}$ denotes the matrix of covariance shift. Suppose $V \in \mathrm{St}(d,\ell)$ satisfies the following conditions:

1. V is a first-order stationary point of the optimization (10) with $\mathrm{grad}\, \Phi_{\upsilon,\eta}(V) = 0$,
2. V is in the admissible set $\mathcal{S}_\Delta(\delta)$, where $\mathcal{S}_\Delta := \{V \in \mathrm{St}(d,\ell) : \|V^\top\Delta\|_{\mathrm{op}}^2 \le 1-\delta\}$.

Then, we have

$$\|V^\top\Delta\|_{\mathrm{op}}^6 \le \frac{\delta^{-1}\lambda_{\max}(\Sigma_\mathcal{S})\|\Sigma_\mathcal{S}^{-1/2}\mathbb{E}_\mathcal{S}[XY]\|^4}{\lambda_{\min}(\Delta^\top D\Delta)^4}\frac{1}{4\upsilon\eta^2}.$$

Remark. Theorem 1 characterizes the optimization landscape on the Stiefel manifold and should be interpreted when the

regularization parameter η is large. It has two noteworthy features. First, we establish a structural result for *all local optima* for the non-convex manifold optimization, which can be found efficiently in practice using Algorithm 1; we do not require the solution V to be the global optima. Second, the result operates even when the invariant subspace overlaps with the endogenous space, namely $\Theta^\top \Delta \neq 0$. In such a case, the learned subspace V will still be approximately orthogonal to Δ, in the sense that $\max_{\mathbf{v}\in V, \mathbf{w}\in\Delta} \cos\angle(\mathbf{v},\mathbf{w}) \leq \text{const.} \times \frac{1}{(\upsilon\eta^2)^{1/6}}$. Conceptually, the optimization procedure will identify a strict subspace inside the invariant subspace to approximately dodge the endogenous subspace Δ, so long as the stability regularization parameter η is large and ridge regularization parameter υ is not too small.

4.2. Stability and Target Risk Bound

Given the alignment with the invariant subspace, what remains to be shown is its implication for the domain adaptation problem, a task for this section. In that regard, the following theorem derives a stability bound between the target and the source risks. Recall the estimator $\beta^{\upsilon,\eta} = V\alpha_V$ with α_V as in (9). The following theorem shows the stability of any stationary point V of the non-convex Stiefel manifold optimization (10).

Theorem 2 (*Stability between Source and Target*). Consider the setting as in Theorem 1. Let $V \in \mathrm{St}(d,\ell)$ be any first-order stationary point of the optimization procedure (10) with regularization parameters υ, η, and $\beta^{\upsilon,\eta}$ be the corresponding linear subspace estimator restricted to V. For any $\epsilon > 0$, define

$$\mathsf{S}_{\epsilon,\delta} := (1+\epsilon^{-1})\delta^{-1/3}\lambda_{\max}(\Sigma_{\mathcal{S}})^{1/3} \times \frac{\lambda_{\max}(\Delta^\top D\Delta)}{\lambda_{\min}(\Delta^\top D\Delta)^{4/3}} \|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^{10/3},$$

the following generalization bound holds

$$R_{\mathcal{T}}(\beta^{\upsilon,\eta}) - R_{\mathcal{S}}(\beta^{\upsilon,\eta}) \leq (1+\epsilon)\cdot\langle\beta^\star + \Delta\gamma, D(\beta^\star+\Delta\gamma)\rangle + \mathsf{S}_{\epsilon,\delta}\cdot\frac{1}{(4\upsilon)^{4/3}\eta^{2/3}}.$$

Remark. Note that the first term is necessary. Even for the oracle V that is perfectly invariant to the environment shift, the oracle gap between the target and source risks equals $\langle\beta^\star+\Delta\gamma, D(\beta^\star+\Delta\gamma)\rangle$, in view of Proposition 4. The above result proves an approximate oracle to the gap between target and source risk, quantified by ϵ. As $\eta\to\infty$ (holding all else fixed), the stability bound $S_{\epsilon,\delta}\cdot\frac{1}{(4\upsilon)^{4/3}\eta^{2/3}}$ decreases with η, which confirms the numerical findings in Section 5. Conceptually, as η increases, the source risk $R_{\mathcal{S}}(\cdot)$ will increase as one tradesoff prediction power for stability; on the plus side, instability $R_{\mathcal{T}}(\cdot) - R_{\mathcal{S}}(\cdot)$ is reduced due to invariance.

