Margherita Lazzaretto

Exploiting invariance to learn under distribution shifts

PHD THESIS

this thesis has been submitted to the phd school of
the faculty of science, University of Copenhagen

Department of Mathematical Sciences
University of Copenhagen

March 2026

Margherita Lazzaretto
mala@math.ku.dk
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Thesis title:	Exploiting invariance to learn under distribution shifts
Supervisor:	Professor Niels Richard Hansen University of Copenhagen
	Professor Jonas Peters ETH Zürich
	Associate Professor Niklas Pfister University of Copenhagen (now Lakera)
Assessment Committee:	Professor Line Clemmensen (chair) University of Copenhagen
	Associate Professor Niki Kilbertus Technical University of Munich
	Associate Professor Søren Wengel Mogensen Copenhagen Business School
Date of Submission:	March 31, 2026
Date of Defense:	May 12, 2026
ISBN:	978-87-7629-232-4

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on 31 March 2026. It was supported by the Novo Nordisk Foundation (research grant 0069071).

a caframa.

Preface

The work presented in this thesis is the result of my time spent as PhD student at the Department of Mathematical Sciences of the University of Copenhagen, under the supervision of Niklas Pfister and Jonas Peters. The thesis contains two manuscripts investigating the importance of invariance in different statistical learning settings. Minor differences may occur between each chapter and the corresponding manuscript. Any typographical or mathematical errors are my own responsibility.

Acknowledgments

I would first like to thank Niklas and Jonas for giving me the opportunity to pursue this PhD and for their continued support, even from a distance. I really enjoyed your positive attitude during our meetings and always left feeling like I had learned something new. I am also grateful to Jonas for welcoming me during my research visit in Zürich, which was both professionally and personally a very valuable experience.

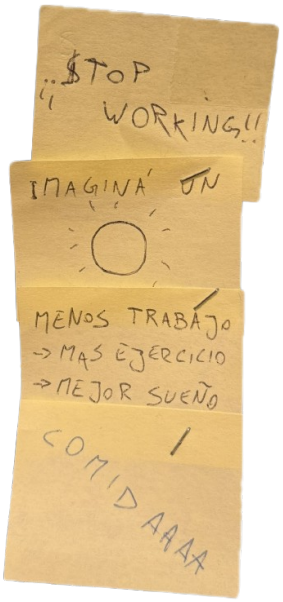
I extend my gratitude to Niels for taking on the role as supervisor during the last part of my PhD. Although we did not end up working together, I really appreciated your support, kindness and willingness to step in at this stage.

I am particularly thankful to all the wonderful people I had the luck to meet at the department every day, some of whom have become—dare I say—a bit more than colleagues. The list is long so I hope I don't forget too many, but thanks to Matt, Pedja, Alex, the other Alex, Shimeng, Phillip, Silvan, Harri, Francesco, Emma, Gabriel, Anton, Myrto, Jin, Nena, Frederik, Christine, Simon, Emanuele, Juan Carlos, Christian, Luigi, Nicola, Lucas, Ella, Jonas, Jeff for making my time in Copenhagen so memorable. An especially special thanks to Ulises, Gabriele, Tessa, and Cecilie for always being there for me in the times of cooking and crying and everything in between. This experience would not have been nearly as fun without you.

I would also like to thank the whole SfS group in Zürich, for being so genuinely welcoming from the very beginning, with a special mention to the G11 people.

Grazie mamma, papà e Pietro for always believing in me, maybe even a bit too much, and for your near-constant presence on video calls which made the distance feel a little bit shorter. A big thanks also to Cate, Reffo and Betta for being like family to me and for coming to visit me in Copenhagen several(!!) times.

Last but surely not least, grazie Gianluca for your love, patience and support, for believing in me more than I do, for involving me in your dreams and for never talking about math. I know it has not been an easy ride and I am very lucky to have you in my life, tvb.


¡¡STOP
WORKING!!
IMAGINÁ UN
MENOS TRABAJO
-> MÁS EJERCICIO
-> MEJOR SUEÑO
COMIDAAAA

Abstract

The concept of invariance plays a central role in robust statistics, serving as a modeling principle to isolate essential structure from variation induced by interventions, environmental heterogeneity, or time. This thesis investigates the advantages of leveraging structural invariance to guide learning in evolving environments. The unifying idea throughout this work is that the presence of invariance supports robust inference, facilitates dimension reduction, and enables adaptation. Chapter 1 elaborates on these ideas and places them within the literature on prediction under distribution shifts. In particular, it highlights strengths and limitations of invariance-based approaches for distribution generalization, thereby motivating the search for forms of invariance that remain useful for adaptation to new environments. It then extends this perspective to online learning and contextual bandits, and discusses limitations of the works presented in this thesis together with future research directions.

Chapter 2 introduces a framework that separates invariant and time-varying effects in the conditional mean of a response in a non-stationary linear system. The framework, called invariant subspace decomposition (ISD), provides a unified approach to both zero-shot prediction and time adaptation. We introduce a practical estimation procedure for the proposed decomposition, establish finite-sample guarantees, and show empirically that exploiting invariance can improve performance over methods that rely only on recent observations or that use invariance without adaptation.

Chapter 3 extends this perspective to stochastic non-stationary contextual bandits. Building on the ISD framework, it introduces an algorithm, ISD-linUCB, which uses offline data to learn an invariant component of the reward model and then performs online adaptation in a lower-dimensional residual subspace. This decomposition leads to regret bounds that scale with the residual dimension rather than the full problem dimension. We complement the theoretical analysis with simulation experiments showing that exploiting invariance can substantially improve online performance, especially in rapidly changing environments and when sufficient offline data is available.

Sammenfatning

Begrebet invarians spiller en central rolle i robust statistik og fungerer som et modelleringsprincip til at isolere essentiel struktur fra variation fremkaldt af interventioner, miljømæssig heterogenitet eller tid. Denne afhandling undersøger fordelene ved at udnytte strukturel invarians til at vejlede læring i foranderlige miljøer. Den gennemgående idé i dette arbejde er, at tilstedeværelsen af invarians understøtter robust inferens, letter dimensionsreduktion og muliggør tilpasning. Kapitel 1 uddyber disse idéer og placerer dem inden for litteraturen om prediktion under fordelingsskift. Særligt fremhæves styrker og begrænsninger ved invariansbaserede tilgange til fordelingsgeneralisering, hvilket motiverer søgningen efter former for invarians, der forbliver nyttige ved tilpasning til nye miljøer. Derefter udvides perspektivet til online læring og "contextual bandits", og begrænsningerne ved de artikler, der præsenteres i denne afhandling, diskuteres sammen med fremtidige forskningsretninger.

Kapitel 2 introducerer en metode, der adskiller invariante og tidsvarierende effekter i det betingede gennemsnit af en responsvariabel i et ikke-stationært lineært system. Metoden, kaldet "invariant subspace decomposition" (ISD), tilbyder en samlet tilgang til både "zero-shot" prediktion og tidstilpasning. Vi introducerer en praktisk estimeringsprocedure for den foreslåede dekomposition, etablerer garantier for endelige stikprøvestørrelser og viser empirisk, at udnyttelse af invarians kan forbedre ydeevnen sammenlignet med metoder, der udelukkende baserer sig på nylige observationer, eller som anvender invarians uden tilpasning.

Kapitel 3 udvider dette perspektiv til stokastiske ikke-stationære "contextual bandits". Med udgangspunkt i ISD-metoden introduceres algoritmen ISD-linUCB, som anvender offline data til at lære en invariant komponent af belønningsmodellen og derefter udfører online tilpasning i et laveredimensionalt residualt underrum. Denne dekomponering fører til fortrydelsesbegrænsninger, der skalerer med residualdimensionen snarere end den fulde problemdimension. Vi supplerer den teoretiske analyse med simuleringseksperimenter, der viser, at udnyttelse af invarians kan forbedre onlineydeevnen væsentligt, især i hurtigt skiftende miljøer og når tilstrækkeligt offline data er tilgængelige.

Contributions and Structure

This thesis is organized into three chapters. Chapter 1 is an introduction, which provides a broad overview on the ideas of domain generalization and invariance, motivating the work in the remainder of the thesis. It also discusses key limitations and future research directions. Chapters 2 and 3 correspond to a paper each. A brief overview of their content is given below.

Chapter 2 formulates a framework—Invariant Subspace Decomposition (ISD)—for separating invariant and time-varying effects in a non-stationary linear model and allowing a sample-efficient adaptation of the invariant component of the system through time. This chapter corresponds to the paper:

[ISD] [Lazzaretto et al., 2025]. M. Lazzaretto, J. Peters, and N. Pfister. Invariant subspace decomposition. *Journal of Machine Learning Research*, 26(95):1–56, 2025.

Chapter 3 proposes to extend the ISD framework to a stochastic contextual bandit setting. In particular, it analyzes how exploiting invariance helps improving dynamic regret bounds. It corresponds to the paper:

[IBCB] [Lazzaretto et al., 2026]. M. Lazzaretto, J. Peters, and N. Pfister. Invariance-based dynamic regret minimization. *arXiv preprint arXiv:2603.03843*, 2026.

Contents

1 Introduction

Often, predictive models are deployed in settings where the data-generating distribution is not stable: observations may come from different environments, evolve through time, or be collected under interventions and changing conditions. In such situations, standard prediction methods optimized for a single training distribution may fail to generalize. This raises the need for models that are robust to distribution shifts while remaining informative for the target task. The central motivation of this thesis is to understand how invariant structure can be isolated from heterogeneous data and exploited to obtain safe predictions, while still retaining enough flexibility to adapt when partial information about new environments becomes available.

The remainder of this introduction places the contributions of this thesis within the broader literature. It is not intended as a comprehensive review, but rather as a focused discussion of the main ideas and methodological perspectives that best contextualize the presented works. Section 1.1 provides an overview on the problem of prediction under distribution shifts. It focuses on the ideas of domain generalization and domain adaptation, and highlights the special role of time-varying heterogeneity. In Section 1.2 we present invariance as a central organizing principle for robust prediction. We address its relation to worst-case optimality and examine its limitations for adaptive prediction. This discussion motivates the invariant subspace decomposition (ISD) framework studied in [**ISD**]. Section 1.3 outlines the extension of these ideas to online learning and contextual bandits. Finally, in Section 1.4 we discuss some limitations of the presented works and possible future research directions.

1.1 Prediction under distribution shifts

Prediction is one of the core problems of statistics, aimed at generalizing knowledge from an observed sample to a larger population. Statistical reasoning behind classic prediction tasks builds on the assumption that the observed data and the target population come from the same underlying distribution [Hastie et al., 2009]. That is, suppose we observe n samples $(X_i, Y_i)_{i\in\{1,\dots,n\}} \overset{\text{iid}}{\sim} P^{\text{train}}$, where P^{train} denotes the joint (training) distribution of a vector of covariates (or predictors) $X \in \mathbb{R}^p$ and a response variable $Y \in \mathbb{R}$. Then, we can use these observations to learn a function $\hat{f}$ that predicts Y from X. For example,

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_{P^{\text{train}}}[(Y - f(X))^2], \tag{1}$$

for some function class $\mathcal{F}$. We would then like to predict an unobserved response Y^{test} from X^{test}, where $(X^{\text{test}}, Y^{\text{test}}) \sim P^{\text{test}}$ for some joint distribution P^{test} describing the

target population. If P^{test} equals P^{train}, using $\hat{f}$ to make such prediction is indeed a sensible choice.

However, if the independent and identically distributed (i.i.d.) assumption is violated or $P^{\text{test}} \neq P^{\text{train}}$, a solution to (1) may no longer capture the right statistical dependencies for accurate prediction. Heterogeneity arises naturally when data are gathered across different locations, experimental conditions, or time periods. For example, data may be collected in several environments, indexed by $e \in \mathcal{E}$, each inducing a different distribution P_e^{train}, and the distribution of the target population may itself differ from the observed training distributions, i.e., $P^{\text{test}} \notin \{(P_e^{\text{train}})_{e\in\mathcal{E}}\}$. Under such distribution shifts, a more appropriate learning objective than (1) is a function that minimizes the expected test error,

$$\arg\min_{f\in\mathcal{F}} \mathbb{E}_{P^{\text{test}}}[(Y - f(X))^2]. \tag{2}$$

Solving (2) can be particularly challenging if only complete samples from P^{train} are observed, while data from P^{test} are fully or partially missing. In particular, if the discrepancy between training and test distributions is entirely unrestricted, no meaningful prediction guarantee on P^{test} is possible. For (2) to be tractable, one must therefore make additional assumptions on how P^{train} relates to P^{test}. The problem of learning a function on training data that performs well on test data coming from a different distribution is widely studied in the literature in different forms under the names of distributional robustness, domain generalization or out of distribution generalization [see, for example, Huber, 1981, Ben-David et al., 2010, Christiansen et al., 2022].

1.1.1 Generalizing out of domain

We first consider the setting in which no observations from P^{test} are available. In this regime, successful generalization to unseen domains or future time points necessarily relies on assumptions about which aspects of the training distribution P^{train} are preserved under the test distribution P^{test}. Such generalization assumptions are generally not verifiable using observational data alone, as they state something about the test distribution without having access to samples from it. In particular, a generalization assumption describes the type of shifts that may be expected between training and test, and that we would like our model to be robust against. Thus, it often requires a trade-off: the more general the allowed shifts are, the more conservative the learned model may become, potentially at the expense of predictive accuracy.

A lot of effort in the literature has gone into designing assumptions that realistically capture the mechanisms responsible for distribution shifts, with the goal of allowing for generalization while maintaining as much predictive capability at test time as possible.

Unlike in the classical i.i.d. setting, observing data coming from different distributions $(P^{\text{train},e})_{e\in\mathcal{E}}$ is beneficial for generalization purposes. If we only observe data from a single distribution, inferring something about a different test distribution can be particularly hard unless strong prior knowledge about their discrepancy is available. Instead, heterogeneity in the training data allows to hypothesize about the types of shifts that one can expect in the test distribution in a more principled way. For example, if we can

detect structured changes among training distributions, it may be reasonable to assume that shifts in the test distribution will follow similar patterns.

In general, the exact test distribution P^{test} is unknown, but assuming that it shares some properties with the distributions of the observed data is equivalent to restricting our attention to a class of possible target distributions $\mathcal{P}$ that we expect to contain P^{test}. This means that in general it is not possible to find an exact solution to (2), and the best we can aim for is minimizing a worst-case loss of the form

$$\min_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(Y - f(X))^2]. \tag{3}$$

Different works in the literature have considered different characterizations of the set $\mathcal{P}$. One of the earliest examples of distributional robustness [Huber, 1981] considers $\mathcal{P}$ as a set of perturbations of the training distribution. This idea was later extended in the distributionally robust optimization (DRO) field, where $\mathcal{P}$ is usually defined as a neighbourhood of the training distribution i.e.,

$$\mathcal{P} = \{P \mid d(P^{\text{train}}, P) < \varepsilon\}, \tag{4}$$

for some some $\varepsilon > 0$ and a distance function d between distributions such as the Wasserstein distance [Mohajerin Esfahani and Kuhn, 2018] or other divergence metrics [Hu and Hong, 2013, Duchi et al., 2021]. This formulation allows for shifts whose magnitude is controlled by the value of ε in (4), but it is in general designed to admit only moderate deviations, or perturbations, between the training and test distributions.

Observing data from multiple training distributions makes it possible to loosen assumptions on how close the test distribution needs to be to the training one. In particular, the more heterogeneous the observed data are, the broader the class of test-time distribution shifts one can hope to guard against is. For example, a natural extension of (4) consists of assuming that P^{test} lies in the convex hull of the training distributions, so $\mathcal{P} = \text{Conv}((P^{\text{train},e})_{e \in \mathcal{E}})$. This is the case for example in group DRO [Sagawa et al., 2020] or in the work by Meinshausen and Bühlmann [2015] on maximin effects.

A key contribution to domain generalization comes from causal approaches [Pearl, 2009, Peters et al., 2017], where distribution shifts are modeled as interventions on an underlying causal system. Intervening on a variable usually refers to its active manipulation, and the corresponding interventional distribution reflects the effects of this manipulation. For instance, in a randomized controlled trial, assigning different treatments to different groups of subjects induces different interventional distributions. More generally, data collected across multiple environments can be viewed as arising from different interventional settings in a shared causal system. Access to interventional data, or data from different environments, makes it possible in particular to infer causal relationships among variables. In predictive settings, the main interest is in recovering the local causal structure around the target variable Y, as recovering the full causal graph is often too difficult [Peters et al., 2016]. To do so, it is commonly assumed that Y itself is not subject to interventions and therefore that the causal mechanism relating Y to X remains stable across environments. Under these assumptions, one can characterize

$\mathcal{P}$ as the set of all distributions generated by interventions on X. Then, the invariant causal function relating Y to its causal parents is robust to distribution shifts induced by interventions on the covariates, and is in particular minimax optimal, i.e., it solves (3), if such shifts can be of arbitrary magnitude [see, for example, Peters et al., 2016, Meinshausen, 2018, Bühlmann, 2020, Christiansen et al., 2022].

1.1.2 Partial target information and domain adaptation

Domain generalization aims to learn a predictive function that works well in unseen test domains. In some settings, however, partial information about the target distribution may be available. Domain adaptation methods aim to leverage such information in order to improve predictive accuracy at test time.

The nature of the adaptation depends on which aspects of the distribution shift. For example, the *covariate shift* problem assumes $P_X^{\text{test}} \neq P_X^{\text{train}}$, while the conditional distribution $P_{Y|X}^{\text{test}} = P_{Y|X}^{\text{train}}$ remains the same. Approaches to the covariate shift problem are based on weighting strategies inspired by importance sampling [Shimodaira, 2000, Sugiyama et al., 2007], assuming the predictors X are observed both for the training and the test distributions, while responses are available only for the training data. The same setup is studied in more recent works on unsupervised domain adaptation, where a common approach is to match training and test distribution by including in the loss function a term that minimizes distributional distance [Ben-David et al., 2010]. Baktashmotlagh et al. [2013], Courty et al. [2017] build on this by learning a shared representation of the covariates that minimizes a similar divergence, and use the available responses from P^{train} to learn a predictive function on the aligned representation space. Magliacane et al. [2018] use instead causal reasoning to select a subset of predictors that make the conditional distribution of Y invariant across source and target domains.

A different scenario arises when test data are observed after training, or data with drifting distribution is observed sequentially. In the problem of *concept shift*, the conditional distribution may change, i.e., $P_{Y|X}^{\text{test}} \neq P_{Y|X}^{\text{train}}$ but the marginal covariate distribution remains the same, $P_X^{\text{test}} = P_X^{\text{train}}$. In this case, matching distributions is not sufficient as the optimal predictive function changes, and adaptation necessarily requires observing some responses in the target domain. Research on concept shifts has focused on drift-detection and adaptive retraining methods based on gradual forgetting mechanisms or sliding windows [Gama et al., 2014], which downweight older observations as the distribution evolves. We further elaborate on the setting of temporal shifts in the box *Time as domain* below.

Finally, settings in which both covariate and target shifts may occur have been analyzed [e.g., by Zhang et al., 2013]. This is the case for example in anti-causal prediction problems, where the predictors X are assumed to be causal effects of Y [Schölkopf et al., 2012]. Under suitable assumptions, domain adaptation can then exploit the stability of $P_{X|Y}$ across domains, even though both P_X and $P_{Y|X}$ may change [Chen and Bühlmann, 2021]. In general, more unstructured changes between training and test distribution fall under the broader transfer learning category, where the overall goal is transferring knowledge acquired from a larger training dataset to a target domain where fewer observations

are available [Weiss et al., 2016, Rojas-Carulla et al., 2018]. Also in the case of domain adaptation, having access to observations from multiple sources allows to make less restrictive assumptions on the unknown aspects of the target distribution. Again, an example is the possibility of inferring the underlying causal structure, as done by Zhang et al. [2015].

Time as domain *When considering heterogeneous data coming from different environments it is usually assumed that such environments are known and in a finite number. This could be the case for example if the same variables are observed in different geographical locations. Without knowing which data points come from which environment, characterizing the structure of distribution shifts across them becomes substantially harder. A special case comes from considering different time periods as different environments, that is, shifts in distribution that occur through time. Indeed, the sequential ordering of time points implies an additional structure in the data, as nearby time points are likely to share similar distributions.*
Despite the additional structure, considering time shifts can come with additional challenges. The timing and abruptness of changes may be unknown, and data within a single (approximately) stable regime may be scarce, precluding the collection of additional observations from past time points. Sliding window and forgetting strategies [Gama et al., 2014] are effective when shifts are slow and stable time periods are long, but struggle when change is rapid or regime-specific data is limited.
At the same time, observing distributional variations through a history of observations may be informative of additional structure in the data, which can be exploited for more robust predictions at future time points. This is the case for example in the work by Pfister et al. [2019a], where sequential observations generated by the same underlying structural causal model are available, and the covariate distribution may shift through time. Heterogeneity, together with the sequential structure, allows to identify the causal parents of the response, which can then be used safely for prediction under unobserved distribution shifts.

1.2 The role of invariance

A unifying idea behind many approaches for robust prediction in the presence of large distribution shifts is the concept of invariance. To guard against changes in distribution it is fundamental to both detect the source of such changes and, concurrently, what remains invariant: it is precisely the invariant aspects across distributions that make generalization possible. We discuss below different notions of invariance proposed in the literature for distribution generalization.

Conditional invariance In the early covariate shift problem analyzed by Shimodaira [2000], Sugiyama et al. [2007], the key assumption is

$$\textit{the distribution of} \quad Y \mid X = x \quad \textit{is invariant across domains.}$$

Under this assumption, a conditional model learned in any single environment generalizes to any domain where only P_X changes.

Causal invariance Assuming that the whole conditional distribution remains invariant may be too restrictive. In particular, changes in the marginal distribution of some of the predictors in $X = (X^1, \ldots, X^p)$ may naturally influence changes in the conditional distribution of Y given such predictors, due to structural properties among variables. For example, if Y *causes* X^i, then changes in the distribution of X^i are likely to be observed simultaneously to changes in the distribution of $Y \mid X^i$. If instead X^i *causes* Y, then changes in the marginal distribution of X^i do not affect the distribution of $Y \mid X^i$, as long as the structural mechanism relating Y to X^i does not change. This is sometimes referred to as autonomy or modularity of the causal mechanisms, and motivates causal approaches to look for a subset of predictors indexed by $S \subseteq \{1, \ldots, p\}$, X^S, such that

$$\textit{the distribution of} \quad Y \mid X^S = x^S \quad \textit{is invariant across domains.} \tag{5}$$

By the autonomy of causal mechanisms, the set of direct causes $X^{\mathrm{pa}(Y)}$ (causal parents) of Y satisfies this invariance under arbitrary shifts in the marginal covariate distribution. This is the target of invariant causal prediction (ICP) by Peters et al. [2016], whose method guarantees to recover a set that is, with high probability, a subset of the causal parents $\mathrm{pa}(Y)$. The goal of ICP is oriented more towards causal discovery, so the method is by construction conservative in the predictors selection and often results in a set of predictors that is indeed robust under interventions but with limited predictive power.

Rojas-Carulla et al. [2018] propose to exploit the same idea of invariant conditionals but with the goal of maximizing predictive performance. This results in selecting the largest subset $X^{S_{\max}}$ of predictors satisfying (5), under the assumption that the invariance observed on training environments extends to the test domain. Thus, $X^{S_{\max}}$ represents the largest set of predictors that are both stable and informative given the observed interventional settings.

Residual invariance Other more prediction-oriented works take inspiration from the instrumental variable setting and assume that distributional shifts across environments are induced by (interventions on) exogenous variables E. Then, an invariant predictive function should remove all influence of E on Y, or equivalently make the residuals insensitive to E. For example, anchor regression [Rothenhäusler et al., 2021] considers a combined loss function that trades off between predictive accuracy and stability of the residuals with respect to E (called anchors), encouraging the minimization of $\mathbb{E}[(Y - f(X))E]$. The resulting predictor is then guaranteed to be robust to shift interventions aligned with directions induced by the anchors, in an amount proportional to the weight assigned to the stability loss. An analogous moment condition is considered also by Christiansen et al. [2022]. Another example is imposing independence between $Y - f(X)$ and E [Saengkyongam et al., 2022]. A more general notion of residual stability is proposed with the boosted control function method by Gnecco et al. [2026], which seeks a function f

such that

$$\textit{the distribution of} \quad Y - f(X) \quad \textit{is invariant across domains.}$$

A similar line of works does not consider environments as exogenous variables but also, in some sense, encourages loss stability. For instance, the target of invariant risk minimization (IRM) by Arjovsky et al. [2019] is a function $\phi : \mathbb{R}^p \to \mathcal{H}$ of the predictors, for some $\mathcal{H} \subseteq \mathbb{R}^p$, that allows to find a shared risk minimizer across all domains $e \in \mathcal{E}$, when regressing the response on $\phi(X)$, i.e., such that

$$\textit{there exists an invariant } \beta : \mathcal{H} \to \mathbb{R} \textit{ such that, } \forall e \in \mathcal{E}, \beta \in \underset{b:\mathcal{H}\to\mathbb{R}}{\arg\min}\, R(b \circ \phi(X_e)), \quad (6)$$

for some loss function R, where X_e denotes the predictors in the environment e and $\circ$ denotes function composition. In practice, the original formulation and implementation of IRM has limitations in terms of generalization capability [Rosenfeld et al., 2021], and later methods have proposed adjustments to the training objective, for example to minimize the complexity of the learned representation ϕ [Ahuja et al., 2021], to minimize the variance of the risks across environments, encouraging constant loss [Krueger et al., 2021] or imposing that, given $\phi(X)$, Y is independent of the environment [Li et al., 2022].

1.2.1 Worst-case optimality and the limits of invariance

All methods discussed above aim to learn a single predictor that remains stable across environments under some notion of invariance. In several cases, this predictor can also be interpreted as worst-case optimal over a suitable class of admissible distribution shifts. This holds, for example, for causality-based approaches such as ICP [Peters et al., 2016], the causal transfer-learning method by Rojas-Carulla et al. [2018], anchor regression [Rothenhäusler et al., 2021] and the boosted control functions [Gnecco et al., 2026]. A worst-case optimal predictor is however, in many cases, suboptimal. Although it guarantees stable performance across environments, in each domain there may exist a predictor with lower prediction error. Moreover, a pooled predictor often achieves stronger average performance [Nastl and Hardt, 2024], and may be preferable when shifts in distribution are mild and robustness is not the primary objective. This is not a contradiction, as the purpose of invariant prediction is not to maximize performance in every domain, but to provide safe predictions under uncertainty, especially when distribution shifts may be severe. In the absence of observations from the target domain, a worst-case optimal predictor may indeed be the safest choice. This is the setting often referred to as *zero-shot generalization.*

A natural limitation of this viewpoint is that, by design, invariance-based methods ignore all information that is not stable across environments. As a consequence, they generally do not offer a straightforward way to *adapt* the learned invariant predictor to a specific—possibly unseen—target environment, without retraining or re-estimating the model. This is the case for domain adaptation methods discussed in Section 1.1.2 as well, some of which also rely on invariance properties [for example, the methods by

Shimodaira, 2000, Sugiyama et al., 2007, Zhang et al., 2015, Magliacane et al., 2018]. In these works, the adaptation is usually done with respect to a specific target domain, and the resulting solution is not generally reusable for a different target domain.