4.3. Finite-Sample Error Analysis

In practice, we have access to labeled and unlabeled samples from the source and target domains, respectively. Accordingly, we approximate the optimization (10) by replacing $\Phi_{\upsilon,\eta}$ with the plug-in analog $\widehat{\Phi}_{\upsilon,\eta}$; see (13). A natural statistical question is whether the plug-in estimator $\widehat{\Phi}_{\upsilon,\eta}$ is a good approximation to the true objective function $\Phi_{\upsilon,\eta}$. To answer this, we bound the uniform deviation of the plug-in estimator $\widehat{\Phi}_{\upsilon,\eta}$ from the true objective function $\Phi_{\upsilon,\eta}$ over the Stiefel manifold $\mathrm{St}(d,\ell)$. This allows us to quantify the gap between the objective value $\Phi_{\upsilon,\eta}$ evaluated at the empirical minimizer, namely, a minimizer of $\widehat{\Phi}_{\upsilon,\eta}$, and the minimum of the population optimization (10).

Theorem 3. Let $\{(x_i,y_i)\}_{i=1}^n$ and $\{\tilde{x}_i\}_{i=1}^n$ be independent samples from $\mathcal{P}_{\mathcal{S}}(X,Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$, respectively. Suppose that the supports of $\mathcal{P}_{\mathcal{S}}(X,Y)$ and $\mathcal{P}_{\mathcal{T}}(X,Y)$ are contained in $\{(x,y)\in\mathbb{R}^d\times\mathbb{R} : \|x\|\leq M, |y|\leq M\}$ for some $M>0$. Let $\widehat{\Phi}_{\upsilon,\eta}$ be the plug-in estimator of $\Phi_{\upsilon,\eta}$ based on the samples: with the sample covariance matrices $\widehat{\Sigma}_{\mathcal{T}}, \widehat{\Sigma}_{\mathcal{S}}$, define

$$\widehat{\Phi}_{\upsilon,\eta}(V) := \min_{\alpha\in\mathbb{R}^\ell} \tfrac{1}{2}\{\tfrac{1}{n}\sum_{i=1}^n (y_i - \langle V\alpha, x_i\rangle)^2 + \upsilon\|\alpha\|^2 + \tfrac{\eta}{2}\|V^\top(\widehat{\Sigma}_{\mathcal{T}} - \widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}}^2\}. \quad (13)$$

Let $\widehat{V} \in \mathrm{St}(d,\ell)$ be a minimizer of $\widehat{\Phi}_{\upsilon,\eta}$, that is, $\widehat{V} \in \arg\min_{V\in\mathrm{St}(d,\ell)} \widehat{\Phi}_{\upsilon,\eta}(V)$. Then, for any $\delta\in(0,1)$, the following inequality holds with probability at least $1-\delta$:

$$\Phi_{\upsilon,\eta}(\widehat{V}) - \min_{V\in\mathrm{St}(d,\ell)} \Phi_{\upsilon,\eta}(V) \leq 9(M + \frac{M^3}{\upsilon})^2\sqrt{\frac{\log\frac{3}{\delta}}{n}} + \mathsf{C}_{M,\Sigma_{\mathcal{T}},\Sigma_{\mathcal{S}}}\cdot\eta\ell\left(\frac{\log\frac{6d}{\delta}}{n} + \sqrt{\frac{\log\frac{6d}{\delta}}{n}}\right),$$

where $\mathsf{C}_{M,\Sigma_{\mathcal{T}},\Sigma_{\mathcal{S}}}$ is a constant depending on M, $\Sigma_{\mathcal{T}}$, and $\Sigma_{\mathcal{S}}$.

The idea of the proof of Theorem 3 is to decompose the deviation $|\Phi_{\upsilon,\eta} - \widehat{\Phi}_{\upsilon,\eta}|$ into the deviation of the empirical risk from the population risk and the deviation of the empirical penalty term from the population penalty term. The first term on the right-hand side of the inequality of Theorem 3 corresponds to the bound on the deviation of the risk, while the other term bounds the deviation of the penalty term. The complete proof of Theorem 3 is provided in Appendix C in the supplementary material together with the exact form of the constant $\mathsf{C}_{M,\Sigma_{\mathcal{T}},\Sigma_{\mathcal{S}}}$.

5. Real-World Data Examples

Having established a framework for the first-order optimization of (10), we apply the method to several real world datasets exhibiting distribution shifts. The following examples naturally admit distinct environments with markedly different covariate distributions. We aim to verify that the causal structure we propose captures real-world relationships in data, and empirically explore the stability risk tradeoff motivated in the preceding section.

5.1. Revisiting the Motivating Example of Section 1.2

We begin by revisiting the example from the introduction. In his work, Tabellini argues *Economic Output* in the 1990s is causally related to various measures of social and economic prosperity. These complex causal relationships lead to a rich data setting for domain adaptation. We consider two environments—historically rural areas (source) and urban areas (target)—and

aim to create a linear predictive model for *Economic Output*. Tabellini's work suggests both covariate and concept shifts may be present. For example, historic *Urbanization* levels likely impacted other features (e.g., "Culture" in the 1990s), and the response variable, *Economic Output*.

A simple regression may be problematic if we want a linear predictor that performs well in both rural and urban areas (Proposition 1). The issue is exacerbated when we lack equal access to information across the urbanization spectrum. Suppose we have access to rural data (*Urbanization* $< 2\%$), but only limited labeled data from urban regions (*Urbanization* $> 10\%$). We can cast this problem in the framework of Figure 1—with X and Y as below, while we take *Urbanization* as an environment variable (a component of E):

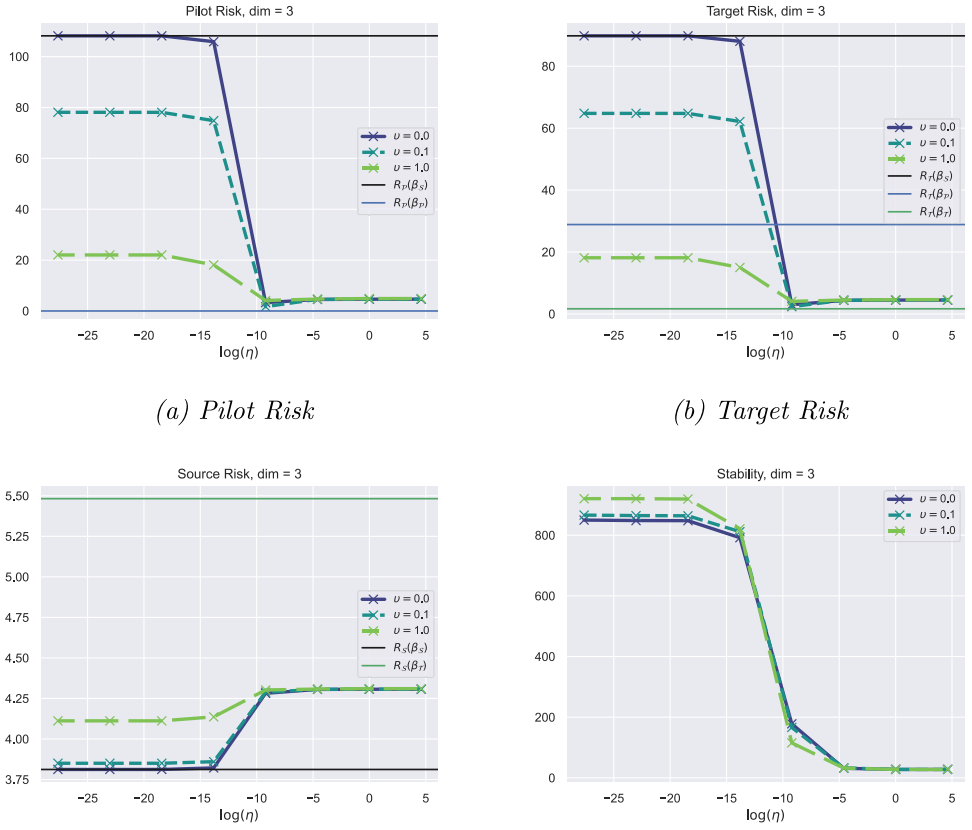
$$X = \{\textit{Institutions}, \textit{Literacy}, \textit{Enrollment}, \textit{Culture}, \textit{Urbanization}\},$$
$$Y = \{\textit{Economic Output}\}\ .$$

Borrowing from Tabellini's work, we would expect indicators of historical *Institutions* (pre 1850), *Literacy* (1880s), and *Enrollment* rates (1960s) to be minimally affected by *Urbanization* levels in the 1850s. Hence, we postulate the existence of a lower-dimensional variable Z with $k = 3$ that is invariant to the shift of *Urbanization*.

As our goal is prediction, not identification, we do not assume knowledge of any causal structure beyond $X \to Y$. Particularly, we do not suppose *Institutions* act as a valid instrument, and do not preempt its role as Z. Instead, we run our method, obtaining a linear subspace predictor $V\alpha_V$ by solving (10) for different regularization parameters υ, η and Stiefel dimension ℓ. For hyperparameter selection, and in line with our data constraints, we use a small amount of pilot data for validation ($\approx$ 20% of target data is labeled).