One final observation is that causality-inspired methods, that provide minimax guarantees for their invariant predictor, in general assume stability of the structural mechanism for the response across domains. This requires in particular the existence of a subset of predictors that satisfy such stability. In contrast, methods such as IRM and its extensions impose invariance indirectly through a training objective, which makes them less tied to a specific shift model, possibly allowing for more general forms of distribution shifts. However, these methods also come with weaker theoretical guarantees and a stronger reliance on the success of the training heuristic. In particular, they do not explicitly look for a worst-case optimal function, i.e., the one that is most predictive among all invariant predictors.

The above observations raise the following question:

When and how can we define a predictive function that

(1) *captures invariance under shifts both in P_X and $P_{Y|X}$, without necessarily requiring stable structural mechanisms or invariant predictors sets,*

(2) *allows for a minimax characterization and*

(3) *can be efficiently adapted using environment-specific information, which may become available after training?*

These points are only partially or not simultaneously addressed by the methods discussed so far, and motivate the work in [**ISD**], as we describe in the following section.

1.2.2 Partial and adaptable invariance

A natural setting in which the distribution we want to adapt to is still unseen during training is when shifts in distribution happen in time (see the window *Time as domain* in Section 1.1.2). The motivation behind [**ISD**] lies in unifying ideas from invariance and adaptation when the latter happens sequentially, and only a small amount of observations is available for such adaptation. This problem is sometimes referred to as *few-shot generalization*. The goal in [**ISD**] is to exploit information that is normally discarded by invariance-based methods because it cannot be transferred across *all* domains—or, in this case, time points—as it may still be relevant for prediction. In particular, [**ISD**] proposes a definition of invariance that allows for sample-efficient adaptation once new data become available after training time: a linear function $\beta^\top(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is called invariant if for all $t \in \mathbb{N}$ it satisfies that,

$$\mathrm{Cov}(Y_t - X_t^\top \beta, X_t^\top \beta) = 0. \tag{7}$$

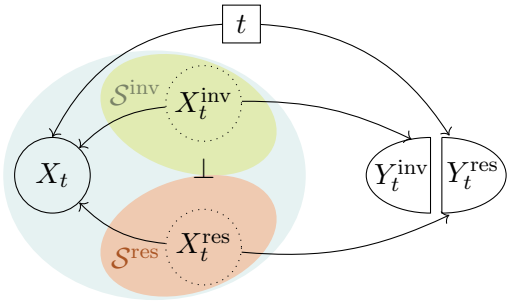


Figure 1: Graphical representation of the separation considered in **[ISD]**. X_t^{inv} and X_t^{res} correspond to the projections of X_t onto the invariant subspace $\mathcal{S}^{\text{inv}}$ and its orthogonal complement $\mathcal{S}^{\text{res}}$, respectively. $Y_t^{\text{inv}} = X_t^\top \beta$ and $Y_t^{\text{res}} = X_t^\top \delta_t$ are the two components in the conditional mean decomposition in (8), and time only enters directly through the latter.

The property in (7) guarantees that (i) $X_t^\top \beta$ has always positive (and stable, given X_t) explained variance equal to $\text{Var}(X_t^\top \beta)$ and that (ii) the residuals can be used to adapt the invariant prediction to shifts in the distribution. This addresses in particular questions (1) and (3) above. Under linearity of $\mathbb{E}[Y_t \mid X_t]$, with time-varying linear relation and covariate distribution, (7) can be interpreted as a partial invariance statement on the conditional mean of Y, which for all $t \in \mathbb{N}$ is split into

$$\mathbb{E}[Y_t \mid X_t] = X_t^\top \beta + X_t^\top \delta_t, \tag{8}$$

with $\text{Cov}(X_t^\top \beta, X_t^\top \delta_t) = 0$ at all times. As our interest is in prediction, the proposed method looks for the most predictive linear function β^{inv} among all linear functions satisfying (7). This component β^{inv} is shown to lie in a subspace $\mathcal{S}^{\text{inv}}$ (the *invariant subspace*) of the originally observed space, $\mathbb{R}^p$, but it does not necessarily correspond to a subset of the predictors (cf. *question (1)*). The invariance property in (7) requires that the projection of the covariates onto $\mathcal{S}^{\text{inv}}$ and onto its orthogonal complement (the *residual subspace*, $\mathcal{S}^{\text{res}}$) remain uncorrelated at all time points. This allows to separately use the orthogonal information to update the invariant predictor using varying, complementary information (cf. *question (3)*). In particular, we do not explicitly have access to known exogenous variables responsible for the shifts in distribution, but separate the sources of variation in $\mathbb{E}[Y_t \mid X_t]$ exploiting observed hererogeneity in a historical sequence. Finally, we mentioned that β^{inv} is the most predictive invariant parameter. This means that, at all time points, β^{inv} is the best linear predictor of Y given the projected covariates onto $\mathcal{S}^{\text{inv}}$, and that it is minimax optimal with respect to arbitrary changes of the linear parameter in $(\mathcal{S}^{\text{inv}})^\perp$ (cf. *question (2)*). A graphical representation of the setting considered in **[ISD]** is shown in Figure 1.

Connections to invariance-based approaches The invariant part of the framework presented in **[ISD]** can be seen as a middle ground between causal and IRM-based ap-

proaches. Compared to causal approaches, it does not require invariance of the causal mechanism or the existence of a subset of predictors with a stable conditional relationship to Y. This flexibility comes at the cost that β^{inv} is a stable predictive parameter rather than a structural quantity, and it does not admit a straightforward causal interpretation. At the same time, compared to IRM-based methods, the framework in [**ISD**] is more structured in the linear setting. In particular, the invariant component β^{inv} can be interpreted as the *most predictive* invariant minimizer of the IRM objective in (6) under squared loss, when the representation ϕ is taken to be the orthogonal projection onto $\mathcal{S}^{\text{inv}}$. In addition, the framework in [**ISD**] comes with an explicit worst-case robustness guarantee and allows for a principled adaptation step.

1.3 Online learning problems and invariant policies

In [**ISD**] we are interested in adapting to a changing mechanism of the response given the covariates when the observed data points are independent sequential observations. Especially for time-evolving systems, this assumption can be too strong since dependence through time often enters the equation. A naive extension, similarly to what proposed by Pfister et al. [2019a], would be to include lagged observations into the predictors vector X to take into account time dependencies up until some past time step. However, more complex dependencies may arise, for example if data are collected as the result of a decision process, and the response is a feedback signal for the quality of the decisions. In such settings, past responses are used to gradually improve the decision policy over time, breaking the independence assumption. In [**IBCB**], we extend the framework proposed in [**ISD**] to an online learning setting.

One of the simplest models for an online decision making problem is given by contextual bandits. In a bandit setting, at each time point a learner is given a decision set. From this set, it selects an action that yields a response quantifying the goodness of such action, namely a reward. The choice of the action may be facilitated by contextual information, for example a vector of predictors observed before acting. Contextual bandits differ from more complex online learning frameworks like reinforcement learning in that the effect of the action is only realized through the response but does not affect the state (context) of the environment.

In a standard bandit setting, learning a good policy requires learning something about the mechanism that assigns a reward to each action. The optimal policy, that is, the one that yields an optimal sequence of actions maximizing the cumulative reward through time, is unknown but fixed, and the learner incrementally learns it by acting throughout some time horizon.

Settings in which the optimal policy may differ across environments, or gradually shifts through time, have been studied in the literature. An example for the former case can be found in the work by Saengkyongam et al. [2023], who consider the problem of offline policy learning using data from different environments. The proposed solution is to learn an invariant policy that is optimal in the worst-case for the observed environments. This corresponds to a policy that only employs a subset of the contextual predictors, whose

influence on the reward through the action is independent of the environment. As in the case of ICP [Peters et al., 2016], this selection of invariant predictors is in practice often too conservative. Saengkyongam et al. [2024] relax this idea to allow for partial invariance of the reward mechanism given a set of predictors, as long as the action only affects the invariant part of the reward mechanism. Both works assume an offline setting, where a treatment policy is learned in retrospect after having observed bandit data from heterogeneous environments.

On the other hand, online settings in which the reward function gradually shifts through time (*non-stationary bandits*) are approached in the literature with adaptive re-training methods based on rolling windows or forgetting strategies [Russac et al., 2019, Cheung et al., 2019], but do not explicitly exploit possible invariances in the data.

In [**IBCB**] we attempt again to unify ideas from invariance and adaptation, this time in a sequential online learning setting, with the goal of gradually adapting online to an invariant policy learned offline. In particular, we consider the online non-stationary contextual bandit setting in the case of linear reward functions, and exploit the framework developed in [**ISD**] to detect partial invariance in the time-varying reward function. In this way, the policy we learn is split into (i) an invariant component which, in lack of good adaptation data (e.g., if the reward function changes too fast), can be used to take conservative—but safe—decisions and (ii) an adaptation component that adjusts the decisions of (i) to time variations in the reward function. We show in particular that finite sample guarantees introduced in [**ISD**] can be extended to the online setting to improve on the regret of standard non-stationary methods.

1.4 Outlooks and limitations

Most of the work discussed in this thesis revolves around the invariant subspace decomposition framework proposed in [**ISD**]. The starting motivation for such framework was an interest in separating stable and changing mechanisms relating a response to some covariates, e.g., to learn two functions $f, g_t : \mathbb{R}^p \to \mathbb{R}$ such that for all t

$$Y_t = f(X_t) + g_t(X_t) + \epsilon_t,$$

without imposing strict invariance of marginal or conditional distributions. A core requirement for these functions is that they should be learnable independently.

In [**ISD**], we restrict our attention to the simpler case of linear functions. The framework we propose first separates the covariate space into orthogonal subspaces, one related to the invariant mechanism and one corresponding to the directions in which the observed mechanism varies. It then relies on OLS solutions on the corresponding subspaces to separately learn the two linear functions. The choice of the linear model gives a tractable framework with well defined theoretical guarantees, but may limit practical applicability. When dealing with complex real world data, one could imagine that the stable and changing parts of the relation between X and Y need not to be well described by linear functions, nor by a decomposition into linear subspaces. Likewise, the orthogonal structure adopted in the subspace decomposition provides a clean interpretation and a

convenient estimation strategy, but it is likely not the only meaningful way to separate invariant from time-varying information. Studying a method, under looser assumptions than linearity, to detach core stable information from varying mechanisms which are still informative but time/environment specific (e.g., answering *questions (1)* to *(3)* in Section 1.2.1) could be an interesting future research direction. In particular, incorporating ideas from causality and representation learning could provide both structural guidance and identifiability tools for such extension.

On a more practical level, the estimation of the subspaces is a central part of the [**ISD**] framework. Although the theoretical framework is not intrinsically tied to a specific estimation procedure, the proposed subspace estimation relies on methods for approximate joint block diagonalization, applied to the estimated covariance matrices of the predictors under the assumption that these matrices are approximately constant within local windows. Thus, the quality of the estimated decomposition also depends on the quality of these estimates. It would be interesting to investigate alternative, more robust procedures for recovering the relevant subspaces, or move toward joint objectives that learn the decomposition and the predictive components simultaneously. Such directions seem especially relevant for extending the framework to settings in which the underlying structure is nonlinear.

These considerations naturally carry over to the contextual bandit setting studied in [**IBCB**]. There, we build on the same structural idea and assume that the context-action feature space admits an invariant subspace decomposition, so that offline historical data can be used to estimate a stable component of the reward model while online learning focuses on the lower-dimensional non-stationary part. This shows how invariance can be exploited not only for prediction, but also for adaptive decision-making. At the same time, the benefits of the approach still depend on the existence of a meaningful linear subspace decomposition as well as on the quality of the learned decomposition and on the amount of available offline data.

In this sense, the works in this thesis highlight the possible advantages of learning invariances that support adaptation. At the same time, some of the modelling assumptions that facilitate the theoretical analysis may limit the practical applicability of the presented methods, suggesting that the proposed framework could indeed benefit from weaker modelling restrictions.

2 Invariant Subspace Decomposition

Margherita Lazzaretto, Jonas Peters and Niklas Pfister

Abstract

We consider the task of predicting a response Y from a set of covariates X in settings where the conditional distribution of Y given X changes over time. For this to be feasible, assumptions on how the conditional distribution changes over time are required. Existing approaches assume, for example, that changes occur smoothly over time so that short-term prediction using only the recent past becomes feasible. To additionally exploit observations further in the past, we propose a novel invariance-based framework for linear conditionals, called Invariant Subspace Decomposition (ISD), that splits the conditional distribution into a time-invariant and a residual time-dependent component. As we show, this decomposition can be utilized both for zero-shot and time-adaptation prediction tasks, that is, settings where either no or a small amount of training data is available at the time points we want to predict Y at, respectively. We propose a practical estimation procedure, which automatically infers the decomposition using tools from approximate joint matrix diagonalization. Furthermore, we provide finite sample guarantees for the proposed estimator and demonstrate empirically that it indeed improves on approaches that do not use the additional invariant structure.

2.1 Introduction

Many commonly studied systems evolve with time, giving rise to heterogeneous data. Indeed, changes over time in the data generating process lead to shifts in the observed data distribution, and make it in general impossible to find a fixed model that consistently describes the system over time.

In this work, we analyze the problem of estimating the relation between a response Y_t and a set of covariates (or predictors) X_t, $t \in \mathbb{N}$, both of which are observed over some period of time, with the goal of predicting an unobserved response, when new observations of the covariates become available. Two types of distribution shifts across time that can arise are (i) variations in the mechanism relating the response to the covariates and (ii) variations in the covariate distribution. Under distribution shifts, successfully predicting an unobserved response from covariates requires assumptions on how the underlying data generating model changes over time.

In regression settings, varying-coefficients models [e.g., Hastie and Tibshirani, 1993, Fan and Zhang, 2008] are a common approach to deal with dynamic system behaviours related to changes in the functional relationship between a response and some covariates. In these works, the model coefficients are usually assumed to change smoothly, and smoothing methods are used in their estimation. Another approach for online estimation of smoothly changing parameters uses state-space models and filtering solutions [see, for example, Durbin and Koopman, 2012]. In general, smooth changes or other kinds of structured variations such as, for example, step-wise changes, allow us to learn a regression function using previous time points from the recent past. To exploit information further in the past, one can look for invariant patterns that persist through time. In this spirit, Meinshausen and Bühlmann [2015] define the maximin effect as a worst case optimal model that maintains good predictive properties at all observed times.

Distribution shifts in time-series models are conceptually similar to distribution shifts across domains. In both cases a key problem is to find invariant parts of the observed data distribution and transfer it to the new domain or time-point. Unsupervised domain adaptation methods aim to predict an unobserved response in the target domain: for the problem to be tractable, they rely, for example, on the invariance of the conditional distribution of the response given the covariates [Sun et al., 2017, Zhao et al., 2019] or on independently changing factors of the joint distribution of the covariates and the response [Zhang et al., 2015, Stojanov et al., 2021]. Solutions to the unsupervised domain adaptation problem are often based on aligning the source and target domains by minimizing the discrepancy either between the covariates' distributions [Zhang et al., 2015] or between the covariates' second order moments [Sun et al., 2017], or by finding low-dimensional invariant transformations of the variables whose distributions match in the domains of interest [Zhao et al., 2019, Stojanov et al., 2021]. Other unsupervised domain adaptation approaches such as the one proposed by Bousmalis et al. [2016] rely on learning disentangled representations where disentangled mean that changes in one representation do not affect the remaining ones. This approach is also explored in non-stationary settings, when changes in the covariate distribution happen over time instead of across different domains, for example in the works by Yao et al. [2022], Li et al. [2024]. The goal of these works is not prediction, but rather the identification of generalizable latent states that could be used for arbitrary downstream tasks.

The field of causality widely explores the concept of invariant and independently changing mechanisms, too. In causal models, changes in the covariate distribution are modeled as interventions, and different interventional settings represent different environments. In this context, Peters et al. [2016] propose to look for a set of stable causal predictors, that is, a set of covariates for which the conditional distribution of the response given such covariates remains invariant across different environments. The same idea is extended by Pfister et al. [2019a] to a sequential setting in which the different environments are implicitly generated by changes of the covariate distribution over time. Such invariances can then be used for prediction in previously unseen environments, too [e.g., Rojas-Carulla et al., 2018, Magliacane et al., 2018, Pfister et al., 2021]. Other works on causal discovery from heterogeneous or nonstationary data, such as the ones by Huang et al. [2020], Günther et al. [2023], study settings in which the skeleton of the causal graph remains

invariant through different contexts or time points, but the causal mechanisms are also allowed to change. Their goal is then to identify the causal graph on observed covariates. To detect whether a mechanism changes, the context or time is included as an additional variable in the graph. As the context or time can be assumed to be exogenous, this often leads to additional identification of causal edges in the graphs. Feng et al. [2022] develop and apply similar ideas in a reinforcement learning setting. Their interest is in learning the causal graph and improving the efficiency of non-stationarity adaptation by disentangling the system variations as latent parameters.

In this work, we build on invariance ideas both from the maximin framework proposed by Meinshausen and Bühlmann [2015] and from the covariates shifts literature, and exploit them for time adaptation. We propose an invariance-based framework, which we call invariant subspace decomposition (ISD), to estimate time-varying regression models in which both the covariate distribution and their relationship with a response can change. In particular, we consider a sequence of independent random vectors $(X_t, Y_t)_{t\in\mathbb{N}} \subseteq \mathbb{R}^p \times \mathbb{R}$ satisfying for all time points $t \in \mathbb{N}$ a linear model of the form

$$Y_t = X_t^\top \gamma_{0,t} + \epsilon_t \tag{2.1.1}$$

with $\mathbb{E}[\epsilon_t \mid X_t] = 0$, where $\gamma_{0,t}$ is the (unknown) *true time-varying parameter*. We allow the covariance matrices $\mathrm{Var}(X_t)$ to change over time, assuming that they are approximately constant in small time windows. Regarding changes in $\gamma_{0,t}$, we only need to assume smoothness during test time (if there is more than one test point)—we provide more details later in Remark 2.2.11. Having observed the X_t and Y_t up until some time point n, our goal is to learn γ_{0,t^*} for some $t^* > n$, which we then use to predict Y_{t^*} from X_{t^*}. We propose to do so by considering the explained variance, which, for all t, is given for some function f by $\Delta\mathrm{Var}_t(f) \coloneqq \mathrm{Var}(Y_t) - \mathrm{Var}(Y_t - f(X_t))$. Using the explained variance as the objective function provides an intuitive evaluation of the predictive quality of a function f: negative explained variance indicates in particular that using f for prediction is harmful, in that it is worse than the best constant function. Under model (2.1.1), f is a linear function fully characterized by a parameter $\beta \in \mathbb{R}^p$, i.e., $f(X_t) = X_t^\top \beta$, and the true time-varying parameter can always be expressed as

$$\gamma_{0,t} = \operatorname*{arg\,max}_{\beta\in\mathbb{R}^p} \Delta\mathrm{Var}_t(\beta). \tag{2.1.2}$$

Key to ISD is the decomposition of the explained variance maximization into a time-invariant and a time-dependent part: we show in Theorem 2.2.7 that, under some assumptions, for all $t \in \mathbb{N}$, $\gamma_{0,t}$ can be expressed as

$$\gamma_{0,t} = \underbrace{\operatorname*{arg\,max}_{\beta\in\mathcal{S}^{\mathrm{inv}}} \overline{\Delta\mathrm{Var}}(\beta)}_{=:\beta^{\mathrm{inv}}} + \underbrace{\operatorname*{arg\,max}_{\beta\in\mathcal{S}^{\mathrm{res}}} \Delta\mathrm{Var}_t(\beta)}_{=:\delta_t^{\mathrm{res}}}, \tag{2.1.3}$$

where $\overline{\Delta\mathrm{Var}}(\beta) \coloneqq \frac{1}{n}\sum_{t=1}^n \Delta\mathrm{Var}_t(\beta)$, and $\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}} \subseteq \mathbb{R}^p$ are two orthogonal linear subspaces, with $\mathcal{S}^{\mathrm{inv}} \oplus \mathcal{S}^{\mathrm{res}} = \mathbb{R}^p$, assumed to be uniquely determined by the distribution

of $(X_1, Y_1), \ldots, (X_n, Y_n)$. The decomposition of $\mathbb{R}^p$ into $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ from observed data is achieved by exploiting ideas from independent subspace analysis [e.g., Gutch and Theis, 2012]. Unlike works on latent representations, we do not reduce the overall dimensionality of the problem but partition the observed space into two lower dimensional subspaces, $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$. This in particular implies that the two sub-problems in (2.1.3) each have less degrees of freedom than the original one. The first sub-problem can be solved using all available heterogeneous observations, which we call *historical data*: its solution β^{inv} is a time-invariant linear parameter that partially describes the dependence of Y on X at all times. The second sub-problem is a time adaptation problem that tunes the invariant component to a time point of interest: its solution δ_t^{res} explains the residuals $Y_t - X_t^\top \beta^{\text{inv}}$; the sum $\beta^{\text{inv}} + \delta_t^{\text{res}}$ gives an estimator for the time-varying linear parameter of interest $\gamma_{0,t}$. In order to estimate the residual component δ_t^{res}, we assume the model is approximately stationary in a small time window preceding t and use this local subset of the data, which we call *adaption data*, for estimation. We distinguish two tasks: (i) the *zero-shot task*, where adaption data is not available and we approximate $\gamma_{0,t}$ by β^{inv} and (ii) the *time-adaption task*, where adaption data is available and we solve both sub-problems to approximate $\gamma_{0,t}$ by $\beta^{\text{inv}} + \delta_t^{\text{res}}$. A two-dimensional example of $\gamma_{0,t}$ and its ISD estimates is shown in Figure 2.1.1. The same example is presented again in more detail in Example 2.2.6 below and used as a running example throughout the paper.

The fundamental assumption we make to apply the ISD framework to new data (Assumption 2) is that the decomposition inferred from available observations generalizes to unseen time points. This guarantees that the estimated invariant component can be meaningfully used for prediction at all new time points, either directly or as part of a two-step estimation that is fine-tuned by solving the time adaptation sub-problem.

The ISD framework allows us to improve on naive methods that directly maximize the explained variance on $\mathbb{R}^p$ using only the most recent available observations and on methods focusing on invariance across environments or time points such as the maximin by Meinshausen and Bühlmann [2015], which do not account for time-adaptation. We show in particular in Theorem 2.4.1 that isolating an invariant component and reducing the dimensionality of the adaptation problem guarantees lower prediction error or, equivalently, higher explained variance compared to naive methods. An example is provided for a simulated setting in Figure 2.1.2, which shows on the left the average explained variance by the invariant component computed at training time on historical data, and on the right the cumulative explained variance by the invariant (zero-shot task) and the adapted invariant component (time-adaptation task) at testing time when new observations become available. For the zero-shot task, we compare the ISD invariant component with the standard OLS solution, computed on all training observations, and with the maximin effect ($\hat{\beta}^{\text{mm}}$) by Meinshausen and Bühlmann [2015], which maximizes the worst case explained variance over the available observations. For the time-adaptation task, we compare the ISD with the standard OLS solution computed on a rolling window over the latest new observations. For both tasks, the ISD estimates achieve higher cumulative explained variance at new time-points.

The remainder of the paper is structured as follows. In Section 2.2 we introduce the

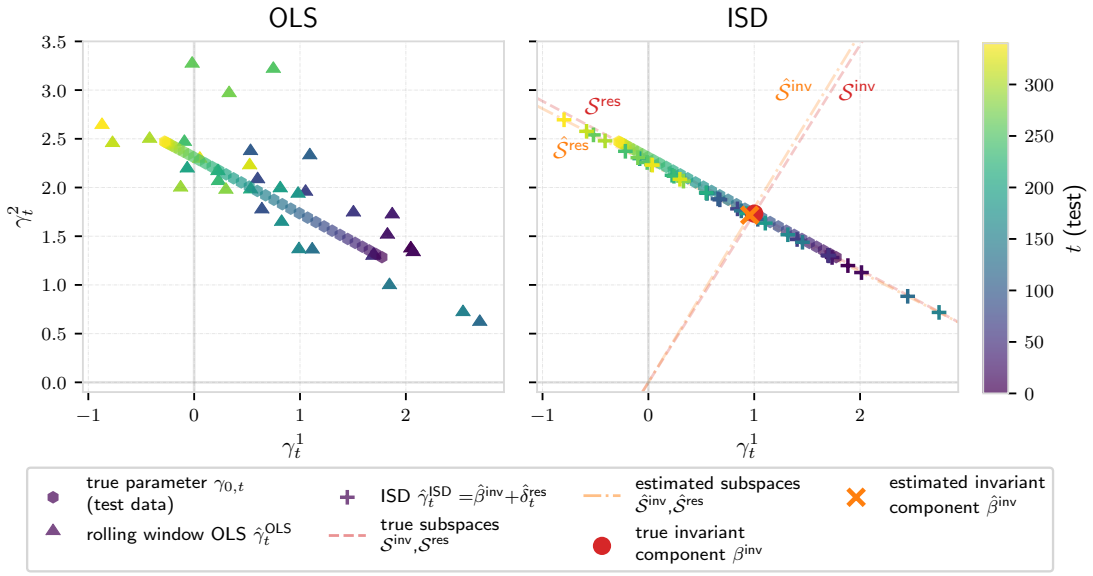


Figure 2.1.1: Example of two-dimensional true parameter $\gamma_{0,t}$ varying on a one-dimensional subspace of $\mathbb{R}^2$, and its estimates using ISD (right) compared to rolling window OLS (left). Time is visually encoded using a color map. (Left) True parameter $\gamma_{0,t}$ (hexagons) on 350 test points and OLS estimates $\hat{\gamma}_t^{\text{OLS}}$ based on rolling windows of size 16. (Right) Same test data and true parameters $\gamma_{0,t}$, but now we additionally use 1000 prior time-points as historical data (not shown) to estimate the decomposition of $\mathbb{R}^2$ into the orthogonal subspaces $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ (dashed lines). Next, we estimate β^{inv} using the historical data and $\hat{\mathcal{S}}^{\text{inv}}$. Then, using the same rolling windows as in the left plot as adaption data, we estimate δ_t^{res} using $\hat{\mathcal{S}}^{\text{res}}$. The ISD estimates are then given by $\hat{\gamma}_t^{ISD} = \hat{\beta}^{\text{inv}} + \hat{\delta}_t^{\text{res}}$. All details on the generative model are provided in Example 2.2.6. The subspaces $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ do not need to be axis aligned, so ISD is applicable even in cases where the conditional of Y_t given X_t and all conditionals of Y_t given subsets of X_t vary over time.