We find that $\ell = 3$ is an optimal lower-dimensional projection through pilot data validation. Figure 2 illustrates the performance of the obtained predictor with this specification. Figure 2 consists of 4 subplots: (a) the pilot risk, accuracy of the obtained predictor $V\alpha_V$ on a minuscule amount of labeled target data, (b) the target risk $R_{\mathcal{T}}(V\alpha_V)$, the primary objective that is inaccessible, (c) the source risk $R_{\mathcal{S}}(V\alpha_V)$, and (d) $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$, a quantifier for the stability/invariance.

While we have assumed no knowledge of the underlying causal structure, by navigating the stability-risk tradeoff, we find synergy with an instrumental variable perspective. With a larger stability parameter η, the lower dimensional subspace prioritizes *Institutions* as a stable feature. This is emphasized in Figure 3, where we optimize for a one-dimensional subspace and track the direction of the found projection with respect to each feature. While we do not claim that maximizing for stability recovers instrumental features, we certainly expect features unaffected by environmental influence to be prominent in the dominant Eigen-directions of V as $\eta \to \infty$. For this example, maximizing stability improves target predictive power. We will see this is



(a) Pilot Risk *(b) Target Risk*

(c) Source Risk *(d) Stability*

Figure 2. Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (10) for different values of (υ, η) with $d = 5$ and $\ell = 3$. The subplots (a)–(c) plot the risk of the estimator on Pilot, Target, and Source data, respectively. (d) displays $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$, a quantifier for the stability. In (a)–(c), the black line tracks the risk of $\beta_{\mathcal{S}}$, the blue line tracks Pilot Data OLS, and the green line tracks the risk of $\beta_{\mathcal{T}}$—a best possible oracle benchmark infeasible in practice.

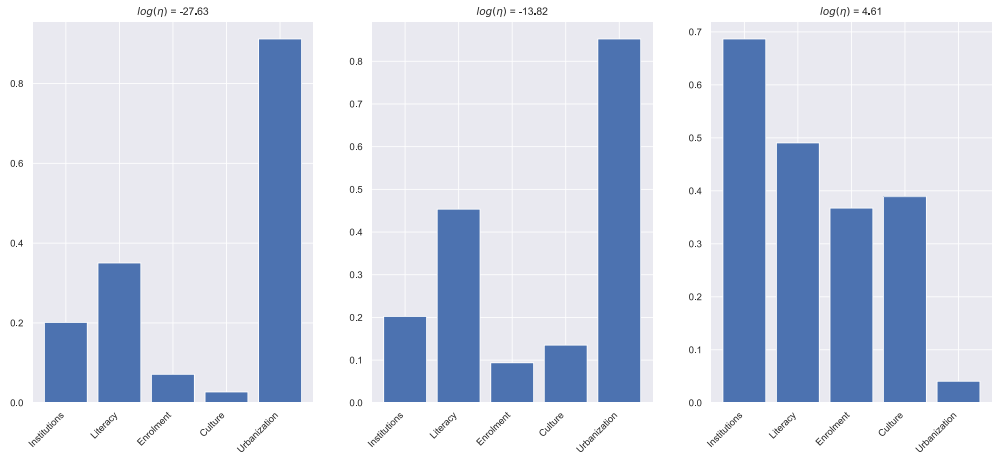


Figure 3. Results for $\upsilon = 0$ and $\ell = 1$ with $\log \eta \in \{-27.63, -13.82, 4.61\}$. We track the weighting placed on each feature by plotting the absolute value of each coefficient in the found $V \in \mathbb{R}^{d\times 1}$. As we increase η (from left to right), greater emphasis is placed on Institutions, the valid instrument.

not always the case. In general, aiming for target prediction involves navigating a tradeoff between source prediction and stability.

5.2. ML Predictions under Distribution Shifts

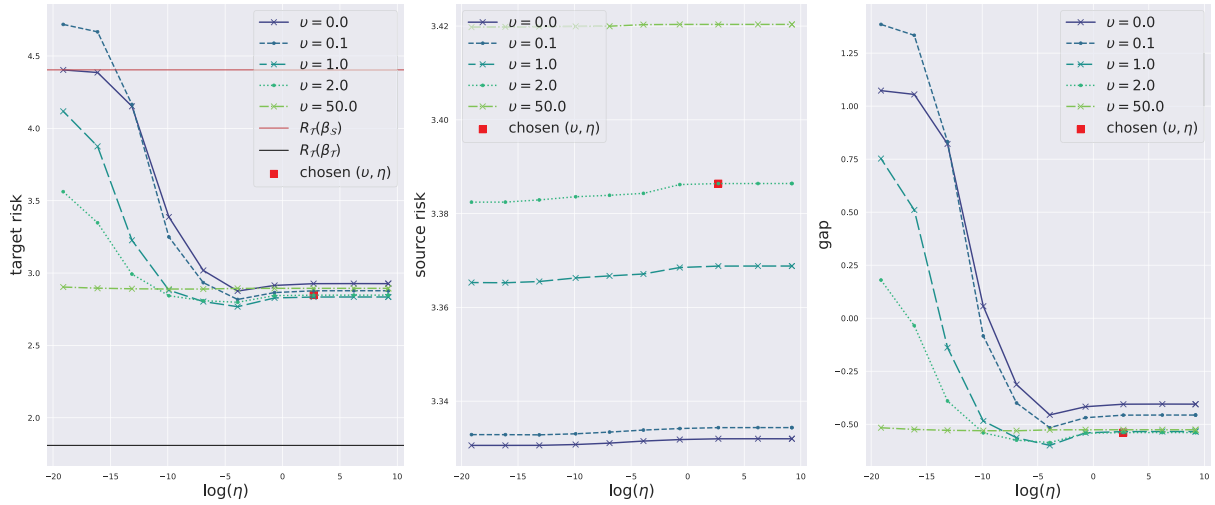
Next, we apply our method to three real-world datasets with less transparent causal relationships. These examples represent typical ML prediction challenges and provide a valuable benchmark for testing the robustness of our method in the absence of rigorously justified SEMs. Additionally, we introduce a data-driven heuristic for hyperparameter selection that does not depend on pilot data.

- *Forest Fires Data* Cortez and Morais (2007): The goal is to predict the burned area of forest fires in the northeast region of Portugal using various meteorological features. We choose seven covariates ($d = 7$): temperature, relative humidity, wind speed, precipitation, and three fire weather indices. We emulate seasonal shifts by considering two environments separated in time; the data corresponding to June, July, and August constitute the target, while we take the remaining data as the source.
- *Bike Sharing Data* Fanaee and Gama (2014): Obtained from the bikeshare system called Capital Bikeshare serving Washington, D.C., USA, the goal of this dataset is to predict hourly counts of bike rentals using features representing weather information. We postulate two different environments depending on the seasons, spring and fall for source and target, respectively. Here, we take a subset of this data by focusing on workdays and take six features ($d = 6$): hour (0 to 23), temperature, feeling temperature, humidity, wind speed, and an indicator for working day.
- *Wine Quality Data* Cortez et al. (2009): This exercise aims to predict the quality of wine using eleven physicochemical features ($d = 11$): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. We use the data corresponding to white and red wine as the source and target, respectively.

Unlike the example in Section 5.1, there are no well-studied models or instruments for these datasets. Accordingly, there is no clear choice for the hyperparameters: ℓ (Stiefel dimension), η (stability parameter) and υ (ridge parameter). It is recommended to use a small amount of pilot data for validation if available. However, when pilot data is unavailable and we have no prior knowledge of the causal structure, it is inevitable to invoke a certain rule of thumb for hyperparameter selection.