ISD framework, starting from the identification of the invariant and residual subspaces for orthogonal covariates transformation and showing the explained variance separation (2.1.3) in Theorem 2.2.7. In Sections 2.2.2 and 2.2.3 we define population estimators for the invariant and residual components, solutions to the two sub-problems in (2.1.3). We then describe, in Section 2.3, the two tasks that ISD solves, namely zero-shot prediction and time-adaptation, and provide a characterization of the invariant component as a worst-case optimal parameter. In Section 2.4 we propose an estimation method for ISD, and provide finite sample generalization guarantees for the proposed estimator. Finally, in Section 2.5, we illustrate ISD based on numerical experiments, both on simulated and on real world data, and validate the theoretical results presented in the paper.

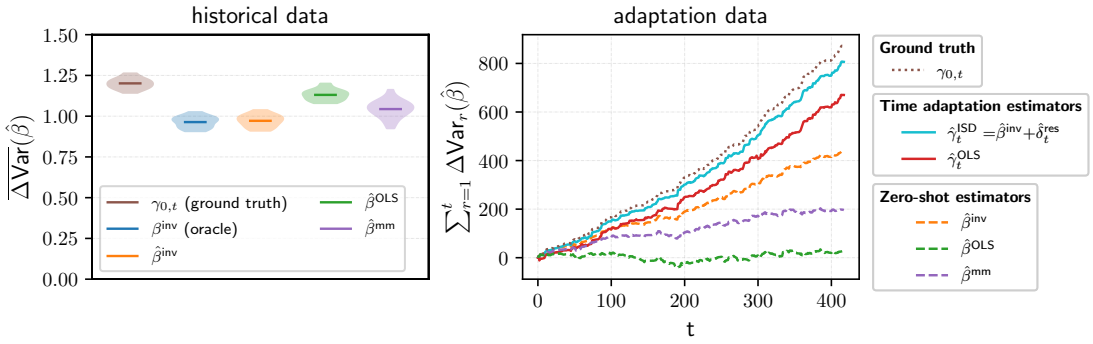


Figure 2.1.2: For the data-generating model in Section 2.5.2, we plot (left) the average explained variance (distribution over 20 runs) obtained at training time (historical data) and (right) the cumulative explained variance obtained testing time (adaptation data) (in one of the 20 runs); in this example, the time-varying components in the historical and adaptation data have disjoint support. The example considers $p = 10$-dimensional predictors and an invariant component of dimension 7. As baselines, we use (i) the true time-varying parameter $\gamma_{0,t}$, which maximizes the explained variance at all observed time points t, and (ii) the oracle invariant component β^{inv}. 6000 historical observations are used to estimate: (iii) the invariant component $\hat{\beta}^{\text{inv}}$ of the ISD framework, (iv) the OLS solution $\hat{\beta}^{\text{OLS}}$, (v) the maximin effect $\hat{\beta}^{\text{mm}}$. Starting from $t = 0$ after the observed history, windows of length $3p$ are used to estimate: (vi) the adaptation parameter $\hat{\delta}_t^{\text{res}}$ for $\hat{\beta}^{\text{inv}}$ to obtain the ISD estimate $\hat{\gamma}^{ISD}$ and (vii) the rolling window OLS solution $\hat{\gamma}_t^{\text{OLS}}$. While at training time on historical data the ISD invariant component $\hat{\beta}^{\text{inv}}$ is the most conservative, with the lowest average explained variance, after a distribution shift (adaptation data) the same component can explain higher variance than other methods based on historical data only ($\hat{\beta}^{\text{OLS}}, \hat{\beta}^{\text{mm}}$), and can be tuned to new time points to improve on estimators based on adaptation data only ($\hat{\gamma}_t^{\text{OLS}}$).

2.2 Invariant subspace decomposition

We formalize the setup described in the introduction in the following setting.

Setting 2.2.1. *Let $(X_t, Y_t)_{t\in\mathbb{N}} \subseteq \mathbb{R}^p \times \mathbb{R}^1$ be a sequence of independent random vectors satisfying for all $t \in \mathbb{N}$ a linear model as in (2.1.1), with $\gamma_{0,t} \in \mathbb{R}^p$ and $\mathbb{E}[\epsilon_t \mid X_t] = 0$. Assume that for all $t \in \mathbb{N}$ the covariance matrix of the predictors $\Sigma_t := \text{Var}(X_t)$ is strictly positive definite. Moreover, for all $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$ and assume we observe n observations $(X_t, Y_t)_{t\in[n]}$ from model (2.1.1), which we call* historical data. *Additionally, let $\mathcal{I}^{\text{ad}} \subseteq \mathbb{N}$ be an interval of consecutive time points with $m := |\mathcal{I}^{\text{ad}}| > p$ and $\min_{t\in\mathcal{I}^{\text{ad}}} t > n$. Assume we observe a second set of observations $(X_t, Y_t)_{t\in\mathcal{I}^{\text{ad}}}$ from the same model but succeeding the historical data, which we call* adaptation data. *Finally,*

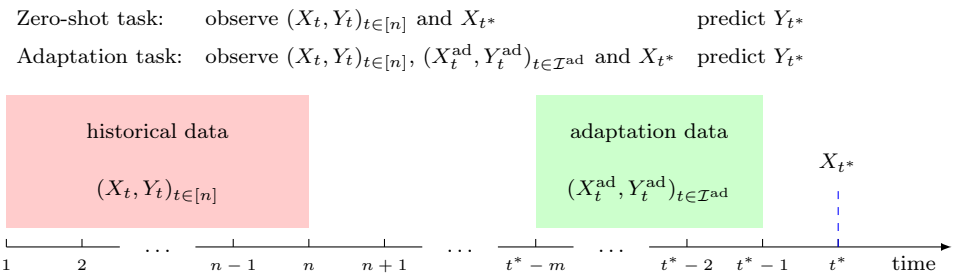


Figure 2.2.3: Illustration of the historical and adaptation data and the zero-shot and adaptation tasks.

denote by $t^* > \max_{t\in\mathcal{I}^{\text{ad}}}$ *a time point of interest that occurs after the adaptation data and assume that, for all* $t \in \mathcal{I}^{\text{ad}} \cup \{t^*\}$*, the quantities* $\gamma_{0,t}$*,* Σ_t *and* $\text{Var}(\epsilon_t)$ *in model* (2.1.1) *are constant (in practice, being approximately constant is sufficient).*

Our goal is to predict Y_{t^*} from X_{t^*}. A naive solution is to only consider the adaptation data in $\mathcal{I}^{\text{ad}}$ (see also the analysis in Section 2.3): we aim to improve on this approach by additionally exploiting historical data in $[n]$. In particular, depending on whether we only have access to the historical data or we additionally have access to the data in $\mathcal{I}^{\text{ad}}$, we can solve (i) the zero-shot task or (ii) the adaptation task, respectively. The full setup including the different tasks is visualized in Figure 2.2.3. The two tasks correspond to the two sub-problems in (2.1.3): the solution to (i) is a time-invariant parameter $\beta^{\text{inv}} \in \mathbb{R}^p$, defined formally below in Section 2.2.2; the solution to (ii) is an adaptation parameter δ_t^{res} formally defined in Section 2.2.3 that satisfies $\beta^{\text{inv}} + \delta_t^{\text{res}} = \gamma_{0,t}$.

We start by defining what a time-invariant parameter is. To do so, we first consider the explained variance for a parameter $\beta \in \mathbb{R}^p$ at time $t \in \mathbb{N}$, defined by

$$\Delta\text{Var}_t(\beta) := \text{Var}(Y_t) - \text{Var}(Y_t - X_t^\top \beta).$$

Under model (2.1.1), the explained variance can be equivalently expressed as

$$\begin{aligned}\Delta\text{Var}_t(\beta) &= 2\,\text{Cov}(Y_t, X_t^\top\beta) - \text{Var}(X_t^\top\beta) \\ &= 2\,\text{Cov}(Y_t - X_t^\top\beta, X_t^\top\beta) + \text{Var}(X_t^\top\beta). \end{aligned} \tag{2.2.4}$$

A desirable property for a non-varying parameter $\beta \in \mathbb{R}^p$ is to guarantee that its explained variance remains non-negative at all time points. We call parameters $\beta \in \mathbb{R}^p$ *never harmful* if, for all $t \in \mathbb{N}$, $\Delta\text{Var}_t(\beta) \geq 0$. Intuitively, parameters with this property at least partially explain Y_t at all time points $t \in \mathbb{N}$, and are therefore meaningful to use for prediction.[1] In this paper, we consider the subset of never harmful parameters $\beta \in \mathbb{R}^p$ that satisfy $\Delta\text{Var}_t(\beta) = \text{Var}(X_t^\top\beta)$ or equivalently, using (2.2.4), $\text{Cov}(Y_t - X_t^\top\beta, X_t^\top\beta) =$

[1] Under an assumption similar to Assumption 2 below, the maximin framework by Meinshausen and Bühlmann [2015] allows us to estimate never harmful parameters: we provide a more detailed comparison between our approach and the maximin in Remark 2.3.2.

0; thus, the explained variance of these parameters reflects changes of $\mathrm{Var}(X_t)$ (over t) but is independent of changes in $\gamma_{0,t}$.

Definition 2.2.2 (time-invariance)**.** *We call a parameter $\beta \in \mathbb{R}^p$* time-invariant (over $[n]$) *if for all $t \in [n]$,*

$$\mathrm{Cov}(Y_t - X_t^\top \beta, X_t^\top \beta) = 0. \tag{2.2.5}$$

We use this definition, in Section 2.2.1, to define the subspaces $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ that allow us to obtain the separation of $\gamma_{0,t}$ as in (2.1.3). In particular, as we show in Proposition 2.2.10, the definition of the invariant subspace $\mathcal{S}^{\mathrm{inv}}$ guarantees that the invariant component β^{inv} maximizing the pooled explained variance over $\mathcal{S}^{\mathrm{inv}}$ is always a time-invariant parameter according to Definition 2.2.2.

While we consider the explained variance here, one can equivalently consider the mean squared error (MSPE).

Remark 2.2.3 (Exchanging explained variance with MSPE)**.** *Under model (2.1.1) maximizing the explained variance at time $t \in [n]$ is equivalent to minimizing the MSPE at t, $\mathrm{MSPE}_t(\beta) \coloneqq \mathbb{E}[(Y_t - X_t^\top \beta)^2]$. More precisely, assuming that all variables have mean zero,*

$$\begin{aligned}\arg\max_{\beta \in \mathbb{R}^p} \Delta\mathrm{Var}_t(\beta) &= \arg\min_{\beta \in \mathbb{R}^p} \mathrm{Var}(Y_t - X_t^\top \beta) \\ &= \mathrm{Var}(X_t)^{-1} \mathrm{Cov}(X_t, Y_t) \\ &= \arg\min_{\beta \in \mathbb{R}^p} \mathrm{MSPE}_t(\beta).\end{aligned}$$

The remaining part of this section is organized as follows. Section 2.2.1 explicitly constructs the spaces $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ and shows that they can be characterized by joint block diagonalization. In Section 2.2.2 we then analyze the time-invariant part of (2.1.3), leading to the optimal time-invariant parameter β^{inv}, and show in Proposition 2.2.10 *(iii)* that its solution corresponds to the maximally predictive time-invariant parameter and is an interesting target of inference in its own right. In Section 2.2.3 we analyze the residual part of (2.1.3), leading to the optimal parameter δ_t^{res}.

2.2.1 Invariant and residual subspaces

We now construct the two linear subspaces $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ that allow us to express the true time-varying parameter as the solution to two separate optimizations as in (2.1.3). For a linear subspace $\mathcal{S} \subseteq \mathbb{R}^p$, denote by $\Pi_{\mathcal{S}} \in \mathbb{R}^{p \times p}$ the orthogonal projection matrix from $\mathbb{R}^p$ onto $\mathcal{S}$, that is, a symmetric matrix such that $\Pi_{\mathcal{S}}^2 = \Pi_{\mathcal{S}}$ and, for all vectors $v \in \mathbb{R}^p$, $\Pi_{\mathcal{S}} v \in \mathcal{S}$ and $(v - \Pi_{\mathcal{S}} v)^\top \Pi_{\mathcal{S}} v = 0$. We call a collection of pairwise orthogonal linear subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_q \subseteq \mathbb{R}^p$ with $\bigoplus_{j=1}^q \mathcal{S}_j = \mathbb{R}^p$ an *orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition (of cardinality q)* if for all $i, j \in \{1, \ldots, q\}$ with $i \neq j$, and for all $t \in [n]$ it holds that

$$\mathrm{Cov}(\Pi_{\mathcal{S}_i} X_t, \Pi_{\mathcal{S}_j} X_t) = 0. \tag{2.2.6}$$

Moreover, we define an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition as *irreducible* if there is no orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition with strictly larger cardinality. We will construct $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ from an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition.

Given a (not necessarily irreducible) orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition $\{\mathcal{S}_j\}_{j=1}^q$ the true time-varying parameter can be expressed as the sum of q orthogonal components each lying in one subspace of the partition. These components can be expressed in terms of the covariates and of the response at time t as shown by the following lemma.

Lemma 2.2.4. *Let $\{\mathcal{S}_j\}_{j=1}^q$ be an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Then it holds for all $t \in [n]$ that $\gamma_{0,t} = \sum_{j=1}^q \Pi_{\mathcal{S}_j}\gamma_{0,t}$ and for all $j \in \{1,\dots,q\}$ that*

$$\Pi_{\mathcal{S}_j}\gamma_{0,t} = \operatorname{Var}\left(\Pi_{\mathcal{S}_j}X_t\right)^\dagger \operatorname{Cov}\left(\Pi_{\mathcal{S}_j}X_t, Y_t\right) \tag{2.2.7}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse.

Each component is in particular the projection of $\gamma_{0,t}$ on the corresponding subspace $\mathcal{S}_j$, as well as a maximizer of the explained variance in $\mathcal{S}_j$, as shown by Lemma 2.2.5.

Lemma 2.2.5. *Let $\mathcal{N} \subseteq \mathbb{N}$ and let $\{\mathcal{S}_j\}_{j=1}^q$ be an orthogonal and $(X_t)_{t\in\mathcal{N}}$-decorrelating partition.*[2] *Then it holds for all $j \in \{1,\dots,q\}$ and for all $t \in \mathcal{N}$ that*

$$\operatorname*{arg\,max}_{\beta\in\mathcal{S}_j} \Delta\operatorname{Var}_t(\beta) = \operatorname{Var}(\Pi_{\mathcal{S}_j}X_t)^\dagger \operatorname{Cov}(\Pi_{\mathcal{S}_j}X_t, Y_t). \tag{2.2.8}$$

Combining Lemmas 2.2.4 and 2.2.5 implies that

$$\gamma_{0,t} = \sum_{j=1}^q \operatorname*{arg\,max}_{\beta\in\mathcal{S}_j} \Delta\operatorname{Var}_t(\beta). \tag{2.2.9}$$

Therefore, any orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition allows us to split the optimization (2.1.2) of the explained variance into separate optimizations over the individual subspaces. In order to leverage all available observations in at least some of the optimizations by pooling the explained variance as in (2.1.3), we need the optimizer to remain constant over time. More formally, we call a linear subspace $\mathcal{S} \subseteq \mathbb{R}^p$ *opt-invariant (optimum invariant) on* $[n]$ if for all $t, s \in [n]$ it holds that

$$\operatorname*{arg\,max}_{\beta\in\mathcal{S}} \Delta\operatorname{Var}_t(\beta) = \operatorname*{arg\,max}_{\beta\in\mathcal{S}} \Delta\operatorname{Var}_s(\beta). \tag{2.2.10}$$

For all terms in (2.2.9) corresponding to an opt-invariant subspace, we can—by definition of opt-invariance—use all time-points in the optimization. For an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition $\{\mathcal{S}_j\}_{j=1}^q$, we therefore define the *invariant subspace*

[2] This is defined analogously to an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition.

$\mathcal{S}^{\text{inv}}$ and the *residual subspace* $\mathcal{S}^{\text{res}}$ by

$$\mathcal{S}^{\text{inv}} := \bigoplus_{\substack{j\in\{1,\dots,q\}:\\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \mathcal{S}_j \quad \text{and} \quad \mathcal{S}^{\text{res}} := \bigoplus_{\substack{j\in\{1,\dots,q\}:\\ \mathcal{S}_j \text{ not opt-invariant on } [n]}} \mathcal{S}_j, \tag{2.2.11}$$

respectively. The invariant subspace $\mathcal{S}^{\text{inv}}$ is opt-invariant (see Lemma 2.6.5). Moreover, by definition, it holds that $\mathcal{S}^{\text{res}} = \left(\mathcal{S}^{\text{inv}}\right)^{\perp}$, where $(\cdot)^{\perp}$ denotes the orthogonal complement in $\mathbb{R}^p$, and that $\{\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}\}$ is an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition (see Lemma 2.6.4). To ensure that these two spaces do not depend on the chosen irreducible orthogonal partition, we introduce the following assumption.

Assumption 1 (uniqueness of the subspace decomposition)**.** *Let $\{\mathcal{S}_j\}_{j=1}^{q}$ and $\{\bar{\mathcal{S}}_j\}_{j=1}^{\bar{q}}$ be two irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partitions. Then, it holds that*

$$\bigoplus_{\substack{j\in\{1,\dots,q\}:\\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \mathcal{S}_j = \bigoplus_{\substack{j\in\{1,\dots,\bar{q}\}:\\ \bar{\mathcal{S}}_j \text{ opt-invariant on } [n]}} \bar{\mathcal{S}}_j.$$

Assumption 1 is for example satisfied if an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition is unique. As we show in Appendix 2.A.1, the uniqueness does not always hold (e.g., if for all $t \in [n]$ the multiplicity of some eigenvalues of Σ_t is larger than one and shared across all such matrices). Assumption 1 is, however, a mild assumption; for example, the uniqueness of an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition is satisfied if there exists at least one $t \in [n]$ such that all eigenvalues of Σ_t, which we have assumed to be non-zero, are distinct (see Lemma 2.6.6 and Proposition 2.2.8 below). Whenever Assumption 1 holds, the invariant and residual spaces $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ do not depend on which irreducible orthogonal partition is used in their construction.

Example 2.2.6. *This example describes the setting used to generate Figure 2.1.1. Consider a 2-dimensional covariate $X_t \in \mathbb{R}^2$, and assume that model (2.1.1) is defined as follows. We take, for all $t \in [n]$*

$$\gamma_{0,t} = \begin{bmatrix} 1.5\sqrt{3} + 1 - \sqrt{3}t/n \\ t/n - 1.5 + \sqrt{3} \end{bmatrix} \quad \text{and} \quad \Sigma_t = \frac{1}{4}\begin{bmatrix} 3\sigma_{1,t} + \sigma_{2,t} & \sqrt{3}(\sigma_{2,t} - \sigma_{1,t}) \\ \sqrt{3}(\sigma_{2,t} - \sigma_{1,t}) & \sigma_{1,t} + 3\sigma_{2,t} \end{bmatrix},$$

where $\sigma_{1,t}$ and $\sigma_{2,t}$ are two fixed sequences sampled as two independent i.i.d. samples from a uniform distribution on $[0, 1]$. In this example, we have that an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition is given by the two spaces

$$\mathcal{S}_1 = \langle \begin{bmatrix} 0.5\sqrt{3} \\ -0.5 \end{bmatrix} \rangle \quad \mathcal{S}_2 = \langle \begin{bmatrix} 0.5 \\ 0.5\sqrt{3} \end{bmatrix} \rangle.$$

Indeed, it holds for all $t \in [n]$ that $\text{Cov}(\Pi_{\mathcal{S}_1} X_t, \Pi_{\mathcal{S}_2} X_t) = \Pi_{\mathcal{S}_1} \Sigma_t \Pi_{\mathcal{S}_2} = \mathbf{0}_{4\times 4}$. Moreover, since $\sigma_{1,t}$ and $\sigma_{2,t}$ are the eigenvalues of Σ_t, the irreducible orthogonal partition defined by $\mathcal{S}_1$ and $\mathcal{S}_2$ is unique (i.e., Assumption 1 is satisfied by Lemma 2.6.6) if $\sigma_{1,t} \neq \sigma_{2,t}$ for some $t \in [n]$. It also holds that

$$\Pi_{\mathcal{S}_2} \gamma_{0,t} = \begin{bmatrix} 1 \\ \sqrt{3} \end{bmatrix},$$

which does not depend on t (it can be verified that the same does not hold for $\Pi_{\mathcal{S}_1}\gamma_{0,t}$). As shown in Lemmas 2.2.4 and 2.2.5, it holds that

$$\arg\max_{\beta\in\mathcal{S}_j} \Delta\mathrm{Var}_t(\beta) = \arg\max_{\beta\in\mathcal{S}_j} \mathrm{Var}(Y_t) - \mathrm{Var}(Y_t - X_t^\top \Pi_{\mathcal{S}_j}\beta) = \Pi_{\mathcal{S}_j}\gamma_{0,t}.$$

It follows that the subspace $\mathcal{S}_2$ is opt-invariant, whereas $\mathcal{S}_1$ is not, and therefore $\mathcal{S}^{\mathrm{inv}} = \mathcal{S}_2$ and $\mathcal{S}^{\mathrm{res}} = \mathcal{S}_1$. The two spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ also appear in Figure 2.1.1: the true time-varying parameter $\gamma_{0,t}$ does not vary with t in the direction of the vector $[0.5, 0.5\sqrt{3}]^\top$ generating $\mathcal{S}_2$; $\mathcal{S}_1$ can be visualized when connecting the circles.

In order for $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ to be useful for prediction on future observations, we assume that the subspace separations we consider remain fixed over time.

Assumption 2 (generalization)**.** *For all irreducible orthogonal and $(X)_{t\in[n]}$-decorrelating partitions, $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ defined in (2.2.11) satisfy the following two conditions: (i) for all $t \in \mathbb{N}$ it holds that $\mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t, \Pi_{\mathcal{S}^{\mathrm{res}}} X_t) = 0$ and (ii) $\mathcal{S}^{\mathrm{inv}}$ is opt-invariant on $\mathbb{N}$.*

As shown in the following theorem, Assumption 2 ensures that the two sets satisfy a separation of the form (2.1.3) for all observed and unobserved time points $t \in \mathbb{N}$, as desired. For unobserved time points $t \in \mathbb{N} \setminus [n]$, Assumption 2 does not require (2.2.6) to hold for all $i, j \in \{1, \dots, q\}$, but only for all $i, j \in \{1, \dots, q\}$ such that $\mathcal{S}_i \subseteq \mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}_j \subseteq \mathcal{S}^{\mathrm{res}}$.

Theorem 2.2.7. *Assume Assumption 2 is satisfied. Let $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ be defined as in (2.2.11) for an arbitrary irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Then, it holds for all $t \in \mathbb{N}$ that*

$$\gamma_{0,t} = \arg\max_{\beta\in\mathcal{S}^{\mathrm{inv}}} \overline{\Delta\mathrm{Var}}(\beta) + \arg\max_{\beta\in\mathcal{S}^{\mathrm{res}}} \Delta\mathrm{Var}_t(\beta), \tag{2.2.12}$$

where $\overline{\Delta\mathrm{Var}}(\beta) \coloneqq \frac{1}{n}\sum_{t=1}^{n} \Delta\mathrm{Var}_t(\beta)$. Moreover, if Assumption 1 is satisfied, the separation in (2.2.12) is independent of the considered irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition.

The proof of Theorem 2.2.7 can be found in Appendix 2.E.1 and relies on the fact that, under Assumption 2, the invariant and residual subspace form an orthogonal and $(X_t)_{t\in\mathbb{N}}$-decorrelating partition.

2.2.1.1 Identifying invariant and residual subspaces using joint block diagonalization

We can characterize an irreducible orthogonal partition using joint block diagonalization of the set of covariance matrices $(\Sigma_t)_{t\in[n]}$. Joint block diagonalization of $(\Sigma_t)_{t\in[n]}$ consists of finding an orthogonal matrix $U \in \mathbb{R}^{p\times p}$ such that the matrices $\tilde{\Sigma}_t \coloneqq U^\top \Sigma_t U$, $t \in [n]$, are block diagonal and we can choose q^U blocks such that the indices of the corresponding submatrices do not change with t (the entries of the blocks may change with t though).

Let $q^U_{\max}$ be the largest number of such blocks and let $(\tilde{\Sigma}_{t,1})_{t\in[n]}, \ldots, (\tilde{\Sigma}_{t,q^U_{\max}})_{t\in[n]}$ denote the corresponding *common blocks* with dimensions $p_1, \ldots, p_{q^U_{\max}}$ that are independent of t. We call U a *joint block diagonalizer* of $(\Sigma_t)_{t\in[n]}$. Moreover, we call U an *irreducible* joint block diagonalizer if, in addition, for all other joint block diagonalizers $U' \in \mathbb{R}^{p\times p}$ the resulting number of common blocks is at most $q^U_{\max}$. Joint block diagonalization has been considered extensively in the literature [see, for example, Murota et al., 2010, Nion, 2011, Tichavsky and Koldovsky, 2012] and various computationally feasible algorithms have been proposed (see Section 2.A.3 for further details).

In our setting, joint block diagonalization can be used to identify the invariant and residual subspaces, since an irreducible joint block diagonalizer U of the covariance matrices $(\Sigma_t)_{t\in[n]}$ corresponds to an irreducible orthogonal partition, as the following proposition shows.