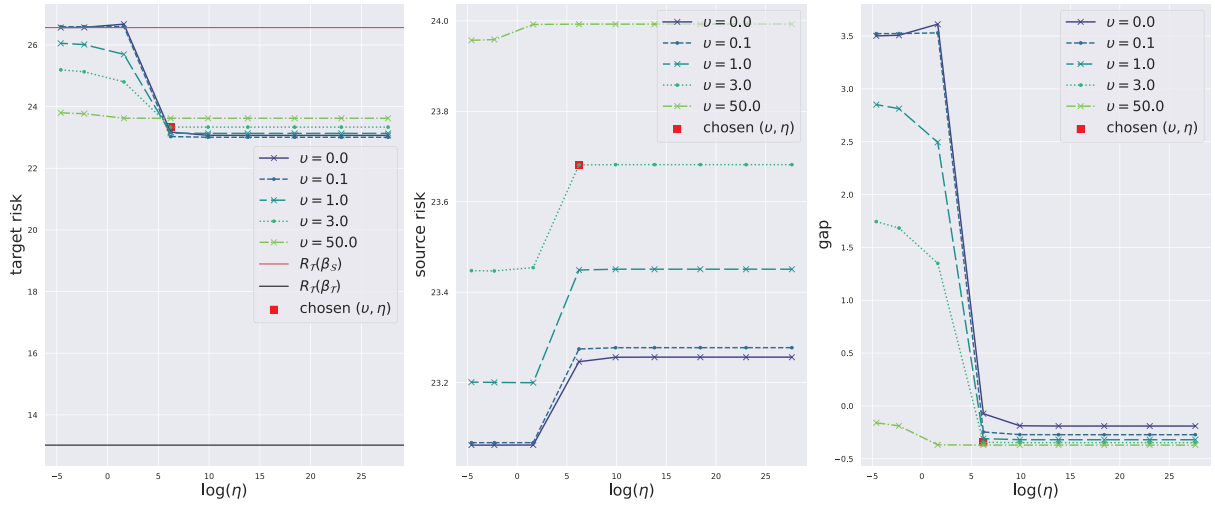
Data-Driven Hyperparameter Selection. For such cases, we provide the following guidelines for hyperparameter selection. First, a principled way to choose the Stiefel dimension ℓ is to use principal component analysis (PCA). To find a stable subspace, it is reasonable to apply PCA to the pooled covariates from both source and target and choose the number of principal components that explain most of the variance; concretely, we recommend the smallest ℓ such that the cumulative explained variance ratio exceeds a certain threshold, say 0.9. For the regularization parameters η, υ, we recommend scaling them by balancing the three terms in (8). For the ridge parameter υ, we first standardize the source covariates and then choose υ based on the ratio between the minimum source risk and the squared norm of the source risk minimizer, namely, $\frac{R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|\beta_{\mathcal{S}}\|^2}$. While letting $\upsilon \approx \frac{R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|\beta_{\mathcal{S}}\|^2}$ balances the source risk and the regularization term, we recommend a smaller value to avoid excessive regularization, say $\upsilon \approx \frac{0.1\cdot R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|\beta_{\mathcal{S}}\|^2}$. For the stability parameter η, we first find $V \in \mathrm{St}(d,\ell)$ that minimizes (8) without the prediction term, namely, $\min_{V\in\mathrm{St}(d,\ell)} \|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2$. Then, we choose η by balancing the obtained stability term and the minimum source risk, namely, $\eta \approx \frac{2R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|V^\top(\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2}$. These data-driven guidelines, albeit heuristic, produce robust empirical performances, as shown below.

Results. Figures 4–6 illustrate the performance of the obtained predictor for the forest fires, bike sharing, and wine quality data, respectively. Each figure consists of three subfigures: (a) the target risk $R_{\mathcal{T}}(V\alpha_V)$, the primary objective that is inaccessible, (b) the source risk $R_{\mathcal{S}}(V\alpha_V)$, and (c) the difference $R_{\mathcal{T}}(V\alpha_V) - R_{\mathcal{S}}(V\alpha_V)$, a quantifier for the stability/invariance. For (a), the solid red horizontal line shows $R_{\mathcal{T}}(\beta_{\mathcal{S}})$, the target



(a) Target Risk (b) Source Risk (c) Target Risk - Source Risk

Figure 4. Forest Fires Data: Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (10) for different values of (υ, η) with $d = 7$ and $\ell = 5$. (a) plots the risk on target dataset $R_{\mathcal{T}}(V\alpha_V)$, where the solid red horizontal line shows $R_{\mathcal{T}}(\beta_{\mathcal{S}})$ and the solid black line shows $R_{\mathcal{T}}(\beta_{\mathcal{T}})$. (b) shows the risk on the source dataset $R_{\mathcal{S}}(V\alpha_V)$. (c) plots the difference $R_{\mathcal{T}}(V\alpha_V) - R_{\mathcal{S}}(V\alpha_V)$.