Proposition 2.2.8. *(i) Let $U \in \mathbb{R}^{p\times p}$ be a joint block diagonalizer of $(\Sigma_t)_{t\in[n]}$. For all $j \in \{1, \ldots, q^U_{\max}\}$, let $S_j \subseteq \{1, \ldots, p\}$ denote the subset of indices corresponding to the j-th common block $\tilde{\Sigma}_{t,j}$. Moreover, let u^k denote the k-th column of U and, for all $j \in \{1, \ldots, q^U_{\max}\}$, define*

$$\mathcal{S}_j \coloneqq \operatorname{span}\{u^k \mid k \in S_j\}.$$

Then, $\{\mathcal{S}_j\}_{j=1}^{q^U_{\max}}$ is an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Moreover, if the joint block diagonalizer is irreducible, then the corresponding orthogonal partition is irreducible.

(ii) The converse is also true. Let $\{\mathcal{S}_j\}_{j=1}^{q}$ be an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Then, there exists a joint block diagonalizer $U \in \mathbb{R}^{p\times p}$ of $(\Sigma_t)_{t\in[n]}$ such that for all $t \in [n]$ the matrix $\tilde{\Sigma}_t \coloneqq U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\widetilde{\Sigma}_{t,j} = (U^{S_j})^\top \Sigma_t U^{S_j}$, $j \in \{1, \ldots, q\}$ of dimension $|S_j| = \dim(\mathcal{S}_j)$, where $S_j \subseteq \{1, \ldots, p\}$ indexes a subset of the columns of U. If the orthogonal partition is irreducible, then U is an irreducible joint block diagonalizer. Moreover, $\Pi_{\mathcal{S}_j} = U^{S_j}(U^{S_j})^\top$.

If Assumption 1 is satisfied, any irreducible joint block diagonalizer U, via its corresponding irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition constructed in Proposition 2.2.8(i), leads to the same (unique) invariant and residual subspaces defined in (2.2.11).

It is clear that Assumption 1 is automatically satisfied whenever the joint block diagonalization is unique up to trivial indeterminacies, that is, if for all orthogonal matrices $U, U' \in \mathbb{R}^{p\times p}$ that jointly block diagonalize the set $(\Sigma_t)_{t\in[n]}$ into $q^U_{\max}$ irreducible common blocks, it holds that U is equal to U' up to block permutations and block-wise isometric transformations. Explicit conditions under which uniqueness of joint block diagonalization is satisfied can be found, for example, in the works by De Lathauwer [2008], Murota et al. [2010]. Intuitively, these conditions are satisfied whenever there is sufficient variability across time in the covariance matrices $(\Sigma_t)_{t\in[n]}$.

Given an irreducible joint block diagonalizer U, the invariant and residual subspaces can be identified using Proposition 2.2.8 and Lemma 2.2.4, by checking for all $j \in \{1, \dots, q^U_{\max}\}$ whether $\Pi_{\mathcal{S}_j}\gamma_{0,t} = U^{S_j}(U^{S_j})^\top\gamma_{0,t}$ remains constant ($\mathcal{S}_j \subseteq \mathcal{S}^{\text{inv}}$) or not ($\mathcal{S}_j \subseteq \mathcal{S}^{\text{res}}$) on $[n]$. We denote by U^{inv} and U^{res} the submatrices of U formed by the columns that span the invariant and the residual subspace respectively.

Example 2.2.6 (Continued). *In Example 2.2.6 so far, we expressed $\mathcal{S}_1$ and $\mathcal{S}_2$ in terms of their generating vectors. These can be retrieved using Proposition 2.2.8 by joint block diagonalizing the matrices $(\Sigma_t)_{t\in[n]}$. In this specific example, an irreducible joint block diagonalizer is given by*

$$U = \begin{bmatrix} 0.5\sqrt{3} & 0.5 \\ -0.5 & 0.5\sqrt{3} \end{bmatrix},$$

which is a (clockwise) rotation matrix of 30 degrees (see Figure 2.1.1 and use $S_1 = \{1\}$ and $S_2 = \{2\}$). In particular, we have that $\tilde{\Sigma}_t = U^\top\Sigma_t U = \text{diag}(\sigma_{1,t}, \sigma_{2,t})$: therefore, $q^U_{\max} = 2$ and each block has dimension 1. Moreover, it holds for all $t \in [n]$ that

$$\Pi_{\mathcal{S}_1}\gamma_{0,t} = U^{S_1}(U^{S_1})^\top\gamma_{0,t} = \begin{bmatrix} 1.5\sqrt{3} - \sqrt{3}t/n \\ t/n - 1.5 \end{bmatrix} \quad \text{and} \quad \Pi_{\mathcal{S}_2}\gamma_{0,t} = U^{S_2}(U^{S_2})^\top\gamma_{0,t} = \begin{bmatrix} 1 \\ \sqrt{3} \end{bmatrix},$$

and therefore $\mathcal{S}^{\text{inv}} = \mathcal{S}_2$ and $\mathcal{S}^{\text{res}} = \mathcal{S}_1$.

2.2.2 Invariant component

In Theorem 2.2.7 we have shown that the true time-varying parameter $\gamma_{0,t}$ can be expressed as the result of two separate optimization problems over the two orthogonal spaces $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$. In this section we analyze the result of the first optimization step over the invariant subspace $\mathcal{S}^{\text{inv}}$. To ensure that the space $\mathcal{S}^{\text{inv}}$ is unique, we assume that Assumption 1 is satisfied throughout Section 2.2.2.

Definition 2.2.9 (Invariant component). *We denote the parameter that maximizes the explained variance over the invariant subspace by*

$$\beta^{\text{inv}} := \arg\max_{\beta\in\mathcal{S}^{\text{inv}}} \overline{\Delta\text{Var}}(\beta). \tag{2.2.13}$$

The parameter β^{inv} corresponds to the pooled OLS solution obtained by regressing Y_t on the projected predictors $\Pi_{\mathcal{S}^{\text{inv}}}X_t$, and can be computed using all observations in $[n]$. The whole procedure to find β^{inv}, by first identifying $\mathcal{S}^{\text{inv}}$ via joint block diagonalization and then using Proposition 2.2.10 (i) below, is summarized in Algorithm 1 (see Section 2.2.4). Proposition 2.2.10 summarizes some of the properties of β^{inv}.

Proposition 2.2.10 (Properties of β^{inv}). *Under Assumption 1, β^{inv} satisfies the following properties:*

(i) For all $t \in [n]$, $\beta^{\text{inv}} = \Pi_{\mathcal{S}^{\text{inv}}}\gamma_{0,t} = \Pi_{\mathcal{S}^{\text{inv}}}\bar{\gamma}_0$, where $\bar{\gamma}_0 := \frac{1}{n}\sum_{t=1}^n \gamma_{0,t}$.

(ii) β^{inv} is time-invariant over $[n]$, see Definition 2.2.2.

(iii) *If, in addition, it holds for all* $\beta \in \mathbb{R}^p$ *time-invariant over* $[n]$ *that* $\beta \in \mathcal{S}^{\text{inv}}$ *then* $\beta^{\text{inv}} = \arg\max_{\beta \in \mathbb{R}^p \text{ time-invariant}} \overline{\Delta\text{Var}}(\beta)$.

Definition 2.2.2 guarantees, for all $t \in [n]$, that the explained variance for all $\beta \in \mathbb{R}^p$ time-invariant over $[n]$ is $\Delta\text{Var}_t(\beta) = \text{Var}(X^\top \beta) = \beta^\top \Sigma_t \beta$. Point *(ii)* of Proposition 2.2.10 therefore implies that, for all $t \in [n]$, $\Delta\text{Var}_t(\beta^{\text{inv}}) = (\beta^{\text{inv}})^\top \Sigma_t \beta^{\text{inv}}$. Under Assumption 2, we have that the same holds for all $t \in \mathbb{N}$: β^{inv} is therefore a never harmful parameter and for all $t \in \mathbb{N}$ it is a solution to $\arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \Delta\text{Var}_t(\beta)$. Moreover, Proposition 2.2.10 *(iii)* implies that, under an additional assumption, the parameter β^{inv} is optimal, i.e., maximizes the explained variance, among all time-invariant parameters. In particular, β^{inv} represents an interesting target of inference: it can be used for zero-shot prediction, if no adaptation data is available at time t to solve the second part of the optimization in (2.2.12) over $\mathcal{S}^{\text{res}}$. Using U^{inv} as defined in Section 2.2.1.1, we can express β^{inv} as follows.

$$\beta^{\text{inv}} = U^{\text{inv}}((U^{\text{inv}})^\top \overline{\text{Var}}(X) U^{\text{inv}})^{-1} (U^{\text{inv}})^\top \overline{\text{Cov}}(X, Y) \tag{2.2.14}$$

where $\overline{\text{Var}}(X) \coloneqq \frac{1}{n}\sum_{t=1}^n \text{Var}(X_t)$ and $\overline{\text{Cov}}(X, Y) \coloneqq \frac{1}{n}\sum_{t=1}^n \text{Cov}(X_t, Y_t)$. We derive this expression in Lemma 2.6.8, and we later use it for estimation in Section 2.4.2.

Example 2.2.6 (Continued)**.** *Considering again Example 2.2.6, we have that* $\mathcal{S}^{\text{inv}} = \mathcal{S}_2$*, and therefore the invariant component is given by*

$$\beta^{\text{inv}} = \arg\max_{\beta \in \mathcal{S}_2} \overline{\Delta\text{Var}}(\beta) = \Pi_{\mathcal{S}_2} \overline{\gamma}_0 = \begin{bmatrix} 1 \\ \sqrt{3} \end{bmatrix}.$$

Moreover, we can express β^{inv} *under the basis for the irreducible subspaces using the irreducible joint block diagonalizer* U*, obtaining* $U^\top \beta^{\text{inv}} = \begin{bmatrix} 0, & 2 \end{bmatrix}^\top$*: the only non-zero component is indeed the one corresponding to the invariant subspace* $\mathcal{S}_2$*.*

2.2.3 Residual component and time adaptation

Under the generalization assumption (Assumption 2), Theorem 2.2.7 implies that we can partially explain the variance of the response Y_t at all—observed and unobserved—time points $t \in \mathbb{N}$ using β^{inv}. It also implies that for all $t \in \mathbb{N}$ we can reconstruct the true time-varying parameter by adding to β^{inv} a residual parameter that maximizes the explained variance at time t over the residual subspace $\mathcal{S}^{\text{res}}$, i.e.,

$$\gamma_{0,t} = \beta^{\text{inv}} + \arg\max_{\beta \in \mathcal{S}^{\text{res}}} \Delta\text{Var}_t(\beta).$$

In this section we focus on the second optimization step over $\mathcal{S}^{\text{res}}$. We assume that Assumptions 1 and 2 hold, and for all $t \in \mathbb{N}$ we define the residual component $\delta_t^{\text{res}} \coloneqq$

$\arg\max_{\beta \in \mathcal{S}^{\text{res}}} \Delta\text{Var}_t(\beta)$. Using (2.2.7), (2.2.8) and (2.2.11), we can express δ_t^{res} as

$$\begin{aligned} \delta_t^{\text{res}} &= \text{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_t)^\dagger \,\text{Cov}(\Pi_{\mathcal{S}^{\text{res}}} X_t, Y_t) \\ &= \text{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_t)^\dagger \,\text{Cov}(\Pi_{\mathcal{S}^{\text{res}}} X_t, Y_t - X_t^\top \beta^{\text{inv}}). \end{aligned} \tag{2.2.15}$$

We can now reduce the number of parameters that need to be estimated by expressing δ_t^{res} as an OLS solution with $\dim(\mathcal{S}^{\text{res}})$ parameters. To see this, we use that, under Assumption 1, Proposition 2.2.8 allows us to express the space $\mathcal{S}^{\text{res}}$ in terms of an irreducible joint block diagonalizer U corresponding to the irreducible orthogonal partition $\{\mathcal{S}_j\}_{j=1}^{q_{\max}^U}$ as

$$\mathcal{S}^{\text{res}} = \text{span}\{u^k \mid \exists j \in \{1, \dots, q_{\max}^U\} : \mathcal{S}_j \text{ not opt-invariant and } k \in S_j\}.$$

Moreover, the orthogonal projection matrix onto $\mathcal{S}^{\text{res}}$ is given by $\Pi_{\mathcal{S}^{\text{res}}} = U^{\text{res}}(U^{\text{res}})^\top$. Hence, using Lemma 2.6.3 we get that

$$\delta_t^{\text{res}} = U^{\text{res}} \,\text{Var}((U^{\text{res}})^\top X_t)^{-1} \,\text{Cov}((U^{\text{res}})^\top X_t, Y_t - X_t^\top \beta^{\text{inv}}), \tag{2.2.16}$$

where $\text{Var}((U^{\text{res}})^\top X_t)^{-1} \,\text{Cov}((U^{\text{res}})^\top X_t, Y_t - X_t^\top \beta^{\text{inv}})$ is the population ordinary least squares solution obtained by regressing $Y_t - X_t^\top \beta^{\text{inv}}$ on $(U^{\text{res}})^\top X_t \in \mathbb{R}^{\dim(\mathcal{S}^{\text{res}})}$.

Example 2.2.6 (Continued)**.** *In Example 2.2.6, we have that for all $t \in [n]$ the residual component δ_t^{res} is given by*

$$\delta_t^{\text{res}} = \underset{\beta \in \mathcal{S}_1}{\arg\max}\, \Delta\text{Var}_t(\beta) = \Pi_{\mathcal{S}_1} \gamma_{0,t} = \begin{bmatrix} 1.5\sqrt{3} - \sqrt{3}t/n \\ t/n - 1.5 \end{bmatrix}.$$

Moreover, we can express δ_t^{res} under the irreducible orthogonal partition basis (given by the two vectors generating $\mathcal{S}_1$ and $\mathcal{S}_2$), obtaining $U^\top \delta_t^{\text{res}} = \begin{bmatrix} 3 - 2t/n, & 0 \end{bmatrix}^\top$, which indeed only has one degree of freedom in the first component. This component takes the following values: for all $t \in [n]$ we have $3 - 2t/n \in [1, 3]$.

2.2.4 Population ISD algorithm

We call the procedure to identify β^{inv} and δ_t^{res} *invariant subspace decomposition* (ISD). By construction, the result of ISD in its population version is equal to the true time-varying parameter at time $t \in \mathbb{N}$, i.e.,

$$\gamma_{0,t} = \beta^{\text{inv}} + \delta_t^{\text{res}}. \tag{2.2.17}$$

The full ISD procedure is summarized in Algorithm 1. In the algorithm, the invariant and residual subspaces are identified through joint block diagoanlization as described at the end of Section 2.2.1.1. The number of subspaces q in the irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition is inferred by the joint block diagonalization algorithm. If Assumption 1 is not satisfied, then the decomposition in (2.2.17) and therefore the output of Algorithm 1 depends on the irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition used.

Algorithm 1 Population ISD

Input: distributions of $(X_t, Y_t)_{t\in[n]}$

Output: β^{inv}, δ_n^{res}

$(\Sigma_t)_{t\in[n]} \leftarrow \{\text{Var}(\text{X}_\text{t})\}_{t\in[n]}$

$\{\gamma_{0,t}\}_{t\in[n]} \leftarrow \{\text{Var}(\text{X}_\text{t})^{-1}\,\text{Cov}(X_t, Y_t)\}_{t\in[n]}$

$U, \{S_j\}_{j=1}^q \leftarrow \text{irreducibleJointBlockDiagonalizer}((\Sigma_t)_{t\in[n]})$ ▷ Prop. 2.2.8

▷ Find the opt-invariant subspaces to identify $\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}$

$S^{\text{inv}}, S^{\text{res}} \leftarrow \emptyset$

for $j = 1, \dots, q$ **do**

  $\Pi_{\mathcal{S}_j} \leftarrow U^{S_j}(U^{S_j})^\top$

  if $\Pi_{\mathcal{S}_j}\gamma_{0,t}$ constant in $[n]$ **then** $S^{\text{inv}} \leftarrow S^{\text{inv}} \cup S_j$ ▷ see Prop. 2.2.10

  else $S^{\text{res}} \leftarrow S^{\text{res}} \cup S_j$

▷ Define estimators for the invariant and residual component of $\gamma_{0,t}$:

$\Pi_{\mathcal{S}^{\text{inv}}} \leftarrow U^{\text{inv}}(U^{\text{inv}})^\top$

$\beta^{\text{inv}} \leftarrow \frac{1}{n}\sum_{t=1}^n \text{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_t)^\dagger\,\text{Cov}(\Pi_{\mathcal{S}^{\text{inv}}} X_t, Y_t)$ ▷ see Prop. 2.2.10, Lemmas 2.2.4,2.2.5

$\delta_n^{\text{res}} \leftarrow U^{\text{res}}\,\text{Var}((U^{\text{res}})^\top X_n)^{-1}\,\text{Cov}((U^{\text{res}})^\top X_n, Y_n - X_n^\top \beta^{\text{inv}})$ ▷ see Eq. (2.2.16)

In Appendix 2.C we show that a decomposition of the true time-varying parameter similar to (2.2.17) is also obtained when considering $(X_t)_{t\in[n]}$-decorrelating partitions that are not orthogonal, i.e., such that (2.2.6) holds but the subspaces in the partition are not necessarily pairwise orthogonal. In this case, in particular, we can still find an invariant and residual parameter as in (2.2.14) and (2.2.16), respectively, where the matrix U is now a non-orthogonal irreducible joint block diagonalizer.

Remark 2.2.11 (Changes of $\gamma_{0,t}$ and Σ_t over time)**.** *For the ISD procedure to work in practice, we assume that the covariance matrices Σ_t are approximately constant in small time windows (e.g., smoothly changing or piece-wise constant). This is required to obtain accurate estimates of Σ_t, which we then (approximately) joint block diagonalize (see Section 2.4.1.1). Similarly, we assume that $\gamma_{0,t}$ is approximately constant in the adaptation window, in order to perform the adaptation step (see Setting 2.2.1). In the historical data, however, $\gamma_{0,t}$ can vary arbitrarily fast in the residual subspace (β^{inv} remains constant, of course). Intuitively, even if $\gamma_{0,t}$ does not change in a structured way, its projection $\Pi_{\mathcal{S}^{\text{inv}}}\gamma_{0,t}$ on the invariant subspace remains constant across time points. We provide an example in which $\gamma_{0,t}$ is quickly varying in Appendix 2.A.2.*

2.3 Analysis of the two ISD tasks: zero-shot generalization and time adaptation

We now analyze the two prediction tasks that can be solved with the ISD framework introduced in the previous sections, namely (i) the *zero-shot task* and (ii) the *adaptation task*. We consider Setting 2.2.1 and assume for (i) that only the historical data is available, while for (ii) we assume to also have access to the adaptation data. Throughout the remainder of this section we assume that the invariant and residual subspaces are known (they can be computed from the joint distributions), which simplifies the theoretical analysis.

2.3.1 Zero-shot task

We start by analyzing the zero-shot task, in which no adaptation data are observed, but only historical data and X_{t^*}. Under the generalization assumption (Assumption 2), $\mathcal{S}^{\text{inv}}$ is opt-invariant on $\mathbb{N}$ and we can characterize all possible models defined as in (2.1.1) by the possible variations of the true time-varying parameter in the residual space $\mathcal{S}^{\text{res}}$, or, equivalently, by all possible values of $\gamma_{0,t} - \beta^{\text{inv}} \in \mathcal{S}^{\text{res}}$. We then obtain that β^{inv} is worst case optimal in the following sense.

Theorem 2.3.1. *Under Assumptions 1 and 2 it holds for all $t \in \mathbb{N}$ that*

$$\beta^{\text{inv}} = \arg\max_{\beta \in \mathbb{R}^p} \inf_{\substack{\gamma_{0,t} \in \mathbb{R}^p: \\ \gamma_{0,t} - \beta^{\text{inv}} \in \mathcal{S}^{\text{res}}}} \Delta\text{Var}_t(\beta).$$

We further obtain from Proposition 2.2.10 (*iii*) that, under an additional condition, β^{inv} is worst case optimal among all time-invariant parameters. This characterization of β^{inv} allows for a direct comparison with the maximin effect by Meinshausen and Bühlmann [2015], which we report in more detail in Remark 2.3.2. In absence of further information on $\delta^{\text{res}}_{t^*}$, Theorem 2.3.1 suggests to use an estimate of $\hat{\beta}^{\text{inv}}$ of β^{inv} to predict Y_{t^*}, i.e., $\hat{Y}_{t^*} = X_{t^*}^\top \hat{\beta}^{\text{inv}}$.

Remark 2.3.2 (Relation to maximin)**.** *The maximin framework introduced by Meinshausen and Bühlmann [2015] considers a linear regression model with varying coefficients, where the variations do not necessarily happen in a structured way, e.g., in time. Translated to our notation and restricting to time-based changes, their model can be expressed for $t \in [n]$ as*

$$Y_t = X_t^\top \gamma_{0,t} + \epsilon_t,$$

where $\mathbb{E}[\epsilon_t \mid X_t] = 0$ and the covariance matrix $\Sigma := \text{Var}(X_t)$ does not vary with t. The maximin estimator maximizes the explained variance in the worst (most adversarial) scenario that is covered in the training data; it is defined as

$$\beta^{\text{mm}} := \arg\max_{\beta \in \mathbb{R}^p} \min_{t \in [n]} \Delta\text{Var}_t(\beta).$$

The maximin estimator guarantees that, for all $t \in [n]$, $\mathrm{Cov}(Y_t - X_t^\top \beta^{\mathrm{mm}}, X_t^\top \beta^{\mathrm{mm}}) \geq 0$ *and therefore* $\Delta\mathrm{Var}_t(\beta^{\mathrm{mm}}) \geq 0$ *[see Meinshausen and Bühlmann, 2015, Equation (8)]. This means that the maximin relies on a weaker notion of invariance that only requires the left-hand side of (2.2.5) to be non-negative instead of zero. This implies that* β^{inv} *is in general more conservative than* β^{mm} *in the sense that, for all* $t \in [n]$, $\Delta\mathrm{Var}_t(\beta^{\mathrm{mm}}) \geq \Delta\mathrm{Var}_t(\beta^{\mathrm{inv}})$. *By finding the invariant and residual subspaces, we determine the domain in which* $\gamma_{0,t}$ *varies and assume (Assumption 2) that this does not change even at unobserved time points. The parameter* β^{inv} *is then worst case optimal over this domain, and guarantees that the explained variance remains positive even for scenarios that are more adversarial than the ones observed in* $[n]$, *i.e., such that* $\Delta\mathrm{Var}_s(\beta^{\mathrm{mm}}) < \min_{t\in[n]} \Delta\mathrm{Var}_t(\beta^{\mathrm{mm}})$ *for some* $s \in \{n+1, n+2, \dots\}$ *(see, for example, the results of the experiment described in Section 2.5.1 and shown in Figure 2.1.2). Furthermore, the decomposition allows for time adaptation (see Section 2.2.3), which would not be possible starting from the maximin effect.*

2.3.2 Adaptation task

We now consider the adaptation task in which, additionally to the historical data, adaptation data are available. Adaptation data can be used to define an estimator for the residual component $\delta_{t^*}^{\mathrm{res}}$, which we denote by $\hat{\delta}_{t^*}^{\mathrm{res}}$. This, together with an estimator $\hat{\beta}^{\mathrm{inv}}$ for β^{inv} fitted on the historical data gives us an estimator $\hat{\gamma}_{t^*}^{\mathrm{ISD}} := \hat{\beta}^{\mathrm{inv}} + \hat{\delta}_{t^*}^{\mathrm{res}}$ for the true time-varying parameter. We compare this estimator with a generic estimator $\hat{\gamma}_{t^*}$ for γ_{0,t^*} which uses only the adaptation data.

To do so, we consider the minimax lower bound provided by Mourtada [2022][Theorem 1] for the expected squared prediction error of an estimator $\hat{\gamma}$ computed using n i.i.d. observations of a random covariate vector $X \in \mathbb{R}^p$ and of the corresponding response Y. It is given by

$$\inf_{\hat{\gamma}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[(X^\top(\hat{\gamma} - \gamma))^2] \geq \sigma^2 \frac{p}{n}, \tag{2.3.18}$$

where $\mathcal{P}$ is the class of distributions over (X, Y) such that $Y = X^\top\gamma + \epsilon$, $\mathbb{E}[\epsilon \mid X] = 0$ and $\mathbb{E}[\epsilon^2 \mid X] < \sigma^2$. It follows from (2.3.18) that, for a generic estimator $\hat{\gamma}_{t^*}$ of γ_{0,t^*} based on the adaptation data alone, we can expect at best to achieve a prediction error of $\sigma_{\mathrm{ad}}^2 \frac{p}{m}$, where σ_{ad}^2 is the variance of ϵ_t for $t \in \mathcal{I}^{\mathrm{ad}}$ (which is assumed to be constant). We can improve on this if we allow the estimator $\hat{\gamma}_{t^*}$ to also depend on the historical data. To see this, observe that we can always decompose $\hat{\gamma}_{t^*} = \hat{\beta} + \hat{\delta}_{t^*}$ with $\hat{\beta} \in \mathcal{S}^{\mathrm{inv}}$ and $\hat{\delta}_{t^*} \in \mathcal{S}^{\mathrm{res}}$. Moreover, under Assumption 2 we can split the expected prediction error of $\hat{\gamma}_{t^*}$ at t^* accordingly as

$$\mathbb{E}[(X_{t^*}^\top(\hat{\gamma}_{t^*} - \gamma_{0,t^*}))^2] = \mathbb{E}[((\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*})^\top(\hat{\beta} - \beta^{\mathrm{inv}}))^2] + \mathbb{E}[((\Pi_{\mathcal{S}^{\mathrm{res}}} X_{t^*})^\top(\hat{\delta}_{t^*} - \delta_{t^*}^{\mathrm{res}}))^2].$$

Then, $\hat{\beta}$ represents an estimator for β^{inv} and can be computed on historical data[3], whereas

[3] In principle, an estimator for β^{inv} could be computed using both historical and adaptation data. However, the ISD procedure is motivated by scenarios in which the size of historical data is very large (and $n \gg m$): this means that it could be computationally costly to update the estimate for the invari-

$\hat{\delta}_{t^*}$ estimates δ_t^{res} and is based on adaptation data: by decomposing $\hat{\gamma}_{t^*}$ in this way, the best prediction error we can hope for is of the order $\frac{\dim(\mathcal{S}^{\text{inv}})}{n} + \frac{\dim(\mathcal{S}^{\text{res}})}{m}$. In Section 2.4, we prove that $\hat{\gamma}_{t^*}^{\text{ISD}}$ indeed achieves this bound in Theorem 2.4.1. If the invariant subspace is non-degenerate and therefore $\dim(\mathcal{S}^{\text{res}}) < p$, and n is sufficiently large, this implies that $\hat{\gamma}_{t^*}^{\text{ISD}}$ has better finite sample performance than estimators based on the adaptation data alone.

2.4 ISD estimator and its finite sample generalization guarantee

We now construct an empirical estimation procedure for the ISD framework, based on the results of Section 2.2.2 and on Algorithm 1 described in Section 2.2.3. Throughout this section, we assume that Assumptions 1 and 2 are satisfied.

We assume that we observe both historical and adaptation data as in the adaptation task. We use the historical data to first estimate the decomposition of $\mathbb{R}^p$ into $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$, employing a joint block diagonalization algorithm (Section 2.4.1). We then use the resulting decomposition to estimate β^{inv}. Finally, we use the adaptation data to construct an estimator for δ_t^{res} in Section 2.4.2. In Section 2.4.3 we then show the advantage of separating the optimization as in (2.1.3) to estimate γ_{0,t^*} at the previously unobserved time point t^*, by providing finite sample guarantees for the ISD estimator.

2.4.1 Estimating the subspace decomposition

2.4.1.1 Approximate joint block diagonalization

We first need to find a good estimator for the covariance matrices Σ_t. Since only one observation (X_t, Y_t) is available at each time step $t \in [n]$, some further assumptions are needed about how Σ_t varies over time. Here, we assume that Σ_t varies smoothly with $t \in [n]$ and is therefore approximately constant in small time windows. We can then consider a rolling window approach, i.e., consider K windows in $[n]$ of length $w \ll n$ over which the constant approximation is deemed valid, and for the k-th time window, $k \in \{1, \dots, K\}$, take the sample covariance $\hat{\Sigma}_k$ as an estimator for Σ_t in such time window.

Given the set of estimated covariance matrices $\{\hat{\Sigma}_k\}_{k=1}^K$, we now need to estimate an orthogonal transformation $\hat{U}$ that approximately joint block diagonalizes them. We provide an overview of joint block diagonalization methods in Section 2.A.3 in the appendix. In our simulated settings, we solve the approximate joint block diagonalization (AJBD) problem via approximate joint diagonalization (AJD), since we found this approach to represent a good trade off between computational complexity and accuracy. More in detail, similarly to what is proposed by Tichavsky and Koldovsky [2012], we start from the output matrix $V \in \mathbb{R}^{p \times p}$ of the `uwedge` algorithm by Tichavsky and Yeredor [2008],

ant component every time new adaptation data are available, without a significant gain in estimation accuracy. For this reason, we only consider estimators for β^{inv} that use historical data.

which solves AJD for the set $\{\hat{\Sigma}_k\}_{k=1}^K$, i.e., it is such that for all $k \in \{1, \dots, K\}$ the matrix $V^\top \hat{\Sigma}_k V$ is approximately diagonal. We then use the off-diagonal elements of the approximately diagonalized covariance matrices to identify the common diagonal blocks. As described more in detail in Section 2.A.3, this is achieved by finding an appropriate permutation matrix $P \in \mathbb{R}^{p \times p}$ for the columns of V such that for all $k \in \{1, \dots, K\}$ the matrix $(VP)^\top \hat{\Sigma}_k (VP)$ is approximately joint block diagonal. The estimated irreducible joint block diagonalizer is then given by $\hat{U} = VP$. We denote by $q_{\max}^{\hat{U}}$ the number of estimated diagonal blocks and by $\{\hat{\mathcal{S}}_|\}_{j=1}^{q_{\max}^{\hat{U}}}$ the estimated irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition.

2.4.1.2 Estimating the invariant and residual subspaces

We now estimate the invariant and residual subspaces using the estimated irreducible joint block diagonalizer $\hat{U}$. To do so, we first estimate the true time-varying parameter $\gamma_{0,t}$ using similar considerations as the ones made in Section 2.4.1.1. We assume for example that $\gamma_{0,t}$ is approximately constant in small windows (for simplicity, we consider the same K windows defined in Section 2.4.1.1). This assumption is helpful to define the estimation procedure, but is not strictly necessary for the ISD framework to work. We show in Appendix 2.A.2 that the procedure can still work if this assumption is violated. We then compute the regression coefficient $\hat{\gamma}_k$ of $\mathbf{Y}_k$ on $\mathbf{X}_k$, where $\mathbf{Y}_k \in \mathbb{R}^{w \times 1}$ and $\mathbf{X}_k \in \mathbb{R}^{w \times p}$ are the observations in the k-th time window[4]. We use the estimates $\hat{\gamma}_k$ to determine which of the subspaces identified by $\hat{U}$ are opt-invariant. For all $j \in \{1, \dots, q_{\max}^{\hat{U}}\}$, we take the sets of indices S_j as defined in Proposition 2.2.8, and consider the estimated orthogonal projection matrices $\Pi_{\hat{\mathcal{S}}_j} = \hat{U}^{S_j}(\hat{U}^{S_j})^\top$ from $\mathbb{R}^p$ onto the j-th subspace $\hat{\mathcal{S}}_|$ of the estimated irreducible orthogonal partition. It follows from Lemma 2.2.4 that we can find the opt-invariant subspaces by checking for all $j \in \{1, \dots, q_{\max}^{\hat{U}}\}$ whether $\Pi_{\hat{\mathcal{S}}_j} \hat{\gamma}_k$ remains approximately constant for $k \in \{1, \dots, K\}$. To do so, let $\hat{\gamma} := (\sum_{k=1}^K \mathrm{Var}(\hat{\gamma}_k)^{-1})^{-1} \sum_{k=1}^K \mathrm{Var}(\hat{\gamma}_k)^{-1} \hat{\gamma}_k$ be the average of the estimated regression coefficients inversely weighted by their variance. We further use that, by Lemma 2.6.7, if $\hat{\mathcal{S}}_j$ is opt-invariant on $[n]$ then the weighted average of the (approximately constant) projected regression coefficient $\Pi_{\hat{\mathcal{S}}_j} \hat{\gamma}_k$, i.e., $\Pi_{\hat{\mathcal{S}}_j} \hat{\gamma}$, approximately satisfies the time-invariance constraint (2.2.5). This approach is motivated by Proposition 2.2.10 (*iii*). If the corresponding assumption cannot be assumed to hold, other methods can alternatively be used to determine whether $\Pi_{\hat{\mathcal{S}}_j} \hat{\gamma}_k$ is constant, e.g., checking its gradient or variance. In (2.2.5), we can equivalently consider the correlation in place of the covariance, that is, $\mathrm{corr}(Y_t - X_t^\top \beta, X_t^\top \beta) = 0$: an estimate of this correlation allows us to obtain a normalized measure of (2.2.5) that is comparable across different

[4] We have so far omitted the intercept in our linear model, but it can be included by adding a constant term to X_t when estimating the linear parameters. We explicitly show how to take the intercept into account in Section 2.B in the appendix.

experiments. Formally, we consider for all $k \in \{1, \dots, K\}$ and for all $j \in \{1, \dots, q_{\max}^{\hat{U}}\}$

$$\hat{c}_k^j := \widehat{\mathrm{Corr}}(\mathbf{Y}_k - \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma}), \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma}))$$

and check, for all $j \in \{1, \dots, q_{\max}^{\hat{U}}\}$, whether

$$\frac{1}{K}\sum_{k=1}^{K}\left|\hat{c}_k^j\right| \leq \lambda \tag{2.4.19}$$

for some small threshold $\lambda \in [0, 1]$. The threshold λ can be chosen, for example, using cross-validation (more details are provided in Section 2.A.4 in the appendix). An estimator of the invariant and residual subspaces is then given by

$$\hat{\mathcal{S}}^{\mathrm{inv}} = \bigoplus_{\substack{j\in\{1,\dots,q_{\max}^{\hat{U}}\}:\\ (2.4.19)\text{ is satisfied}}} \hat{\mathcal{S}}_j, \quad \hat{\mathcal{S}}^{\mathrm{res}} = \bigoplus_{\substack{j\in\{1,\dots,q_{\max}^{\hat{U}}\}:\\ (2.4.19)\text{ is not satisfied}}} \hat{\mathcal{S}}_j, \tag{2.4.20}$$

where we approximate opt-invariance with the inequality (2.4.19) being satisfied.

2.4.2 Estimating the invariant and residual components

Let $\hat{U}^{\mathrm{inv}}$ and $\hat{U}^{\mathrm{res}}$ be the submatrices of $\hat{U}$ whose columns span $\hat{\mathcal{S}}^{\mathrm{inv}}$ and $\hat{\mathcal{S}}^{\mathrm{res}}$, respectively. We propose to estimate β^{inv} using the following plug-in estimator for (2.2.14)

$$\hat{\beta}^{\mathrm{inv}} := \hat{U}^{\mathrm{inv}}((\hat{U}^{\mathrm{inv}})^\top \mathbf{X}^\top \mathbf{X} \hat{U}^{\mathrm{inv}})^{-1}(\hat{U}^{\mathrm{inv}})^\top \mathbf{X}^\top \mathbf{Y}. \tag{2.4.21}$$

We consider now the new observation of the covariates X_{t^*} at time t^* and the adaptation data $(X_t, Y_t)_{t\in\mathcal{I}^{\mathrm{ad}}}$ introduced in Setting 2.2.1, and denote by $\mathbf{X}^{\mathrm{ad}} \in \mathbb{R}^{m\times p}$ and $\mathbf{Y}^{\mathrm{ad}} \in \mathbb{R}^{m\times 1}$ the matrices containing this adaptation data. Similarly to $\hat{\beta}^{\mathrm{inv}}$, using (2.2.16) we obtain the following plug-in estimator for $\delta_{t^*}^{\mathrm{res}}$

$$\hat{\delta}_{t^*}^{\mathrm{res}} := \hat{U}^{\mathrm{res}}((\hat{U}^{\mathrm{res}})^\top (\mathbf{X}^{\mathrm{ad}})^\top \mathbf{X}^{\mathrm{ad}} \hat{U}^{\mathrm{res}})^{-1}(\hat{U}^{\mathrm{res}})^\top (\mathbf{X}^{\mathrm{ad}})^\top (\mathbf{Y}^{\mathrm{ad}} - \mathbf{X}^{\mathrm{ad}}\hat{\beta}^{\mathrm{inv}}). \tag{2.4.22}$$

We can now define the ISD estimator for the true time-varying parameter at t^* as

$$\hat{\gamma}_{t^*}^{\mathrm{ISD}} := \hat{\beta}^{\mathrm{inv}} + \hat{\delta}_{t^*}^{\mathrm{res}}. \tag{2.4.23}$$

A prediction of the response Y_{t^*} is then given by $\hat{Y}_{t^*} = X_{t^*}^\top \hat{\gamma}_{t^*}^{\mathrm{ISD}}$.

We provide the pseudocode summarizing the whole ISD estimation procedure in the Appendix 2.B.

Example 2.2.6 (Continued)**.** *Assume that in Example 2.2.6 the true time-varying parameter at time points $t \in \{n+1, n+2, \dots\}$ is given by*

$$\gamma_{0,t}^{\mathrm{ad}} = \begin{bmatrix} 1 + 0.5\sqrt{3} - 1.5\sqrt{3}\frac{t-n}{T-n}\sin^2(\frac{t-n}{T-n}+1) \\ \sqrt{3} - 0.5 + 1.5\frac{t-n}{T-n}\sin^2(\frac{t-n}{T-n}+1) \end{bmatrix}$$

for some $T \in \mathbb{N}$. We can verify that $\Pi_{\mathcal{S}_2}\gamma_{0,t}^{\text{ad}} = \Pi_{\mathcal{S}_2}\gamma_{0,t} = \beta^{\text{inv}}$, and therefore the invariant and residual subspaces defined on $[n]$ generalize to $\{n+1, n+2, \dots\}$ and Assumption 2 is satisfied. Moreover, we obtain that for $t \in \{n+1, \dots\}$ the residual component expressed under the irreducible orthogonal partition basis is

$$U^\top \delta_t^{\text{res}} = \begin{bmatrix} 1 - 3\frac{t-n}{T-n}\sin^2(\frac{t-n}{T-n} + 1) \\ 0 \end{bmatrix}.$$

The first entry now takes values in $[-1.5, 1]$, which is disjoint from the range of values of the first coordinate observed in $[n]$ (which was $[1, 3]$).

We now take $T = 350$ and consider an online setup in which we sequentially observe X_t^* at a new time point $t^* \in \{n+1, \dots\}$, and assume that Y_t is observed until $t = t^* - 1$. We take as historical data the observations on $[n]$, and consider as adaptation data the observations in windows $\mathcal{I}^{\text{ad}} = \{t^* - m, \dots, t^* - 1\}$ of length $m = 16$. After estimating β^{inv} on historical data, we use the adaptation data to estimate $\delta_{t^*}^{\text{res}}$ and the OLS solution $\gamma_{t^*}^{\text{OLS}}$. We repeat this online step 350 times (each time increasing t^* by one). Figure 2.1.1 shows the results of this experiment.

2.4.3 Finite sample generalization guarantee

Considering the setting described in Section 2.2.3, we now compare the ISD estimator in (2.4.23) with the OLS estimator computed on $\mathcal{I}^{\text{ad}}$, i.e., $\hat{\gamma}_{t^*}^{\text{OLS}} := ((\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}(\mathbf{X}^{\text{ad}})^\top \mathbf{Y}^{\text{ad}}$. We assume that we are given the (oracle) subspaces $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$, and consider the expected explained variance at t^* of $\hat{\gamma}_{t^*}^{\text{ISD}}$ and $\hat{\gamma}_{t^*}^{\text{OLS}}$. More in detail, the expected explained variance at t^* of an arbitrary estimator $\hat{\gamma}$ of γ_{0,t^*} is given by

$$\mathbb{E}[\Delta\text{Var}_{t^*}(\hat{\gamma})] := \mathbb{E}[\text{Var}(Y_{t^*}) - \text{Var}(Y_{t^*} - X_{t^*}^\top \hat{\gamma} \mid \hat{\gamma})]. \tag{2.4.24}$$

Evaluating (2.4.24) allows us to obtain a measure of the prediction accuracy of $\hat{\gamma}$: the higher the expected explained variance, the more predictive the estimator is. This also becomes clear by isolating in (2.4.24) the term $\mathbb{E}[\text{Var}(Y_{t^*} - X_{t^*}^\top \hat{\gamma} \mid \hat{\gamma})]$, which represents the mean square prediction error obtained by using $\hat{\gamma}$ (see Remark 2.2.3). Our goal is to show that the explained variance by $\hat{\gamma}_{t^*}^{\text{ISD}}$ is always greater than that of the OLS estimator.

Theorem 2.4.1. *Assume Assumption 2 and that, in model (2.1.1), $\gamma_{0,t}$ and the variances of X_t and ϵ_t do not change with respect to $t \in \mathcal{I}^{\text{ad}} \cup t^*$, and denote them by γ_{0,t^*}, Σ_{t^*} and σ_{ad}^2, respectively. Moreover, let $c, \sigma_{\epsilon,\max}^2 > 0$ be constants such that for all $n \in \mathbb{N}$ it holds that $\sigma_{\epsilon,\max}^2 \geq \max_{t\in[n]} \text{Var}(\epsilon_t)$ and for all $m \in \mathbb{N}$ with $m \geq p$, $\lambda_{\min}(\frac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}}) \geq c$ almost surely, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. Further assume that the invariant and residual subspaces $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ are known. Then, there exist $C_{\text{inv}}, C_{\text{res}} > 0$ constants such that for all $n, m \in \mathbb{N}$ with $n, m \geq p$ it holds that*

$$\text{MSPE}(\hat{\gamma}_{t^*}^{\text{ISD}}) := \mathbb{E}[(X_{t^*}(\hat{\gamma}_{t^*}^{\text{ISD}} - \gamma_{0,t^*}))^2] \leq \sigma_{\epsilon,\max}^2 \frac{\dim(\mathcal{S}^{\text{inv}})}{n} C_{\text{inv}} + \sigma_{\text{ad}}^2 \frac{\dim(\mathcal{S}^{\text{res}})}{m} C_{\text{res}}.$$

Furthermore, for all $n, m \in \mathbb{N}$ with $n, m \geq p$ it holds that

$$\mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{OLS}}) - \mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{ISD}}) \geq \sigma_{\mathrm{ad}}^2 \frac{\dim(\mathcal{S}^{\mathrm{inv}})}{m} - \sigma_{\epsilon,\max}^2 \frac{\dim(\mathcal{S}^{\mathrm{inv}})}{n} C_{\mathrm{inv}}.$$

From the proof of Theorem 2.4.1 it follows that C_{res} can be chosen close to 1 if we only consider sufficiently large m. Moreover, because $\mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{OLS}}) - \mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{ISD}}) = \mathbb{E}[\Delta\mathrm{Var}_{t^*}(\hat{\gamma}_{t^*}^{\mathrm{ISD}})] - \mathbb{E}[\Delta\mathrm{Var}_{t^*}(\hat{\gamma}_{t^*}^{\mathrm{OLS}})]$, Theorem 2.4.1 implies that if $\dim(\mathcal{S}^{\mathrm{inv}}) \geq 1$ and n sufficiently large then

$$\mathbb{E}[\Delta\mathrm{Var}_{t^*}(\hat{\gamma}_{t^*}^{\mathrm{ISD}})] > \mathbb{E}[\Delta\mathrm{Var}_{t^*}(\hat{\gamma}_{t^*}^{\mathrm{OLS}})].$$

The first term in the difference between the expected explained variances (or MSPEs) depends on the dimensions of the invariant subspace and of the time-adaptation window $\mathcal{I}^{\mathrm{ad}}$: there is a higher gain in using $\hat{\gamma}_{t^*}^{\mathrm{ISD}}$ instead of $\hat{\gamma}_{t^*}^{\mathrm{OLS}}$ if the dimension of $\mathcal{S}^{\mathrm{inv}}$ is large and only a small amount of time-points are available are available in the adaptation data.

2.5 Experiments

To show the effectiveness of the ISD framework we report the results of two simulation experiments and one real data experiment. The first simulation experiment evaluates the estimation accuracy of the invariant component $\hat{\beta}^{\mathrm{inv}}$ for increasing sample size n of the historical data. The second simulation experiment compares the predictive accuracy of $\hat{\gamma}_t^{\mathrm{ISD}}$ and $\hat{\gamma}_t^{\mathrm{OLS}}$ for different sizes m of the adaptation dataset, to empirically investigate the dependence of the MSPE difference on the size of the adaptation data shown in Theorem 2.4.1.

In both simulation experiments we let the dimension of the covariates be $p = 10$, and $\dim(\mathcal{S}^{\mathrm{inv}}) = 7$, $\dim(\mathcal{S}^{\mathrm{res}}) = 3$, and generate data as follows. We sample a random orthogonal matrix U, and sample the covariates X_t from a normal distribution with zero mean and covariance matrix $U\tilde{\Sigma}_t U^\top$, where $\tilde{\Sigma}_t$ is a block-diagonal matrix with four blocks of dimensions $2, 4, 3$ and 1, and random entries that change 10 times in the observed time horizon n. We take as true time-varying parameter the rotation by U of the parameter with constant entries equal to 0.2 (we set these entries all to the same value for simplicity, but they need not to be equal in general) corresponding to the blocks of sizes 4 and 3, and time-varying entries—corresponding to the blocks of sizes 2 and 1—equal to $(1 - 1.5t\sin^2(it/n + i))/n$, where $i \in \{2, ..., 8\}$ is the entry index (the values of these coefficients range between -0.25 and 1). The noise terms ϵ_t are sampled i.i.d. from a normal distribution with zero mean and variance $\sigma_{\epsilon_t}^2 = 0.64$. The dimensions of the subspaces, $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$, and the true time-varying parameter $\gamma_{0,t}$ are chosen to ensure that, in the historical data, $X_t^\top \beta^{\mathrm{inv}}$ and $X_t^\top \delta_t^{\mathrm{res}}$ explain approximately half of the variance of Y_t each. This choice allows for better visualization in the described experiments.

In Section 2.5.3 we repeat the same experiments on real data coming from a controlled physical system (light tunnel) developed by Gamella et al. [2025]. The data used in this experiment is available at `https://github.com/juangamella/causal-chamber`. The

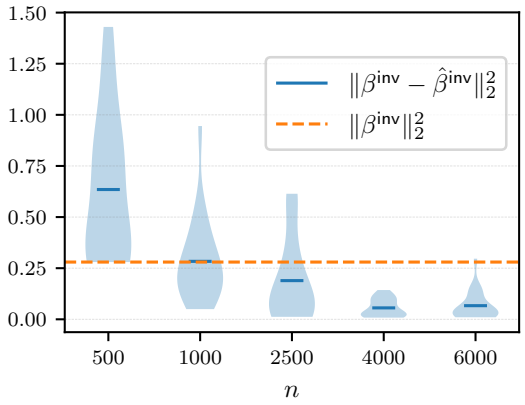


Figure 2.5.4: MSE of $\hat{\beta}^{\text{inv}}$ for increasing size of the historical data n (see Section 2.5.1). For larger values of n, the estimation of the invariant subspace decomposition becomes more precise and leads to smaller errors in the estimated invariant component $\hat{\beta}^{\text{inv}}$.

code for the presented experiments is available at `https://github.com/mlazzaretto/Invariant-Subspace-Decomposition`. The implementation of the `uwedge` algorithm is taken from the Python package `https://github.com/sweichwald/coroICA-python` developed by Pfister et al. [2019b].

2.5.1 Invariant decomposition and zero-shot prediction

We first estimate the time invariant parameter β^{inv} for different sample sizes n of the historical data. We consider $n \in \{500, 1000, 2500, 4000, 6000\}$, and repeat the experiment 20 times for each n. To compute $\hat{\beta}^{\text{inv}}$, we use $K = 25$ equally distributed windows of length $n/8$ (see Section 2.4). Figure 2.5.4 shows that the mean squared error (MSE) $\|\beta^{\text{inv}} - \hat{\beta}^{\text{inv}}\|_2^2$ converges to zero for increasing values of n.

We then consider a separate time window of 250 observations in which the value of the time-varying coefficients (before the transformation using U) is set to -1. We use these observations to test the zero-shot predictive capability of the estimated invariant component, i.e., they can be seen as realizations of the variable X_{t^*} introduced in Setting 2.2.1 (we refer to this window as test data). We compare the predictive performance of the parameter $\hat{\beta}^{\text{inv}}$ on the historical and test data with that of the oracle invariant parameter β^{inv}, the maximin effect $\hat{\beta}^{\text{mm}}$ [computed using the magging estimator proposed by Bühlmann and Meinshausen, 2016], and the OLS solution $\hat{\beta}^{\text{OLS}}$, both computed using the historical data. We show in Figure 2.5.5 the results in terms of the R^2 coefficient, given by $R^2 = \frac{\sum_{t=1}^{n}(\widehat{\text{Var}}(Y_t) - \widehat{\text{Var}}(Y_t - X_t^\top \hat{\beta}))}{\sum_{t=1}^{n} \widehat{\text{Var}}(Y_t)}$.

Figure 2.5.5 shows that the R^2 coefficient of the oracle invariant component β^{inv} remains positive even for values of $\gamma_{0,t}$ in the test data that lie outside of the observed support in the historical data; for increasing n the same holds for the estimated $\hat{\beta}^{\text{inv}}$.

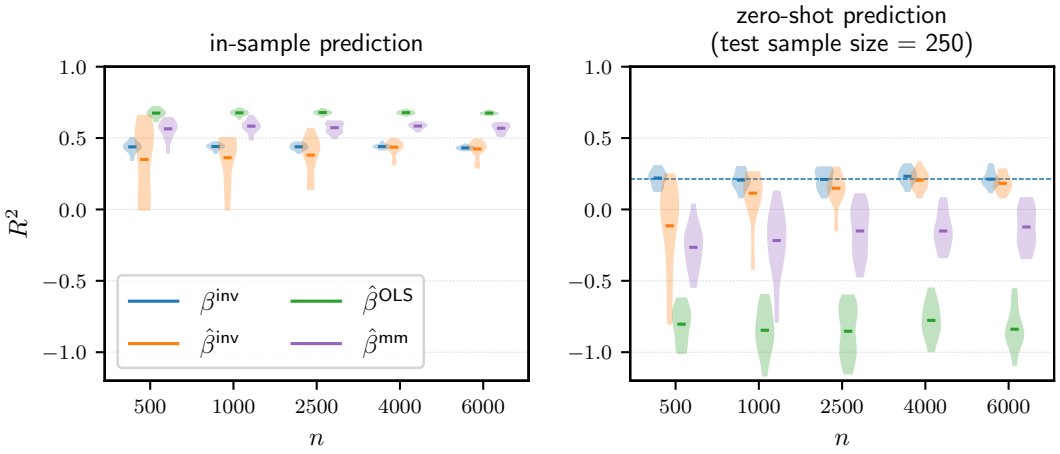


Figure 2.5.5: Normalized explained variance (R^2) by $\hat{\beta}^{\text{inv}}$ and comparison with β^{inv}, $\hat{\beta}^{\text{mm}}$ and $\hat{\beta}^{\text{OLS}}$: training (historical data, left) and zero-shot generalization (test data, right), for different sizes n of the historical data (see Section 2.5.1). The dashed line indicates the population value of the (normalized) explained variance by β^{inv}.

Using $\hat{\beta}^{\text{OLS}}$ or $\hat{\beta}^{\text{mm}}$ leads instead to negative explained variance in this experiment.

The main limitation of the ISD method lies in the estimation of the invariant and residual subspaces. As outlined in Section 2.4, this process consists of two main steps, approximate joint block diagonalization and selection of the invariant blocks, both of which are in practice sensitive to noise. In both steps, we implement our estimator to be as conservative as possible, that is, such that it does not on average overestimate the number of common diagonal blocks or the dimension of $\mathcal{S}^{\text{inv}}$, to avoid including part of the residual subspace into $\hat{\mathcal{S}}^{\text{inv}}$. This behavior is however hard to avoid if the size of the historical dataset is not sufficiently large, therefore requiring large values of n for the ISD framework to work effectively, see Figures 2.5.4 and 2.5.5.