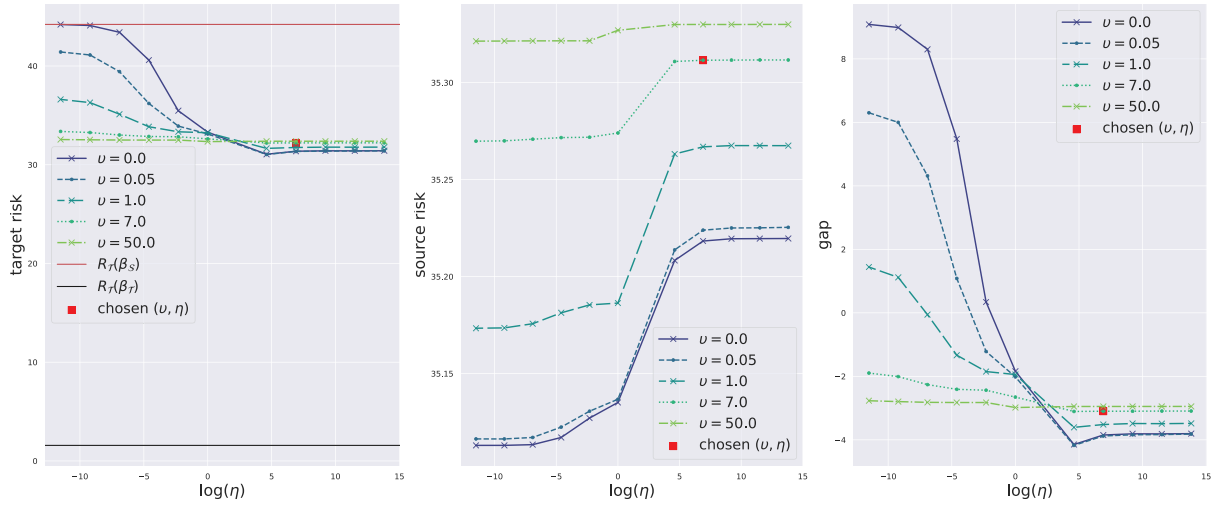
(a) Target Risk (b) Source Risk (c) Target Risk - Source Risk

Figure 5. Bike Sharing Data: Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (10) with $d = 6$ and $\ell = 5$. The subplots (a)–(c) plot the same quantities as in Figure 4.

risk of the vanilla source risk minimizer $\beta_{\mathcal{S}}$, and the solid black line shows $R_{\mathcal{T}}(\beta_{\mathcal{T}})$, the target risk of the target risk minimizer $\beta_{\mathcal{T}}$ if we were given the labeled access to the target distribution $\mathcal{P}_{\mathcal{T}}(X, Y)$—a best possible oracle benchmark infeasible in practice.

From Figures 4(a), 5(a), and 6(a), we can see that for each υ, there is a range of η such that the resulting target risk, $R_{\mathcal{T}}(V\alpha_V)$, is smaller than the target risk of the source risk minimizer $R_{\mathcal{T}}(\beta_{\mathcal{S}})$ (visually below the red solid line). This indicates that the proposed procedure can improve the target risk by balancing the stability and the predictive power of the estimator for certain combinations of the parameters υ and η. Notably, for the wine quality data, we can see improvement over the source risk minimizer $\beta_{\mathcal{S}}$ for any pair (υ, η), where we also achieve significant dimension reduction $\frac{\ell}{d} \approx 0.64$. Meanwhile, for the forest fires and bike sharing data, certain combinations of (υ, η) did not improve the target risk over the source risk. Based on the aforementioned hyperparameter selection guidelines, we obtain $(\upsilon, \eta) \approx (2, 15)$ for the forest fires data, $(\upsilon, \eta) \approx (3, 500)$ for the bike sharing data, and $(\upsilon, \eta) \approx (7, 1000)$ for the wine quality data. For these choices of hyperparameters, we find that the target risk is improved over the source risk for all datasets.

Meanwhile, we observe the non-monotonic relationship between target risk and η for fixed υ, which is more pronounced in Figures 4(a) and 6(a). This U-shaped behavior as a function of η results from the conflicting monotone behaviors of source risk and the risk gap: source risks shown in Figures 4(b) and 6(c) monotonically increase as η increases, while the risk gap shown in the corresponding (c) is almost monotone decreasing, which is explored in Theorem 2. Figures 4(c), 5(c), and 6(c) numerically validate the stability bound in Theorem 2: as η increases, the risk gap tends to decrease in most cases.



(a) Target Risk (b) Source Risk (c) Target Risk - Source Risk

Figure 6. Wine Quality Data: Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (10) with $d = 11$ and $\ell = 7$. The subplots (a)–(c) plot the same quantities as in Figure 4.

6. Conclusion

This article investigates domain adaptation for observational data while addressing the challenge of unobserved confounding. Unobserved confounding complicates domain adaptation by (i) altering the optimal statistical models across environments and (ii) extrapolating the model to the unseen domain of covariates. To address these challenges, we propose a causal model for distribution shifts in the presence of confounding. We highlight the limitations of traditional source risk minimization and the advantages of using an invariant subspace to stabilize learning and reduce target risk.

Methodologically, we introduce a domain adaptation algorithm that learns a representation through a lower-dimensional projection and tackles a non-convex manifold optimization problem using Riemannian optimization techniques. This algorithm acts as a stability-regularized source risk minimization that aligns with existing literature on stability and predictability in domain adaptation. Theoretically, we analyze the non-convex manifold optimization landscape and establish guarantees for nearly all local optima. With proper regularization, local optima from first-order Riemannian optimization will converge to an invariant linear subspace resistant to concept and covariate shifts. We also prove a target risk bound, showing that a predictive model derived from the learned subspace can achieve a nearly ideal gap between target and source risks.