2.5.2 Time adaptation

In the same setting, we now fix the size of the historical dataset to $n = 6000$, which we use to estimate $\hat{\beta}^{\text{inv}}$, and consider a test dataset in which the time-varying coefficients (before the transformation using U) undergo two shifts and take values -0.5 and -2 on two consecutive time windows, each containing 1000 observations. We assume that the test data are observed sequentially, and take as adaptation data a rolling window of length m contained in the test data and shifting by one time point at the time. We use these sequential adaptation datasets to estimate the residual parameter $\hat{\delta}_t^{\text{res}}$ and the OLS solution $\hat{\gamma}_t^{\text{OLS}}$. We then compute the squared prediction error of $\hat{\gamma}_t^{\text{ISD}} = \hat{\beta}^{\text{inv}} + \hat{\delta}_t^{\text{res}}$ and $\hat{\gamma}_t^{\text{OLS}}$ on the next data point X_{t+1}, i.e., $(X_{t+1}^\top(\gamma_{0,t+1} - \hat{\gamma}_t))^2$ and approximate the corresponding MSPE using a Monte-Carlo approximation with 1000 draws from X_{t+1}

(these correspond to the 1000 sequential observations in each of the two windows of the test data). We repeat the simulation 20 times for different sizes of the adaptation window, $m \in \{1.5p, 2p, 5p, 10p\}$, and plot the obtained MSPE values for $\hat{\gamma}_t^{\text{OLS}}$ against $\hat{\gamma}_t^{\text{ISD}}$. The result is shown in Figure 2.5.6, and empirically supports Theorem 2.4.1, in particular that the difference in the MSPEs of the OLS and ISD estimators is proportional to the ratio $\frac{\dim(\mathcal{S}^{\text{inv}})}{m}$ and is therefore larger for small values of m and shrinks for increasing size of the adaptation data (additional details on this simulation are provided in Section 2.A.5 in the Appendix). The ISD framework is particularly helpful in scenarios in which the size

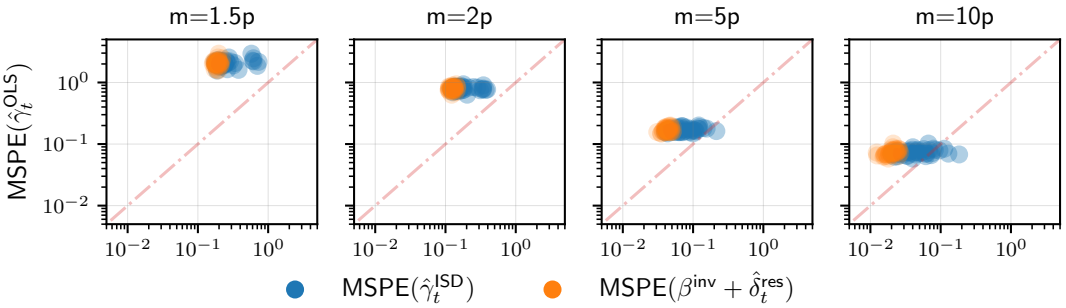


Figure 2.5.6: MSPE comparison: $\hat{\gamma}_t^{\text{ISD}}$ (blue dots) vs. $\hat{\gamma}_t^{\text{OLS}}$ for $p = 10$ and various adaptation window lengths m (see Section 2.5.2). The ISD estimator achieves lower MSPE than the OLS for smaller sizes m of the adaptation window, while the two estimators become comparable for increasing m. The orange dots show the MSPE of β^{inv}(oracle) $+ \hat{\delta}_t^{\text{res}}$ vs. $\hat{\gamma}_t^{\text{OLS}}$: if the subspace decomposition is known, then the ISD always achieves lower MSPE than the OLS.

of the available adaptation window is small (two first plots from the left in Figure 2.5.6). Indeed, from Theorem 2.4.1 it also follows that the larger the dimension of the invariant subspace the greater the advantage in using the ISD framework for prediction rather than a naive OLS approach. A further benefit of the ISD estimator is that it allows us to estimate $\hat{\delta}_t^{\text{res}}$ for small lengths m of the adaptation window where $\dim(\mathcal{S}^{\text{res}}) < m < p$ and OLS is not feasible.

We run a similar experiment to show (Figure 2.5.7) the average cumulative explained variance on the adaptation data over 20 runs, both by estimators computed only on the historical data and estimators that use the adaptation data. For visualization purposes, we now consider the time-varying coefficients (before the transformation using U) equal to $(0.5 - t\sin^2(it/n + i))/n$, where $i \in \{2, ..., 8\}$ is the coefficient index, in the historical data, and constantly equal to -0.3, -0.65 and -1 in three consecutive time windows of size 150 on the time points after the historical data. We estimate $\hat{\delta}_t^{\text{res}}$ and $\hat{\gamma}_t^{\text{OLS}}$ on a rolling adaptation window of size $m = 3p$. The plot in Figure 2.5.7 shows that, on average, the ISD framework, by exploiting invariance properties in the observed data, allows us to accurately explain the variance of the response by using small windows for time adaptation, significantly improving on the OLS solution in the same time windows.

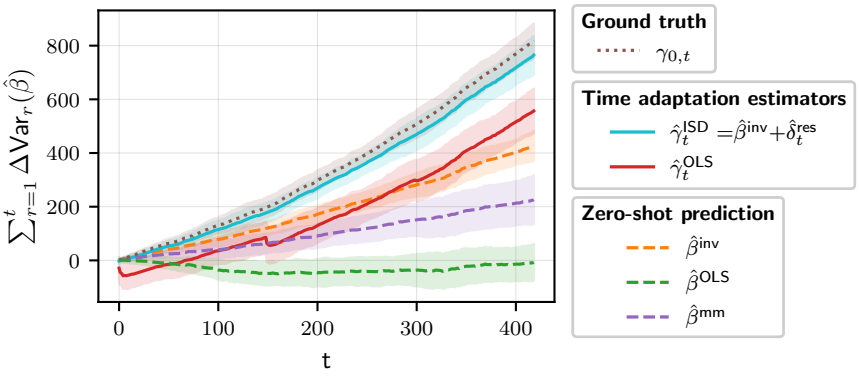


Figure 2.5.7: Average cumulative explained variance on the adaptation data and one standard deviation intervals (over 20 runs) by the true time-varying parameter $\gamma_{0,t}$ and various estimators (see Section 2.5.2). When few data points are available for adaptation at time t, the explained variance of the ISD estimator is significantly higher than that of the rolling window OLS, and improves on invariance-based estimators such as the invariant component $\hat{\beta}^{\text{inv}}$ or the maximin $\hat{\beta}^{\text{mm}}$. Due to a shift of $\gamma_{0,t}$ in the residual subspace, the OLS computed on historical data can perform worse than the zero function.

2.5.3 Real data example

We now present a toy example that applies ISD to real data. The data used for this experiment are collected using a controlled physical system developed by Gamella et al. [2025]. The system, shown in Figure 2.5.8, consists of a light tunnel with a light source X_{RGB} whose emitted light passes through two polarizers with relative angle θ between them and is captured by a sensor placed at the end of the tunnel, behind the polarizers. At the end of the tunnel there are two additional LED light sources, $X_{L_{31}}$ and $X_{L_{32}}$, whose emitted light is unaffected by the polarizers. The sensor $Y := \tilde{I}_3$ measures the overall infrared light at the end of the tunnel, which is affected by the intensity of the RGB source and by the two LEDs. As described by Gamella et al. [2025], the effect of X_{RGB} on Y is linear and depends on θ, more precisely, it holds that $Y \propto \cos^2(\theta) X_{\text{RGB}}$ The dependence of Y on the two LEDs is instead independent of θ. We take Y as our response, consider the covariates vector $[X_{\text{RGB}}, X_{L_{31}}, X_{L_{32}}]^\top \in \mathbb{R}^3$ and assume that the angle θ is unknown. Since we control the three light sources independently, and we expect the dependence of the infrared light Y on the two LEDs to remain the same across time, we hope to detect a nontrivial invariant subspace related to the two LED covariates.

The available dataset contains 8000 observations, collected under changing values of the angle θ. The historical dataset contains the first 7000 observations, and the test dataset the remaining 1000. Figure 2.5.9 shows the dependence of the response on the three covariates, as well as the values of the response and of $\cos^2(\theta)$ through time.

We apply ISD on the historical data to find an invariant component $\hat{\beta}^{\text{inv}}$. For comparison, we also compute the maximin $\hat{\beta}^{\text{mm}}$[Bühlmann and Meinshausen, 2016] and the OLS

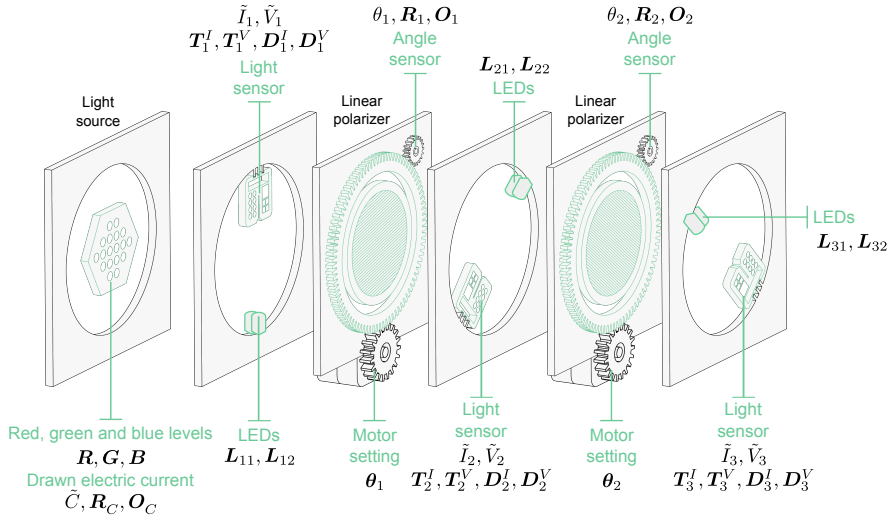


Figure 2.5.8: Illustration of the light tunnel, see Section 2.5.3. The figure is taken from Gamella et al. [2025] (published under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/)). The variables of interest in our experiment are the RGB values of the light source, the LEDs intensities L_{31} and L_{32}, and the measurement of the light sensor at the end of the tunnel $\tilde{I}_3$.

R^2	historical data	test data (zero-shot prediction)
$\hat{\beta}^{\text{inv}}$	0.019	0.062
$\hat{\beta}^{\text{OLS}}$	0.558	-10.703
$\hat{\beta}^{\text{mm}}$	0.477	-2.194

Table 2.5.1: Normalized explained variance (R^2) by $\hat{\beta}^{\text{inv}}$ and comparison with $\hat{\beta}^{\text{OLS}}$ and $\hat{\beta}^{\text{mm}}$: training (historical data) and zero-shot generalization (test data); see Section 2.5.3.

solution $\hat{\beta}^{\text{OLS}}$ on historical data. We then use these estimated parameters for zero-shot prediction on test data. Table 2.5.1 shows the R^2 coefficient, defined as in Section 2.5.1 as the fraction of explained variance $R^2 = \frac{\sum_{t=1}^{n}(\widehat{\text{Var}}(Y_t) - \widehat{\text{Var}}(Y_t - X_t^\top \hat{\beta}))}{\sum_{t=1}^{n} \widehat{\text{Var}}(Y_t)}$, on historical and test data. The invariant component $\hat{\beta}^{\text{inv}}$ is the only estimator that achieves positive explained variance on test data. This is because $\hat{\beta}^{\text{inv}}$ only captures the parts of the variance that can be transferred to the test data, as can be seen from the lower explained variance compared to $\hat{\beta}^{\text{OLS}}$ on the historical data. The estimated invariant subspace $\hat{\mathcal{S}}^{\text{inv}} = \text{span}\{[-0.13910168, -0.95836843, -0.24936055]^\top\}$ has dimension 1 and shows in particular that most of the invariant information is encoded in the two LEDs, with highest weight given to $X_{L_{31}}$. This is expected since the LEDs intensities are not affected by the changing angle between the two polarizers. Also the relatively small R^2 is ex-

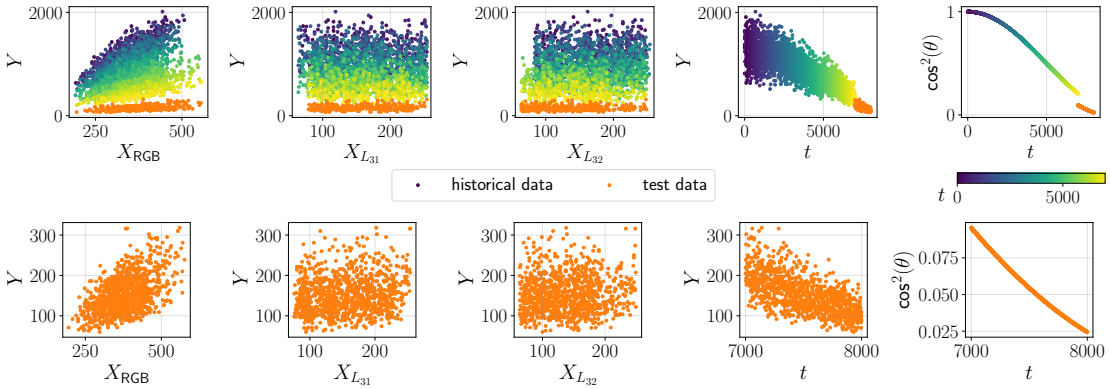


Figure 2.5.9: Dataset of the experiment discussed in Section 2.5.3. The top four figures from the left show the dependence of the infrared measurement Y on the covariates and on time t (encoded by the colormap) for historical data. The angle of the polarizers changes over time (top right) and thus has an influence on the linear relationship between Y and X_{RGB} (top left). For this experiment we assume that the angle θ is unknown. The second row shows the same quantities during test time, where the polarization angle is much smaller (bottom right). The dependence between Y and the two LEDs is small but significant (testing a reduced linear model without either L_{31} or L_{32} against the full model results in p-values smaller than 10^{-4}, both for historical and test data).

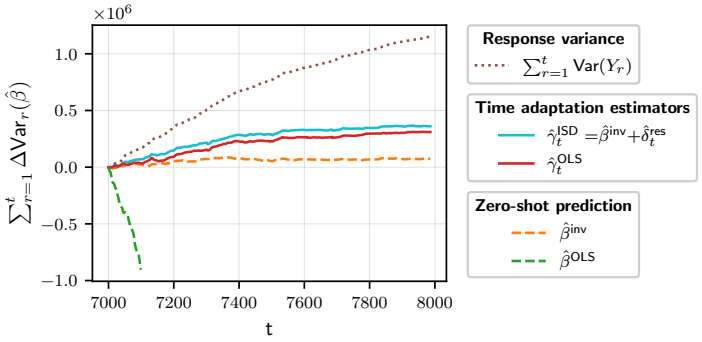


Figure 2.5.10: Cumulative explained variance on the adaptation data by the time adaptation estimators $\hat{\gamma}^{\text{ISD}}$ and $\hat{\gamma}^{\text{OLS}}$ and by the zero-shot predictors $\hat{\beta}^{\text{inv}}$ and $\hat{\beta}^{\text{OLS}}$; see Section 2.5.3.

pected: since the RGB light source is stronger than the two LEDs at the end of the tunnel, most of the variance in Y is explained by the non-invariant component X_{RGB} (see Figure 2.5.9). However, it is not the case that the whole subspace spanned by the two covariates corresponding to the LEDs is invariant, as we would have assumed from

the knowledge of the physical system. An explanation is that due to the data collection process there is nonzero observed correlation between X_{RGB} and $X_{L_{32}}$, but not between X_{RGB} and $X_{L_{31}}$ (on the historical data, we have $\text{corr}(\text{X}_{\text{RGB}}, \text{X}_{\text{L}_{31}}) = 0.004$ with p-value 0.723 and $\text{corr}(\text{X}_{\text{RGB}}, \text{X}_{\text{L}_{32}}) = -0.107$ with p-value smaller than 10^{-15}).

We also run the adaptation step considering as adaptation data a rolling window of size $m = 8$ shifting through the test data. We show in Figure 2.5.10 the cumulative explained variance obtained by ISD, by the OLS solution computed on the same adaptation data and by the OLS solution and the invariant component computed on historical data. The plot shows that $\hat{\gamma}^{ISD}$ achieves the highest explained variance, with a small improvement on the rolling window OLS $\hat{\gamma}^{OLS}$. Indeed, in this particular example the size of the invariant subspace is small ($\dim(\mathcal{S}^{\text{inv}}) = 1$) and by Theorem 2.4.1, we expect only a small improvement.

2.6 Summary

We propose Invariant Subspace Decomposition (ISD), a framework for invariance-based time adaptation. Our method relies on the orthogonal decomposition of the parameter space into an invariant subspace $\mathcal{S}^{\text{inv}}$ and a residual subspace $\mathcal{S}^{\text{res}}$, such that the maximizer of the explained variance over $\mathcal{S}^{\text{inv}}$ is time-invariant. The estimation of the invariant component β^{inv} on a large historical dataset and the reduced dimensionality of $\mathcal{S}^{\text{res}}$ with respect to the original parameter space $\mathbb{R}^p$ allow the ISD estimator to improve on the prediction accuracy of existing estimation techniques. We provide finite sample guarantees for the proposed estimation method and additionally support the validity of our theoretical results through simulated experiments and one real world data experiment.

Future developments of this work may investigate the presented problem in the case of nonlinear time-varying models, and study how to incorporate the ISD framework in specific applied settings such as contextual bandits.

Acknowledgments

We thank Juan Gamella for generating the data for the presented real world example, and the three anonymous reviewers for the valuable comments. NP and ML are supported by a research grant (0069071) from Novo Nordisk Fonden.

Supplement to 'Invariant Subspace Decomposition'

2.A. Supporting examples and remarks

2.A.1 Example of non-uniqueness of an irreducible orthogonal partition

Assume that for all $t \in [n]$ the covariance matrix Σ_t of X_t takes one of the following two values (and each value is taken at least once in $[n]$)

$$\Sigma^1 := \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \Sigma^2 := \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Define for all $j \in \{1, 2, 3\}$ the linear spaces $\mathcal{S}_j = \langle e_j \rangle$, where e_j is the j-th vector of the canonical basis. Then since Σ^1 and Σ^2 are (block) diagonal the partition $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ is an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Consider now the orthonormal matrix

$$U := \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

It holds that

$$U^\top \Sigma^1 U = \Sigma^1$$
$$U^\top \Sigma^2 U = \Sigma^2.$$

Therefore, the spaces

$$\tilde{\mathcal{S}}_1 = \langle \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix} \rangle, \quad \tilde{\mathcal{S}}_2 = \langle \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} \rangle, \quad \tilde{\mathcal{S}}_3 = \langle \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \rangle$$

also form an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition (this follows, for example, from Proposition 2.2.8) but $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\} \neq \{\tilde{\mathcal{S}}_1, \tilde{\mathcal{S}}_2, \tilde{\mathcal{S}}_3\}$.

Then, if we assume for example that $\gamma_{0,t} = \begin{bmatrix} 1, & t, & 1 \end{bmatrix}^\top$, given the first partition we obtain $\mathcal{S}^{\text{inv}} = \mathcal{S}_1 \oplus \mathcal{S}_3$, whereas given the second partition it holds that $\Pi_{\tilde{\mathcal{S}}_1} \gamma_{0,t} = \begin{bmatrix} \frac{1-t}{2}, & \frac{t-1}{2}, & 0 \end{bmatrix}^\top$ and $\Pi_{\tilde{\mathcal{S}}_2} \gamma_{0,t} = \begin{bmatrix} \frac{1+t}{2}, & \frac{1+t}{2}, & 0 \end{bmatrix}^\top$ and therefore $\tilde{\mathcal{S}}^{\text{inv}} = \tilde{\mathcal{S}}_3$, leading to Assumption 1 not being satisfied.

2.A.2 Example of quickly varying $\gamma_{0,t}$ and zero-shot generalization

We repeat the same simulation described in Section 2.5.1 but now consider non-smooth variations of $\gamma_{0,t}$. More specifically, the only difference from the simulation described in Section 2.5.1 is that we let the 3 time-varying entries of $\gamma_{0,t}$ (prior to its rotation by U)

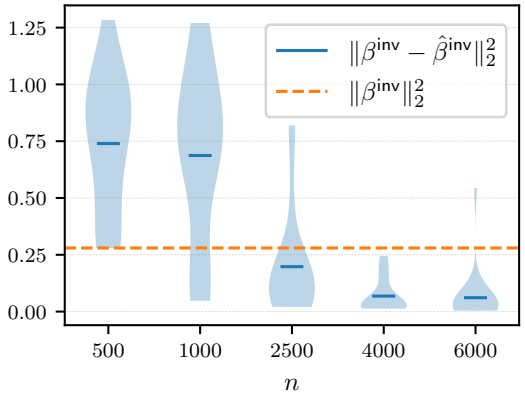


Figure 2.A.1: MSE of $\hat{\beta}^{\text{inv}}$ for increasing size of the historical data n. For larger values of n, the estimation of the invariant subspace decomposition becomes more precise and leads to smaller errors in the estimated invariant component $\hat{\beta}^{\text{inv}}$.

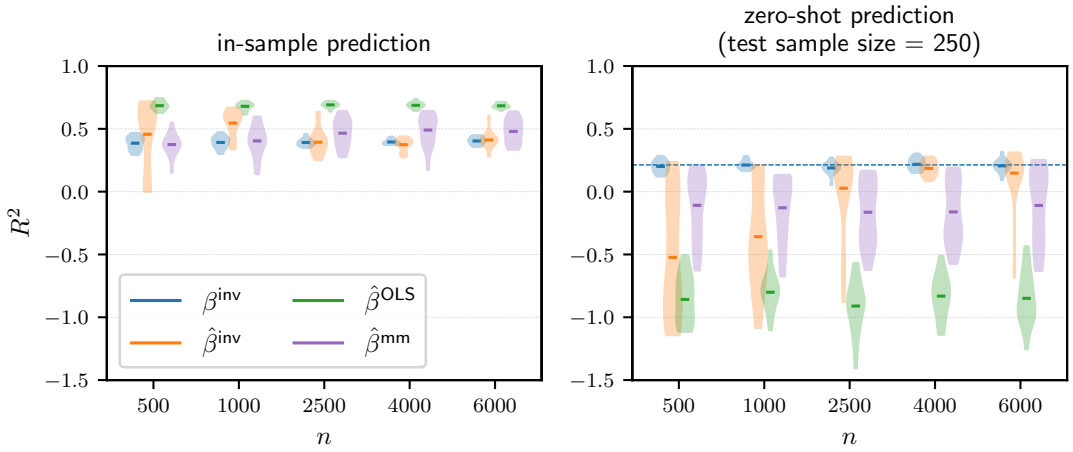


Figure 2.A.2: Normalized explained variance (R^2) by $\hat{\beta}^{\text{inv}}$ and comparison with β^{inv}, $\hat{\beta}^{\text{mm}}$ and $\hat{\beta}^{\text{OLS}}$: training (historical data, left) and zero-shot generalization (test data, right), for different sizes n of the historical data. The dashed line indicates the population value of the (normalized) explained variance by β^{inv}. In the historical data, $\gamma_{0,t}$ is quickly varying.

vary quickly with t. More specifically, at each time point each one of the 3 entries is sampled uniformly in an interval of width 1 centered around a value which changes 20 times in the observed time horizon n. These centers are randomly sampled in $[0, 1.2]$. These choices of the sampling intervals ensure that the experiments are comparable

for different sizes n of the historical data and that there is a shift in $\gamma_{0,t}$ outside of the observed support in the test data (which is generated as in Section 2.5.1). This experiment supports Remark 2.2.11, showing that we do not need assumptions on the type of changes in $\gamma_{0,t}$ in the historical data.

In particular, Figure 2.A.1 shows that the MSE of $\hat{\beta}^{\text{inv}}$ converges to zero for increasing values of n, that is, we are able to estimate the invariant component even when $\gamma_{0,t}$ is quickly varying in the historical data.

Moreover, Figure 2.A.2 shows the results in terms of the R^2 coefficient. In particular, for n large enough, the R^2 coefficient of the estimated invariant component $\hat{\beta}^{\text{inv}}$ remains positive even for values of the test data that lie outside of the observed support.

2.A.3 Methods and computational complexity of joint block diagonalization

In the context of joint block diagonalization, we can differentiate between methods that solve the exact problem (JBD), i.e., are such that the transformed matrices have exactly zero off-block diagonal entries, and approximate methods (AJBD), which assume the presence of noise and aim to minimize the off-block diagonal entries, without necessarily setting them to zero.

JBD is in general an easier problem, and algorithms that solve it have been shown to achieve polynomial complexity [see, e.g., Murota et al., 2010, Tichavskỳ et al., 2012]. Many of these methods, e.g., the one presented by Murota et al. [2010], are based on eigenvalue decompositions. Alternatively, as shown by Tichavskỳ et al. [2012], some algorithms that solve the problem of approximate joint diagonalization (AJD) of a set of matrices, such as `uwedge` developed by Tichavsky and Yeredor [2008], can also be used for JBD. More in detail, a solution to AJD is a matrix that maximally jointly diagonalizes a set of matrices by minimizing the average value of the off-diagonal entries; if the matrices in the set cannot be exactly jointly diagonalized, the transformed matrices will have some non-zero off-diagonal elements. Adding an appropriate permutation of the columns of the joint diagonalizer allows to reorganize the non-zero off-diagonal elements into blocks, leading to jointly block diagonal matrices: in Section 4.3 of their work Gutch and Theis [2012] argue that if the matrices to be block diagonalized are symmetric, then the solution found in this way is also an optimal JBD solution.

However, in general, methods for JBD cannot be directly applied to solve AJBD. Algorithms that solve AJBD directly—based on the iterative optimization of a cost function via matrix rotations—have been developed, for example, by Tichavsky and Koldovsky [2012] and Févotte and Theis [2007], but require the number of diagonal blocks to be known in advance. Alternatively and similarly to how JBD can be solved by AJD methods, one can also, with some slight modifications, use AJD methods to solve AJBD. More specifically, some heuristics need to be used to determine the size of the blocks: these can consist, for example, in setting a threshold for the non-zero off-block-diagonal elements in the transformed matrices.

In the (orthogonal) settings considered here, we have found the last approach to work effectively. More specifically, in Section 2.4.1.1, we have denoted by V the AJD solution

for the set of estimated covariance matrices $\{\hat{\Sigma}_k\}_{k=1}^K$. Similarly to how Tichavsky and Koldovsky [2012] suggest to determine a permutation of the AJD result, we proceed in the following way. To discriminate the non-zero off-block-diagonal entries in these matrices, we start by computing the following auxiliary matrix using V

$$\Sigma := \max_{k\in\{1,\dots,K\}} |V^\top \hat{\Sigma}_k V|,$$

where the maximum is taken element-wise. The matrix Σ captures in its off-diagonal entries the residual correlation among the components identified by AJD, for all the K jointly diagonalized matrices. For all thresholds $\tau \in \mathbb{R}$, we let $P(\tau) \in \mathbb{R}^{p\times p}$ denote one of the permutation matrices satisfying that $P(\tau)^\top \Sigma P(\tau)$ is a block diagonal matrix if all entries smaller than τ are considered zero. We then define the optimal threshold τ^* by

$$\tau^* \in \arg\min_\tau \frac{1}{K}\sum_{k=1}^K \mathrm{OBD}_\tau(|(VP(\tau))^\top \hat{\Sigma}_k (VP(\tau))|) + \nu \frac{p_{bd}(\tau)}{p^2}$$

where $\mathrm{OBD}_\tau(\cdot)$ denotes the average value of off-block-diagonal entries (determined by the threshold τ) of a matrix, $p_{bd}(\tau)$ is the total number of entries in the blocks induced by τ and $\nu \in \mathbb{R}$ is a regularization parameter. The penalization term $\frac{p_{bd}(\tau)}{p^2}$ is introduced to avoid always selecting a zero threshold, and the regularization parameter is set to $\nu = \frac{1}{K}\sum_{k=1}^K \lambda_{\min}(\hat{\Sigma}_k)$ where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue. The optimal permutation is then $P^* := P(\tau^*)$ and the estimated irreducible joint block diagonalizer is $\hat{U} = VP^*$.

2.A.4 Threshold selection for opt-invariant subspaces

In the simulations, we select the threshold λ in (2.4.19) by cross-validation. More in detail, we define the grid of possible thresholds λ by

$$\Lambda := \left\{0, \tfrac{1}{K}\textstyle\sum_{k=1}^K |\hat{c}_k^1|, \dots, \tfrac{1}{K}\sum_{k=1}^K |\hat{c}_k^{q_{\max}^{\hat{U}}}|\right\},$$

where, for all $j \in \{1,\dots,q_{\max}^{\hat{U}}\}$ and $k \in \{1,\dots,K\}$, $\hat{c}_k^j := \widehat{\mathrm{Corr}}(\mathbf{Y}_k - \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma}), \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma}))$. We then split the historical data into $L = 10$ disjoint blocks of observations, and for all $j \in \{1,\dots,J\}$ denote by $\mathbf{X}_\ell$ and $\mathbf{Y}_\ell$ the observations in the ℓ-th block and by $\mathbf{X}_{-\ell}$ and $\mathbf{Y}_{-\ell}$ the remaining historical data. For all possible thresholds $\lambda \in \Lambda$ we proceed in the following way. For all folds $\ell \in \{1,\dots,L\}$, we compute an estimate for the invariant component as in (2.4.20) using $\mathbf{X}_{-\ell}$ and $\mathbf{Y}_{-\ell}$, which we denote by $\hat{\beta}^{\mathrm{inv},-\ell}(\lambda)$. Inside the left-out ℓ-th block of observations, we then consider a rolling window of length $d = 2p$ and the observation at t^* immediately following the rolling window: we compute the residual parameter $\hat{\delta}_{t^*}^{\mathrm{res}}(\lambda)$ as in (2.4.22) using the d observations in the rolling window, and evaluate the empirical explained variance by $\hat{\beta}^{\mathrm{inv},-\ell}(\lambda) + \hat{\delta}_{t^*}^{\mathrm{res}}(\lambda)$ on the observation at t^*, i.e.,

$$\widehat{\Delta\mathrm{Var}}_{t^*}(\lambda) := (Y_{t^*})^2 - (Y_{t^*} - X_{t^*}^\top(\hat{\beta}^{\mathrm{inv},-\ell}(\lambda) + \hat{\delta}_{t^*}^{\mathrm{res}}(\lambda)))^2.$$

We repeat this computation for all possible $t^* \in \mathcal{I}_\ell$, where $\mathcal{I}_\ell$ denotes the time points in the ℓ-th block of observations excluding the first d observations, and define $\overline{\Delta\text{Var}}_\ell^\lambda := \frac{1}{|\mathcal{I}_\ell|}\sum_{t^*\in\mathcal{I}_\ell}\widehat{\Delta\text{Var}}_{t^*}(\lambda)$. For all $\lambda \in \Lambda$, we denote the average explained variance over the L folds by $\overline{\Delta\text{Var}}^\lambda := \frac{1}{L}\sum_{\ell=1}^L \overline{\Delta\text{Var}}_\ell^\lambda$ and the standard error (across the L folds) of such explained variance as $\text{se}(\overline{\Delta\text{Var}}^\lambda) := \frac{1}{L}\sqrt{\sum_{\ell=1}^L(\overline{\Delta\text{Var}}_\ell^\lambda - \overline{\Delta\text{Var}}^\lambda)^2}$. Moreover, let $\lambda^{\max} := \arg\max_{\lambda\in\Lambda}\overline{\Delta\text{Var}}^\lambda$. Then, we choose the optimal threshold as

$$\lambda^* := \min\left\{\lambda \in \Lambda \,\middle|\, \overline{\Delta\text{Var}}^\lambda > \overline{\Delta\text{Var}}^{\lambda^{\max}} - t_{\text{se}}\,\text{se}(\overline{\Delta\text{Var}}^{\lambda^{\max}})\right\},$$

which is the most conservative (lowest) threshold such that the corresponding explained variance is within t_{se} (in our simulations, we choose $t_{\text{se}} = 1$) standard errors (computed across the folds) of the maximal explained variance.

2.A.5 Further simulation details: MSPE comparison

In Section 2.5.2 we present a simulated experiment in which we compare the ISD estimator and the OLS estimator on the time adaptation task. Figure 2.5.6 shows that the difference in the MSPE for $\hat{\gamma}_t^{\text{OLS}}$ and $\hat{\gamma}_t^{\text{ISD}}$ is positive and decreases for increasing values of m. To further support the statement of Theorem 2.4.1, we show in Figure 2.A.3 the value of such difference against $\sigma_{\text{ad}}^2\frac{\dim(\mathcal{S}^{\text{inv}})}{m}$, when computing $\hat{\gamma}_t^{\text{ISD}}$ both with the estimated and oracle invariant component. The figure shows that the difference in the MSPEs indeed satisfies the bound stated in Theorem 2.4.1, i.e., it is always greater than $\sigma_{\text{ad}}^2\frac{\dim(\mathcal{S}^{\text{inv}})}{m}$. Moreover, it shows that for small values of m the gain in using the ISD estimator over the OLS is even higher than what the theoretical bound suggests, indicating that it is not sharp for small m. We further show in Figure 2.A.4 the MSPE of $\hat{\gamma}_t^{\text{ISD}}$ (again computed both with the estimated and oracle invariant component) against $\sigma_{\text{ad}}^2\frac{\dim(\mathcal{S}^{\text{res}})}{m}$, obtaining in this case an empirical confirmation of the first bound presented in Theorem 2.4.1.

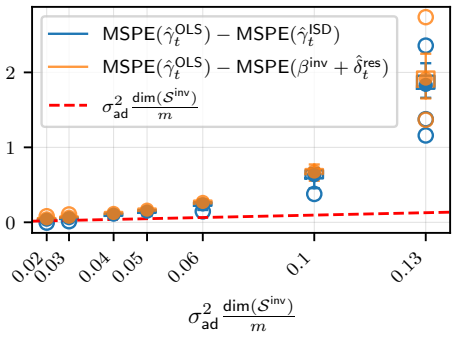


Figure 2.A.3: Difference in the MSPE of $\hat{\gamma}_t^{\text{OLS}}$ and $\hat{\gamma}_t^{\text{ISD}}$ (and an oracle version of $\hat{\gamma}_t^{\text{ISD}}$ based on the true β^{inv}) with respect to $\sigma_{\text{ad}}^2 \frac{\dim(\mathcal{S}^{\text{inv}})}{m}$ for different values of m. The computed difference is larger than $\sigma_{\text{ad}}^2 \frac{\dim(\mathcal{S}^{\text{inv}})}{m}$ for all values of m (in the oracle case), empirically confirming the lower bound obtained in Theorem 2.4.1. The filled dots in the boxplots show the mean over 20 runs (while the empty dots represent the outliers).

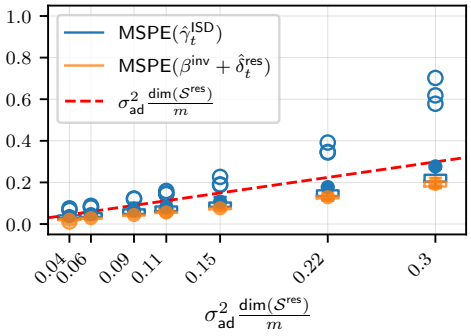


Figure 2.A.4: MSPE of $\hat{\gamma}_t^{\text{ISD}}$ (and an oracle version of $\hat{\gamma}_t^{\text{ISD}}$ based on the true β^{inv}) with respect to $\sigma_{\text{ad}}^2 \frac{\dim(\mathcal{S}^{\text{res}})}{m}$ for different values of m. The computed MSPE is smaller than $\sigma_{\text{ad}}^2 \frac{\dim(\mathcal{S}^{\text{res}})}{m}$ for all values of m (in the oracle case), empirically confirming the upper bound obtained in Theorem 2.4.1. The filled dots in the boxplots show the mean over 20 runs (while the empty dots represent the outliers).

2.B. ISD estimation algorithm

We provide here the pseudocode summarizing the ISD procedure described in Section 2.4. The algorithm includes the estimation of the intercept, that is, it considers for all $t \in \mathbb{N}$ the model

$$Y_t = \gamma^0_{0,t} + X_t^\top \gamma_{0,t} + \epsilon_t \tag{2.B.1}$$

satisfying the assumptions of Setting 2.2.1 but with the addition of $\gamma^0_{0,t} \in \mathbb{R}$. In the

Algorithm 2 ISD: estimation

Input: observations $(X_t, Y_t)_{t\in[n]\cup\mathcal{I}_{\text{ad}}}$, X_{t^*}, number of windows K, $\lambda \in [0,1]$

Output: $\hat{\beta}^{\text{inv}}$, $\hat{\delta}^{\text{res}}_{t^*}, \hat{\gamma}_{t^*}, \hat{\gamma}^0_{t^*}$

$(\mathbf{X}_k)_{k\in[K]} \leftarrow ([X_{\frac{(k-1)n}{K}}, \dots, X_{\frac{kn}{K}}]^\top)_{k\in[K]}$

$(\mathbf{Y}_k)_{k\in[K]} \leftarrow ([Y_{\frac{(k-1)n}{K}}, \dots, Y_{\frac{kn}{K}}]^\top)_{k\in[K]}$

$(\hat{\Sigma}_k)_{k\in[K]} \leftarrow \{\widehat{\text{Var}}(\mathbf{X}_k)\}_{k\in[K]}$

$\left\{\begin{bmatrix}\hat{\gamma}^0_k \\ \hat{\gamma}_k\end{bmatrix}\right\}_{k\in[K]} \leftarrow \left\{\text{OLS}(\mathbf{Y}_k, \begin{bmatrix}\mathbf{1}_{\frac{n}{K}} & \mathbf{X}_k\end{bmatrix})\right\}_{k\in[K]}$

$\begin{bmatrix}\hat{\gamma}^0 \\ \hat{\gamma}\end{bmatrix} \leftarrow \frac{1}{K}\sum_{k=1}^K \begin{bmatrix}\hat{\gamma}^0_k \\ \hat{\gamma}_k\end{bmatrix}$

$\hat{U}, \{\hat{S}_j\}_{j=1}^q \leftarrow \text{approxIrreducibleJointBlockDiagonalizer}((\hat{\Sigma}_t)_{t\in[n]})$ ▷ see Sec. 2.4.1.1

▷ Find the opt-invariant subspaces to estimate $\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}$:

$\hat{S}^{\text{inv}}, \hat{S}^{\text{res}} \leftarrow \emptyset$

for $j = 1, \dots, q$ **do**

  $\Pi_{\hat{\mathcal{S}}_j} \leftarrow \hat{U}^{S_j}(\hat{U}^{S_j})^\top$

  $(\hat{c}^j_k)_{k\in[K]} \leftarrow (\widehat{\text{Corr}}(\mathbf{Y}_k - \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma}), \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma})))_{k\in[K]}$

  if $\frac{1}{K}\sum_{k=1}^K \left|\hat{c}^j_k\right| \le \lambda$ **then** $\hat{S}^{\text{inv}} \leftarrow \hat{S}^{\text{inv}} \cup \hat{S}_j$ ▷ see Eq. (2.4.19)

  else $\hat{S}^{\text{res}} \leftarrow \hat{S}^{\text{res}} \cup \hat{S}_j$

▷ Invariant component estimation:

$\mathbf{X} \leftarrow [X_1 \dots X_n]^\top$, $\mathbf{Y} \leftarrow [Y_1 \dots Y_n]^\top$

$\hat{\beta}^{\text{inv}} \leftarrow \hat{U}^{\text{inv}}\widehat{\text{Var}}(\mathbf{X}\hat{U}^{\text{inv}})^{-1}\widehat{\text{Cov}}(\mathbf{X}\hat{U}^{\text{inv}}, \mathbf{Y})$ ▷ see Eq. (2.4.21)

▷ Adaptation step:

$\mathbf{X}^{\text{ad}} \leftarrow [(X_t)_{t\in\mathcal{I}^{\text{ad}}}]^\top$, $\mathbf{Y}^{\text{ad}} \leftarrow [(Y_t)_{t\in\mathcal{I}^{\text{ad}}}]^\top$

$\hat{\delta}^{\text{res}}_{t^*} \leftarrow \hat{U}^{\text{res}}\widehat{\text{Var}}(\mathbf{X}^{\text{ad}}\hat{U}^{\text{res}})^{-1}\widehat{\text{Cov}}(\mathbf{X}^{\text{ad}}\hat{U}^{\text{res}}, \mathbf{Y}^{\text{ad}} - \mathbf{X}^{\text{ad}}\hat{\beta}^{\text{inv}})$ ▷ see Eq. (2.4.22)

▷ Intercept estimation:

if $\{\hat{\gamma}^0_k\}_{k\in[K]}$ approximately constant **then** $\hat{\gamma}^0_{t^*} \leftarrow \hat{\gamma}^0$

else $\hat{\gamma}^0_{t^*} \leftarrow \hat{\mathbb{E}}[\mathbf{Y}^{\text{ad}} - \mathbf{X}^{\text{ad}}\hat{\beta}^{\text{inv}}] - \hat{\mathbb{E}}[\mathbf{X}^{\text{ad}}]\hat{\delta}^{\text{res}}_{t^*}$

$\hat{\gamma}_{t^*} \leftarrow \hat{\beta}^{\text{inv}} + \hat{\delta}^{\text{res}}_{t^*}$ ▷ see Eq. (2.4.23)

estimation of the intercept, the algorithm distinguishes between two cases: (a) the intercept remains approximately constant in $[n] \cup \mathcal{I}_{\text{ad}}$ and (b) the intercept changes with time but is assumed to be approximately constant in $\mathcal{I}_{\text{ad}}$. In case (a), the computation of $\hat{\gamma}^0_{t^*}$ in line 21 of Algorithm 2 is done by averaging the estimated values of the (approximately constant) intercept on historical data. Alternatively, in case (a) we could estimate the intercept simultaneously to the invariant component β^{inv}: the whole vector (including the intercept) can be computed by taking the OLS solution of regressing $\mathbf{Y}$ on $\begin{bmatrix}\mathbf{1}_n & \mathbf{X}\hat{U}^{\text{inv}}\end{bmatrix}$ and premultiplying it by $\begin{bmatrix}\mathbf{e}_1 & \hat{U}^{\text{inv}}\end{bmatrix}$ (where $\mathbf{e}_1$ is the p-dimensional vector with the first entry equal to 1 and the remaining equal to zero) in place of lines 16 and 21 of Algorithm 2 (see Remark 2.6.1). In case (b), the intercept can instead be estimated simultaneously to the residual component $\hat{\delta}^{\text{res}}_t$: the computation in line 22 of the algorithm is equivalent to taking the first component of the OLS solution of regressing $\mathbf{Y}^{\text{ad}} - \mathbf{X}^{\text{ad}}\hat{\beta}^{\text{inv}}$ on $\begin{bmatrix}\mathbf{1}_m & \mathbf{X}^{\text{ad}}\hat{U}^{\text{res}}\end{bmatrix}$, premultiplied by $\begin{bmatrix}\mathbf{e}_1 & \hat{U}^{\text{res}}\end{bmatrix}$.

Remark 2.6.1. *Recall that the population OLS solution for the linear model (2.B.1) at time t can be found by adding a 1 to the vector X_t and solving*

$$\arg\min_{\gamma^0,\gamma} \mathbb{E}\left[Y_t - [1\ X_t^\top]\begin{bmatrix}\gamma^0\\ \gamma\end{bmatrix}\right]^2.$$

The problem has closed form solution

$$\begin{bmatrix}\gamma^0\\ \gamma\end{bmatrix} = \begin{bmatrix}1 & \mathbb{E}[X_t]^\top\\ \mathbb{E}[X_t] & \mathbb{E}[X_tX_t^\top]\end{bmatrix}^{-1}\begin{bmatrix}\mathbb{E}[Y_t]\\ \mathbb{E}[X_tY_t]\end{bmatrix} = \begin{bmatrix}\mathbb{E}[Y_t] - \mathbb{E}[X_t]^\top \operatorname{Var}(X_t)^{-1}\operatorname{Cov}(X_t,Y_t)\\ \operatorname{Var}(X_t)^{-1}\operatorname{Cov}(X_t,Y_t)\end{bmatrix}.$$

2.C. Extension to non-orthogonal subspaces

In Section 2.2.1 we have defined an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition as a collection $\{\mathcal{S}_j\}_{j\in\{1,\dots,q\}}$ of pairwise orthogonal linear subspaces of $\mathbb{R}^p$ satisfying (2.2.6). Finding the orthogonal subspaces that form the partition means in particular finding a rotation of the original X-space such that each subspace is spanned by a subset of the rotated axes, and the coordinates of the projected predictors onto one subspace are uncorrelated with the ones in the remaining subspaces. In this section, we show that orthogonality of the subspaces in the partition is not strictly required to obtain a separation of the true time-varying parameter of the form (2.2.7), that is, more general invertible linear transformations can be considered besides rotations. In particular, we briefly present results similar to the ones obtained throughout Section 2.2 but where the subspaces in the partition are not necessarily orthogonal. To do so, we define a collection of (not necessarily orthogonal) linear subspaces $\mathcal{S}_1,\dots,\mathcal{S}_q \subseteq \mathbb{R}^p$ with $\bigoplus_{j=1}^q \mathcal{S}_j = \mathbb{R}^p$ and satisfying (2.2.6) a *$(X_t)_{t\in[n]}$-decorrelating partition (of cardinality q)*, and further call it *irreducible* if it is of maximal cardinality. A $(X_t)_{t\in[n]}$-decorrelating partition can still be identified through a joint block diagonalization of the covariance matrices $(\Sigma_t)_{t\in[n]}$ as described in Section 2.2.1.1 but with an adjustment. More specifically, instead of assuming that the joint diagonalizer U is an orthogonal matrix, we only assume it is

invertible and for all $j \in \{1, \dots, q\}$, the columns of U^{S_j} are orthonormal vectors. A version of Proposition 2.2.8 in which the resulting partition is not necessarily orthogonal follows with the same proof. Moreover, similarly to the orthogonal case, the uniqueness of an irreducible $(X_t)_{t\in[n]}$-decorrelating partition is implied by the uniqueness of an irreducible non-orthogonal joint block diagonalizer for $(\Sigma_t)_{t\in[n]}$; explicit conditions under which such uniqueness holds can be found for example in the work by Nion [2011]. In the results presented in the remainder of this section we adopt the same notation introduced in Section 2.2.1.1, and we additionally define the matrix $W \coloneqq U^{-\top}$.

A $(X_t)_{t\in[n]}$-decorrelating partition of cardinality q allows us to decompose the true time-varying parameter into the sum of q components. To obtain such a decomposition via non-orthogonal subspaces, oblique projections need to be considered in place of orthogonal ones. Oblique projections are defined [see, e.g., Schott, 2016] for two subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^p$ such that $\mathcal{S}_1 \bigoplus \mathcal{S}_2 = \mathbb{R}^p$ and a vector $x \in \mathbb{R}^p$ as the vectors $x_1 \in \mathcal{S}_1$ and $x_2 \in \mathcal{S}_2$ such that $x = x_1 + x_2$: x_1 is called the projection of x onto $\mathcal{S}_1$ along $\mathcal{S}_2$, and x_2 the projection of x onto $\mathcal{S}_2$ along $\mathcal{S}_1$. For a $(X_t)_{t\in[n]}$-decorrelating partition $\{\mathcal{S}_j\}_{j=1}^q$, we denote by $P_{\mathcal{S}_j|\mathcal{S}_{-j}}$ the oblique projection matrix onto $\mathcal{S}_j$ along $\bigoplus_{i\in\{1,\dots,q\}\setminus\{j\}} \mathcal{S}_i$: this can be expressed in terms of a (non-orthogonal) joint block diagonalizer U corresponding to the partition as $P_{\mathcal{S}_j|\mathcal{S}_{-j}} = U^{S_j}(W^{S_j})^\top$. Orthogonal and $(X_t)_{t\in[n]}$-decorrelating partitions $\{\mathcal{S}_j\}_{j=1}^q$ are a special case of $(X_t)_{t\in[n]}$-decorrelating partitions. In particular, if the subspaces are pairwise orthogonal, it holds for all $j \in \{1, \dots, q\}$ that $P_{\mathcal{S}_j|\mathcal{S}^{-j}} = \Pi_{\mathcal{S}_j}$.

By definition of oblique projections, for all $t \in [n]$, we can express $\gamma_{0,t}$ as

$$\gamma_{0,t} = \sum_{j=1}^{q} P_{\mathcal{S}_j|\mathcal{S}_{-j}} \gamma_{0,t}.$$

Similarly to how we define opt-invariance on $[n]$ for orthogonal subspaces in Section 2.2.1, we say that a subspace $\mathcal{S}_j$ in a $(X_t)_{t\in[n]}$-decorrelating partition is *proj-invariant* on $[n]$ if it satisfies for all $t, s \in [n]$ that

$$P_{\mathcal{S}_j|\mathcal{S}_{-j}} \gamma_{0,t} = P_{\mathcal{S}_j|\mathcal{S}_{-j}} \gamma_{0,s}.$$

By Lemma 2.2.4 it follows that for orthogonal partitions proj-invariance is equivalent to opt-invariance. For an irreducible $(X_t)_{t\in[n]}$-decorrelating partition, we now define the invariant subspace $\mathcal{S}^{\text{inv}}$ and residual subspace $\mathcal{S}^{\text{res}}$ as

$$\mathcal{S}^{\text{inv}} \coloneqq \bigoplus_{\substack{j\in\{1,\dots,q\}: \\ \mathcal{S}_j \text{ proj-invariant on } [n]}} \mathcal{S}_j \quad \text{and} \quad \mathcal{S}^{\text{res}} \coloneqq \bigoplus_{\substack{j\in\{1,\dots,q\}: \\ \mathcal{S}_j \text{ not proj-invariant on } [n]}} \mathcal{S}_j.$$

It follows directly by the definition of partitions that $\{\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}\}$ is a $(X_t)_{t\in[n]}$-decorrelating partition. Moreover, $\mathcal{S}^{\text{inv}}$ is proj-invariant on $[n]$ since $P_{\mathcal{S}^{\text{inv}}|\mathcal{S}^{\text{res}}}\gamma_{0,t} = \sum_{\substack{j\in\{1,\dots,q\}: \\ \mathcal{S}_j \text{ proj-invariant on } [n]}} P_{\mathcal{S}_j|\mathcal{S}_{-j}}\gamma_{0,t}$. We finally define the invariant and residual components by

$$\beta^{\text{inv}} \coloneqq P_{\mathcal{S}^{\text{inv}}|\mathcal{S}^{\text{res}}}\bar{\gamma}_0 \quad \text{and} \quad \delta_t^{\text{res}} \coloneqq P_{\mathcal{S}^{\text{res}}|\mathcal{S}^{\text{inv}}}\gamma_{0,t}.$$

We show in the following proposition that the expressions (2.2.14) and (2.2.16) used to construct our estimators for the invariant and residual component in the case of orthogonal partitions, remain valid in the case of non-orthogonal partitions.

Proposition 2.6.1. *Let $\{\mathcal{S}_j\}_{j=1}^q$ be a $(X_t)_{t\in[n]}$-decorrelating partition and let U be a joint block diagonalizer corresponding to that partition. Then, the following results hold.*

(i) For all $t \in [n]$ and for all $j \in \{1, \dots, q\}$

$$P_{\mathcal{S}_j|\mathcal{S}_{-j}}\gamma_{0,t} = U^{S_j}\operatorname{Var}((U^{S_j})^\top X_t)^{-1}\operatorname{Cov}((U^{S_j})^\top X_t, Y_t). \tag{2.C.1}$$

(ii) β^{inv} is time-invariant over $[n]$ and

$$\beta^{\text{inv}} = U^{\text{inv}}((U^{\text{inv}})^\top \overline{\operatorname{Var}}(X)U^{\text{inv}})^{-1}(U^{\text{inv}})^\top \overline{\operatorname{Cov}}(X, Y).$$

(iii)

$$\delta_t^{\text{res}} = U^{\text{res}}((U^{\text{res}})^\top \operatorname{Var}(X_t)U^{\text{res}})^{-1}(U^{\text{res}})^\top \operatorname{Cov}(X_t, Y_t - X_t^\top \beta^{\text{inv}}).$$

Proposition 2.6.1 implies in particular that, apart from the joint block diagonalization differences, estimating β^{inv} and δ_t^{res} in the non-orthogonal case can be done as described in Section 2.4. Moreover, it holds that the parameter β^{inv} defined for non-orthogonal $(X_t)_{t\in[n]}$-decorrelating partitions is still a time-invariant parameter. Under a generalization assumption analogous to Assumption 2, β^{inv} has positive explained variance at all time points $t \in \mathbb{N}$ and can be used to at least partially predict $\gamma_{0,t}$ (we have not added this result explicitly). In addition, also in the non-orthogonal case the estimation of the residual component only requires to estimate a reduced number of parameters, that is, $\dim(\mathcal{S}^{\text{res}})$.

In the case of non-orthogonal partitions, however, we cannot directly interpret ISD as separating the true time-varying parameter into two separate optimizations of the explained variance over $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$.