The low-dimensional projection our method seeks will implicitly balance invariance and source domain predictive power to minimize target risk. This will be governed by hyperparameters (η, υ) and Stiefel dimension ℓ. Therefore, in practice, the choice of these hyperparameters is crucial. Hyperparameter selection guidelines in Section 5.2 provide actionable heuristics, which we verify to be effective from the examples in Section 5, but we believe more systematic hyperparameter selection rules could be more helpful. We leave this for future work. The current method is based on linear projections and linear regression; see Appendix B in the supplementary material for further discussions on the testability of the main assumption and comparison to the usual instrumental variable setting. It would be interesting to extend the proposed framework to partially linear or nonlinear models for full generality, as well as to the setting with multiple source environments. We leave these as future directions.

In summary, our method is a data-driven approach that extends the utility of instrumental variables for identification problems to domain adaptation by automatically finding the stable representation. Amid increasing interest in ML/AI, our method shows how to use traditional statistical/econometric techniques for data discovery in the AI context by properly designing automated procedures.

Supplementary Materials

The supplementary material discusses related literature, provides a discussion of the assumptions, and contains proofs of all theoretical results presented in the paper.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was supported by the NSF Career Award (DMS-2042473) and the Wallman Society of Fellows from the University of Chicago.

ORCID

YoonHaeng Hur http://orcid.org/0000-0001-6308-8896

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2008), *Optimization Algorithms on Matrix Manifolds*, Princeton: Princeton University Press. [1005]

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019), "Invariant Risk Minimization," arXiv preprint arXiv:1907.02893. [1000]

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010), "A Theory of Learning from Different Domains," *Machine Learning*, 79, 151–175. [1000,1006]

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2006), "Analysis of Representations for Domain Adaptation," in *Advances in Neural Information Processing Systems*. [1000,1001,1006]

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. (2007), "Learning Bounds for Domain Adaptation," in *Advances in Neural Information Processing Systems*. [1000]

Blitzer, J., McDonald, R., and Pereira, F. (2006), "Domain Adaptation with Structural Correspondence Learning," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 120–128, Association for Computational Linguistics. [1000]

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling Wine Preferences by Data Mining from Physicochemical Properties," *Decision Support Systems*, 47, 547–553. [1009]

Cortez, P., and Morais, A. (2007), "A Data Mining Approach to Predict Forest Fires Using Meteorological Data," in *Portuguese Conference on Artificial Intelligence*. [1009]

Fanaee-T. H., and Gama, J. (2014), "Event Labeling Combining Ensemble Detectors and Background Knowledge," *Progress in Artificial Intelligence*, 2, 113–127. [1009]

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016), "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, 17, 1–35. [1000,1006]

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2017), "Invariant Causal Prediction for Nonlinear Models," *Journal of Causal Inference*, 6, 20170016. [1000]

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015), "Learning Transferable Features with Deep Adaptation Networks," in *International Conference on Machine Learning*. [1000,1006]

Pearl, J. (2009), *Causality* (2nd ed.), Cambridge: Cambridge University Press. [1001]

Peters, J., Bühlmann, P., and Meinshausen, N. (2016), "Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals," *Journal of the Royal Statistical Society*, Series B, 78, 947–1012. [1000]

Peters, J., Janzing, D., and Schölkopf, B. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*, Cambridge, MA: The MIT Press. [1000]

Pfister, N., Bühlmann, P., and Peters, J. (2019), "Invariant Causal Prediction for Sequential Data," *Journal of the American Statistical Association*, 114, 1264–1276. [1000]

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2021), "Anchor Regression: Heterogeneous Data Meet Causality," *Journal of the Royal Statistical Society*, Series B, 83, 215–246. [1000]

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018), "Wasserstein Distance Guided Representation Learning for Domain Adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*. [1006]

Shimodaira, H. (2000), "Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function," *Journal of Statistical Planning and Inference*, 90, 227–244. [1003]

Stock, J. H., and Trebbi, F. (2003), "Retrospectives: Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives*, 17, 177–194. [1001]

Tabellini, G. (2010), "Culture and Institutions: Economic Development in the Regions of Europe," *Journal of the European Economic Association*, 8, 677–716. [1002]

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014), "Deep Domain Confusion: Maximizing for Domain Invariance," arXiv preprint arXiv:1412.3474. [1000,1006]