Example 2.6.2 (non-orthogonal irreducible partition)**.** *Let $X_t \in \mathbb{R}^3$ with covariance matrix Σ_t that for all $t \in [n]$ takes one of the following values*

$$\begin{bmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 4 & -2 & 0 \\ -2 & 7/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

These matrices are in block diagonal form and in particular the 2×2 submatrices forming the first diagonal block do not commute. This implies that an irreducible (orthogonal) joint block diagonalizer for these matrices is the three-dimensional identity matrix I_3, and the diagonal blocks cannot be further reduced into smaller blocks using an orthogonal transformation. It also implies that an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition is given by $\{\mathcal{S}_j\}_{j=1}^2$ with $\mathcal{S}_1 := \langle e_1, e_2\rangle$ and $\mathcal{S}_2 = \langle e_3\rangle$, where e_j is the j-th vector of the canonical basis for $\mathbb{R}^3$.

There exists, however, a non-orthogonal joint diagonalizer for these matrices, i.e., a non-orthogonal matrix U such that $\tilde{\Sigma}_t = U^\top \Sigma_t U$ is diagonal. It is given by

$$U = \begin{bmatrix} 1/\sqrt{5} & 1 & 0 \\ 2/\sqrt{5} & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

U induces an irreducible $(X_t)_{t\in[n]}$-decorrelating partition by

$$\bar{\mathcal{S}}_1 \coloneqq \langle \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{bmatrix} \rangle = \langle U^{S_1} \rangle, \quad \bar{\mathcal{S}}_2 \coloneqq \langle \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \rangle = \langle U^{S_2} \rangle, \quad \bar{\mathcal{S}}_3 \coloneqq \langle \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \rangle = \langle U^{S_3} \rangle.$$

Let $\mathcal{S}^{\mathrm{inv}}$ and $\bar{\mathcal{S}}^{\mathrm{inv}}$ be the invariant subspaces associated with the irreducible orthogonal partition $\{\mathcal{S}_j\}_{j=1}^3$ and the irreducible partition $\{\bar{\mathcal{S}}_j\}_{j=1}^3$, respectively. As any irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition is also a $(X_t)_{t\in[n]}$-decorrelating partition, it in general holds that

$$\dim(\mathcal{S}^{\mathrm{inv}}) \leq \dim(\bar{\mathcal{S}}^{\mathrm{inv}}).$$

Moreover, as in the explicit example above, the inequality can be strict.

2.D. Auxiliary results

Lemma 2.6.1. *Let $B \in \mathbb{R}^{m\times m}$ be a symmetric invertible matrix, $U \in \mathbb{R}^{m\times m}$ an orthogonal block diagonalizer of B and $S_1, \ldots, S_q \subseteq \{1, \ldots, m\}$ disjoint subsets satisfying*

$$U^\top B U = \begin{bmatrix} (U^{S_1})^\top B U^{S_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & (U^{S_q})^\top B U^{S_q} \end{bmatrix}.$$

Then it holds for all $j \in \{1, \ldots, q\}$ that

$$(\Pi_{\mathcal{S}_j} B \Pi_{\mathcal{S}_j})^\dagger = \Pi_{\mathcal{S}_j} B^{-1} \Pi_{\mathcal{S}_j},$$

where $\Pi_{\mathcal{S}_j} \coloneqq U^{S_j}(U^{S_j})^\top$.

Proof. The pseudo-inverse $A^\dagger$ of a matrix A is defined as the unique matrix satisfying: (i) $AA^\dagger A = A$, (ii) $A^\dagger A A^\dagger = A^\dagger$, (iii) $(AA^\dagger)^\top = AA^\dagger$ and (iv) $(A^\dagger A)^\top = A^\dagger A$.

Fix $j \in \{1 \ldots, q\}$ and define $A \coloneqq \Pi_{\mathcal{S}_j} B \Pi_{\mathcal{S}_j}$ and $A^\dagger \coloneqq \Pi_{\mathcal{S}_j} B^{-1} \Pi_{\mathcal{S}_j}$. Moreover, define $\widetilde{B} \coloneqq U^\top B U$ and for all $k \in \{1, \ldots, q\}$, $\widetilde{B}_k \coloneqq (U^{S_k})^\top B U^{S_k}$. We now verify that conditions (i)-(iv) hold for A and $A^\dagger$ and hence $A^\dagger$ is indeed the pseudo-inverse. Conditions (ii) and (iv) hold by symmetry of B and Π_S. For (i) and (iii), first observe that by the properties of orthogonal matrices it holds that $\widetilde{B}^{-1} = U^\top B^{-1} U$, and, due to the block

diagonal structure of $\widetilde{B}$,

$$\widetilde{B}^{-1} = \begin{bmatrix} \widetilde{B}_1^{-1} & & \\ & \ddots & \\ & & \widetilde{B}_q^{-1} \end{bmatrix} = \begin{bmatrix} (U^{S_1})^\top \\ \vdots \\ (U^{S_q})^\top \end{bmatrix} B^{-1} \begin{bmatrix} U^{S_1} & \dots & U^{S_q} \end{bmatrix}. \tag{2.D.1}$$

Hence we get that $\widetilde{B}_j^{-1} = (U^{S_j})^\top B^{-1} U^{S_j}$. For (i), we now get

$$\begin{aligned} \Pi_{\mathcal{S}_j} B \Pi_{\mathcal{S}_j} \Pi_{\mathcal{S}_j} B^{-1} \Pi_{\mathcal{S}_j} \Pi_{\mathcal{S}_j} B \Pi_{\mathcal{S}_j} &= U^{S_j} (U^{S_j})^\top B U^{S_j} (U^{S_j})^\top B^{-1} U^{S_j} (U^{S_j})^\top B U^{S_j} (U^{S_j})^\top \\ &= U^{S_j} \widetilde{B}_j \widetilde{B}_j^{-1} \widetilde{B}_j (U^{S_j})^\top \\ &= \Pi_{\mathcal{S}_j} B \Pi_{\mathcal{S}_j}. \end{aligned}$$

Similarly, for (iii) we get

$$\begin{aligned} (\Pi_{\mathcal{S}_j} B \Pi_{\mathcal{S}_j} \Pi_{\mathcal{S}_j} B^{-1} \Pi_{\mathcal{S}_j})^\top &= (U^{S_j} (U^{S_j})^\top B U^{S_j} (U^{S_j})^\top B^{-1} U^{S_j} (U^{S_j})^\top)^\top \\ &= (U^{S_j} \widetilde{B}_j \widetilde{B}_j^{-1} (U^{S_j})^\top)^\top \\ &= (\Pi_{\mathcal{S}_j})^\top \\ &= \Pi_{\mathcal{S}_j}. \end{aligned}$$

This completes the proof of Lemma 2.6.1.

□

Lemma 2.6.2. *Let $\mathcal{N} \subseteq \mathbb{N}$ and let $\{\mathcal{S}_j\}_{j=1}^q$ be an orthogonal and $(X_t)_{t\in\mathcal{N}}$-decorrelating partition. Then, there exists a joint block diagonalizer of $(\Sigma_t)_{t\in\mathcal{N}}$. More precisely, there exists an orthonormal matrix $U \in \mathbb{R}^{p\times p}$ such that for all $t \in \mathcal{N}$ the matrix $\widetilde{\Sigma}_t \coloneqq U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\widetilde{\Sigma}_{t,j} = (U^{S_j})^\top \Sigma_t U^{S_j}$, $j \in \{1,\dots,q\}$ and of dimension $|S_j| = \dim(\mathcal{S}_j)$, where $S_j \subseteq \{1,\dots,p\}$ indexes a subset of the columns of U. Moreover, $\Pi_{\mathcal{S}_j} = U^{S_j}(U^{S_j})^\top$.*

Proof. Let $U = (u_1,\dots,u_p) \in \mathbb{R}^{p\times p}$ be an orthonormal matrix with columns $u_1,\dots,u_p$ such that for all $j \in \{1,\dots,q\}$ there exists $S_j \subseteq \{1,\dots,p\}$ such that $\mathcal{S}_j = \operatorname{span}(\{u_k \mid k \in S_j\})$. Such a matrix U can be constructed by selecting an orthonormal basis for each of the disjoint subspaces $\{\mathcal{S}_j\}_{j=1}^q$. Furthermore, assume that the columns of U are ordered in such a way that for all $i,j \in \{1,\dots,q\}$ with $i < j$ it holds for all $k \in S_i$ and $\ell \in S_j$ that $k < \ell$. As the matrix U is orthogonal, it holds for all $j \in \{1,\dots,q\}$ that the projection matrix $\Pi_{\mathcal{S}_j}$ can be expressed as $\Pi_{\mathcal{S}_j} = U^{S_j}(U^{S_j})^\top$ and hence using the definition of orthogonal partition (see (2.2.6)) it holds that, for all $t \in \mathcal{N}$,

$$\widetilde{\Sigma}_t \coloneqq U^\top \Sigma_t U = \begin{bmatrix} \widetilde{\Sigma}_{t,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \widetilde{\Sigma}_{t,q} \end{bmatrix},$$

where for all $j \in \{1, \dots, q\}$ we defined $\widetilde{\Sigma}_{t,j} \coloneqq (U^{S_j})^\top \Sigma_t U^{S_j}$. □

Lemma 2.6.3. *Let $\{\mathcal{S}_j\}_{j=1}^q$ be an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Then it holds for all $t \in [n]$ and, for all $j \in \{1, \dots, q\}$, that*

$$\operatorname{Var}\left(\Pi_{\mathcal{S}_j} X_t\right)^\dagger = \Pi_{\mathcal{S}_j} \operatorname{Var}(X_t)^{-1} \Pi_{\mathcal{S}_j}. \tag{2.D.2}$$

Proof. Let $U \in \mathbb{R}^{p\times p}$ be an orthonormal matrix such that, for all $t \in [n]$, $\tilde{\Sigma}_t \coloneqq U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\widetilde{\Sigma}_{t,1}, \dots, \widetilde{\Sigma}_{t,q}$ of dimensions $\dim(\mathcal{S}_1), \dots, \dim(\mathcal{S}_q)$. Such a matrix exists by Lemma 2.6.2 and each diagonal block is given by $\widetilde{\Sigma}_{t,j} \coloneqq (U^{S_j})^\top \Sigma_t U^{S_j}$ and the projection matrix $\Pi_{\mathcal{S}_j}$ can be expressed as $\Pi_{\mathcal{S}_j} = U^{S_j}(U^{S_j})^\top$. The statement then follows from Lemma 2.6.1. □

Proof of Lemma 2.2.4

Proof. Let $U \in \mathbb{R}^{p\times p}$ be an orthonormal matrix such that, for all $t \in [n]$, $\tilde{\Sigma}_t \coloneqq U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\widetilde{\Sigma}_{t,1}, \dots, \widetilde{\Sigma}_{t,q}$ of dimensions $\dim(\mathcal{S}_1), \dots, \dim(\mathcal{S}_q)$. Such a matrix exists by Lemma 2.6.2 and each diagonal block is given by $\widetilde{\Sigma}_{t,j} \coloneqq (U^{S_j})^\top \Sigma_t U^{S_j}$ and the projection matrix $\Pi_{\mathcal{S}_j}$ can be expressed as $\Pi_{\mathcal{S}_j} = U^{S_j}(U^{S_j})^\top$.

Now, for an arbitrary $t \in [n]$ it holds using the linear model (2.1.1) that $\gamma_{0,t} = \Sigma_t^{-1} \operatorname{Cov}(X_t, Y_t)$ and hence we use U to get the following expansion

$$\begin{aligned}
\gamma_{0,t} &= \Sigma_t^{-1} \operatorname{Cov}(X_t, Y_t) \\
&= U\tilde{\Sigma}_t^{-1} U^\top \operatorname{Cov}(X_t, Y_t) \\
&= U \begin{bmatrix} \tilde{\Sigma}_{t,1}^{-1} & & \\ & \ddots & \\ & & \tilde{\Sigma}_{t,q}^{-1} \end{bmatrix} U^\top \operatorname{Cov}(X_t, Y_t) \\
&= U \begin{bmatrix} \tilde{\Sigma}_{t,1}^{-1} (U^{S_1})^\top \operatorname{Cov}(X_t, Y_t) \\ \vdots \\ \tilde{\Sigma}_{t,q}^{-1} (U^{S_q})^\top \operatorname{Cov}(X_t, Y_t) \end{bmatrix}.
\end{aligned}$$

By the properties of orthogonal matrices it holds that $\tilde{\Sigma}_t^{-1} = U^\top \Sigma_t^{-1} U$, and, due to the block diagonal structure of $\tilde{\Sigma}_t$,

$$\tilde{\Sigma}_t^{-1} = \begin{bmatrix} \tilde{\Sigma}_{t,1}^{-1} & & \\ & \ddots & \\ & & \tilde{\Sigma}_{t,q}^{-1} \end{bmatrix} = \begin{bmatrix} (U^{S_1})^\top \\ \vdots \\ (U^{S_q})^\top \end{bmatrix} \Sigma_t^{-1} \begin{bmatrix} U^{S_1} & \dots & U^{S_q} \end{bmatrix}. \tag{2.D.3}$$

This implies that $\tilde{\Sigma}_{t,j}^{-1} = (U^{S_j})^\top \Sigma_t^{-1} U^{S_j}$ and therefore

$$\begin{aligned}
\gamma_{0,t} &= U \begin{bmatrix} (U^{S_1})^\top \Sigma_t^{-1} U^{S_1} (U^{S_1})^\top \operatorname{Cov}(X_t, Y_t) \\ \vdots \\ (U^{S_q})^\top \Sigma_t^{-1} U^{S_q} (U^{S_q})^\top \operatorname{Cov}(X_t, Y_t) \end{bmatrix} \\
&= \begin{bmatrix} U^{S_1} & \cdots & U^{S_q} \end{bmatrix} \begin{bmatrix} (U^{S_1})^\top \Sigma_t^{-1} U^{S_1} (U^{S_1})^\top \operatorname{Cov}(X_t, Y_t) \\ \vdots \\ (U^{S_q})^\top \Sigma_t^{-1} U^{S_q} (U^{S_q})^\top \operatorname{Cov}(X_t, Y_t) \end{bmatrix} \\
&= \sum_{j=1}^{q} U^{S_j} (U^{S_j})^\top \Sigma_t^{-1} U^{S_j} (U^{S_j})^\top \operatorname{Cov}(X_t, Y_t) \\
&= \sum_{j=1}^{q} \Pi_{\mathcal{S}_j} \Sigma_t^{-1} \Pi_{\mathcal{S}_j} \operatorname{Cov}(X_t, Y_t) \\
&= \sum_{j=1}^{q} \Pi_{\mathcal{S}_j} \Sigma_t^{-1} \Pi_{\mathcal{S}_j} \Pi_{\mathcal{S}_j} \operatorname{Cov}(X_t, Y_t) \\
&= \sum_{j=1}^{q} \operatorname{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \operatorname{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t).
\end{aligned}$$

In the second to last equality, we have used that $\Pi_{\mathcal{S}_j}$ is a projection matrix and thus idempotent. In the last equality we have used that $\operatorname{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger = \Pi_{\mathcal{S}_j} \Sigma_t^{-1} \Pi_{\mathcal{S}_j}$, by Lemma 2.6.3. It now suffices to show that $\Pi_{\mathcal{S}_j} \gamma_{0,t} = \operatorname{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \operatorname{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t)$, which follows from the following computation

$$\begin{aligned}
\Pi_{\mathcal{S}_j} \gamma_{0,t} &= \Pi_{\mathcal{S}_j} \sum_{i=1}^{q} \Pi_{\mathcal{S}_i} \operatorname{Var}(X_t)^{-1} \Pi_{\mathcal{S}_i} \operatorname{Cov}(\Pi_{\mathcal{S}_i} X_t, Y_t) \\
&= \Pi_{\mathcal{S}_j} \operatorname{Var}(X_t)^{-1} \Pi_{\mathcal{S}_j} \operatorname{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) \\
&= \operatorname{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \operatorname{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t),
\end{aligned}$$

where the first equality follows from the first part of this proof and the third equality follows from Lemma 2.6.3. □

Proof of Lemma 2.2.5

Proof. For all $j \in \{1, \ldots, q\}$, for all $\beta \in \mathcal{S}_j$ and for all $t \in \mathcal{N}$ it holds that

$$\begin{aligned}
\Delta \operatorname{Var}_t(\beta) &= 2 \operatorname{Cov}(Y_t, X_t)\beta - \beta^\top \operatorname{Var}(X_t)\beta \\
&= 2 \operatorname{Cov}(Y_t, (\textstyle\sum_{k=1}^{q} \Pi_{\mathcal{S}_k} X_t))\Pi_{\mathcal{S}_j}\beta - \beta^\top \Pi_{\mathcal{S}_j} \operatorname{Var}(\textstyle\sum_{k=1}^{q} \Pi_{\mathcal{S}_k} X_t)\Pi_{\mathcal{S}_j}\beta \\
&= 2 \operatorname{Cov}(Y_t, (\Pi_{\mathcal{S}_j} X_t))\Pi_{\mathcal{S}_j}\beta - \beta^\top \Pi_{\mathcal{S}_j} \operatorname{Var}(\Pi_{\mathcal{S}_j} X_t)\Pi_{\mathcal{S}_j}\beta \\
&= 2 \operatorname{Cov}(Y_t, (\Pi_{\mathcal{S}_j} X_t))\beta - \beta^\top \operatorname{Var}(\Pi_{\mathcal{S}_j} X_t)\beta, \qquad (2.D.4)
\end{aligned}$$

where the first equality follows by (2.2.4) and the third equality follows from the definition of an orthogonal partition. It follows that

$$\nabla(\Delta\mathrm{Var}_t)(\beta) = 2\,\mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) - 2\,\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\beta,$$

where ∇ denotes the gradient. The equation $\nabla(\Delta\mathrm{Var}_t)(\beta) = 0$ has a unique solution in $\mathcal{S}_j$ given by $\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t)$. To see this, observe that all other solutions in $\mathbb{R}^p$ are given, for an arbitrary vector $w \in \mathbb{R}^p$, by

$$\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) + (I_p - \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t))w \tag{2.D.5}$$

where I_p denotes the identity matrix. Let $U \in \mathbb{R}^{p\times p}$ and $(\tilde{\Sigma}_t)_{t\in\mathcal{N}}$ be defined for the orthogonal and $(X_t)_{t\in\mathcal{N}}$-decorrelating partition as in Lemma 2.6.2. We now observe that

$$\begin{aligned}
\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t) &= \Pi_{\mathcal{S}_j}\Sigma_t^{-1}\Pi_{\mathcal{S}_j}\Pi_{\mathcal{S}_j}\Sigma_t\Pi_{\mathcal{S}_j} \\
&= U^{S_j}(U^{S_j})^\top\Sigma_t^{-1}U^{S_j}(U^{S_j})^\top\Sigma_t U^{S_j}(U^{S_j})^\top \\
&= U^{S_j}\tilde{\Sigma}_{t,j}^{-1}\tilde{\Sigma}_{t,j}(U^{S_j})^\top \\
&= \Pi_{\mathcal{S}_j}.
\end{aligned} \tag{2.D.6}$$

where the first equality follows from Lemma 2.6.1, since U jointly block diagonalizes the matrices $(\Sigma_t)_{t\in\mathcal{N}}$ by Lemma 2.6.2. We can therefore rewrite (2.D.5) as

$$\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) + (I_p - \Pi_{\mathcal{S}_j})w.$$

For all $w \in \mathcal{S}_j$, this expression equals $\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t)$. For all $w \notin \mathcal{S}_j$, it is not in $\mathcal{S}_j$. This concludes the proof of Lemma 2.2.5. □

Lemma 2.6.4. *$\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ is an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition.*

Proof. Let $\{\mathcal{S}_j\}_{j=1}^q$ be a fixed irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition according to which $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ are defined. By orthogonality of the subspaces in the partition and by definition of $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$, $\mathcal{S}^{\mathrm{res}} = (\mathcal{S}^{\mathrm{inv}})^\perp$. Moreover, by definition of orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition, it holds that

$$\mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t, \Pi_{\mathcal{S}^{\mathrm{res}}} X_t) = \mathrm{Cov}\Big(\sum_{\substack{j\in\{1,\dots,q\}:\\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \Pi_{\mathcal{S}_j} X_t,\ \sum_{\substack{j\in\{1,\dots,q\}:\\ \mathcal{S}_j \text{ not opt-invariant on } [n]}} \Pi_{\mathcal{S}_j} X_t\Big) = 0.$$

□

Lemma 2.6.5. *$\mathcal{S}^{\mathrm{inv}}$ is opt-invariant on $[n]$.*

Proof. That $\mathcal{S}^{\mathrm{inv}}$ is opt-invariant on $[n]$ can be seen from the following computations. It

holds for all $t, s \in [n]$ that

$$\begin{aligned}
\arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \Delta\text{Var}_t(\beta) &= \text{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_t)^\dagger \,\text{Cov}(\Pi_{\mathcal{S}^{\text{inv}}} X_t, Y_t) \\
&= \sum_{\substack{j \in \{1,\dots,q\}: \\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \text{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\text{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) \\
&= \sum_{\substack{j \in \{1,\dots,q\}: \\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \arg\max_{\beta \in \mathcal{S}_j} \Delta\text{Var}_t(\beta) \\
&= \sum_{\substack{j \in \{1,\dots,q\}: \\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \arg\max_{\beta \in \mathcal{S}_j} \Delta\text{Var}_s(\beta) \\
&= \arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \Delta\text{Var}_s(\beta).
\end{aligned}$$

The first equality holds by Lemma 2.2.5, since $\{\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}\}$ is indeed an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition, see Lemma 2.6.4. The second equality follows from the definition of $\mathcal{S}^{\text{inv}}$ and can be proved by Lemma 2.2.4 and observing that the set of subspaces $\{\mathcal{S}_j \mid j \in \{1,\dots,q\} : \mathcal{S}_j \text{ opt-invariant on } [n]\}$ is an orthogonal and $(\Pi_{\mathcal{S}^{\text{inv}}} X_t)_{t\in[n]}$-decorrelating partition. The third equality holds by Lemma 2.2.5 and the fourth by definition of an opt-invariant subspace on $[n]$. □

Lemma 2.6.6. *Let $\{A_t\}_{t=1}^n$ be a set of n symmetric strictly positive definite matrices. If there exists a matrix $A \in \{A_t\}_{t=1}^n$ that has all distinct eigenvalues, then any two irreducible joint block diagonalizers $U, \tilde{U}$ for the set $\{A_t\}_{t=1}^n$ are equal up to block permutations and block-wise isometric transformations.*

Proof. We start by observing that if a matrix is symmetric and has all distinct eigenvalues, then its eigenvectors are orthogonal to each other and unique up to scaling. We define $Q \in \mathbb{R}^{p\times p}$ as the orthonormal matrix whose columns are eigenvectors for A: Q is then uniquely defined up to permutations of its columns.

We now exploit the results by Murota et al. [2010] used in the construction of an irreducible orthogonal joint block diagonalizer for a set of symmetric matrices $\{A_t\}_{t=1}^n$ (not necessarily containing a matrix with all distinct eigenvalues). In the following, we translate all the useful results by Murota et al. [2010] in our notation introduced for joint block diagonalizers in Section 2.2.1.1. Murota et al. [2010] show in Theorem 1 that there exists an irreducible orthogonal joint block diagonalizer U such that, for all $t \in [n]$, $U^\top A_t U = \bigoplus_{j=1}^q (I_{\bar{m}_j} \otimes A_{t,j})$, where $q, \bar{m}_j \in \mathbb{N}$, $0 < q, \bar{m}_j \leq p$ are such that $q^U_{\max} = \sum_{j=1}^q \bar{m}_j$, and $A_{t,j}$ are square matrices (common diagonal blocks). Here $\oplus$ denotes the direct sum operator for matrices and $\otimes$ the Kronecker product. They further propose to partition the columns of the matrix U into q subsets, each denoted by U^{S_j}, with $j \in \{1,\dots,q\}$ indexing the diagonal blocks and $S_j \subseteq \{1,\dots,p\}$ denoting the subset of indexes corresponding to the selected columns in U^{S_j}, such that, for all $t \in [n]$, $(U^{S_j})^\top A_t U^{S_j} = I_{\bar{m}_j} \otimes A_{t,j}$. They then argue that, as a consequence of Theorem 1, the

spaces spanned by such subsets of columns, i.e.,

$$\mathcal{U}_j \coloneqq \operatorname{span}\{u^k \mid k \in S_j\}$$

are uniquely defined. We therefore only need to prove that, if at least one matrix in the set $\{A_t\}_{t=1}^n$ has all distinct eigenvalues, then for all $j \in \{1,\dots,q\}$, $\bar{m}_j = 1$. Such condition implies that the diagonal blocks indexed by S_j, $j \in \{1,\dots,q\}$, are irreducible, and $q_{\max}^U = q$.

The uniqueness of the spaces $\mathcal{U}_j$ then implies the uniqueness of the irreducible joint block diagonalizer U up to block permutations and block-wise isometric transformations. To show this result, we now use Murota et al. [2010, Proposition 1]. In particular, if the set $\{A_t\}_{t=1}^n$ contains at least one matrix A with all distinct eigenvalues then the assumptions of Proposition 1 are satisfied. Let $\{\mathcal{Q}_1,\dots,\mathcal{Q}_p\}$ be the set of eigenspaces of A and for all $i \in \{1,\dots,p\}$ let $m_i \coloneqq \dim(\mathcal{Q}_i) = 1$. The proposition then implies that for all $i \in \{1,\dots,p\}$ there exists $j \in \{1,\dots,q\}$ such that $\mathcal{Q}_i \subseteq \mathcal{U}_j$. Moreover, for all i such that $\mathcal{Q}_i \subseteq \mathcal{U}_j$ it holds that $m_i = \bar{m}_j$. As a consequence, we obtain that, since the eigenvalues of A are distinct, $m_1 = \cdots = m_p = 1$ and therefore for all $j \in \{1,\dots,q\}$, $\bar{m}_j = 1$. This concludes the proof for Lemma 2.6.6. □

Lemma 2.6.7. *Let $\{\mathcal{S}_j\}_{j=1}^q$ be an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Then, for all $j \in \{1,\dots,q\}$ it holds that*

$$\operatorname*{arg\,max}_{\beta\in\mathcal{S}_j} \overline{\Delta\mathrm{Var}}(\beta) = \left(\frac{1}{n}\sum_{t=1}^n \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\right)^\dagger \frac{1}{n}\sum_{t=1}^n \mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t).$$

In addition, for all $j \in \{1,\dots,q\}$ such that $\mathcal{S}_j$ is opt-invariant on $[n]$, it holds for all $t \in [n]$ that

$$\operatorname*{arg\,max}_{\beta\in\mathcal{S}_j} \overline{\Delta\mathrm{Var}}(\beta) = \Pi_{\mathcal{S}_j}\gamma_{0,t} = \Pi_{\mathcal{S}_j}\bar{\gamma}_0.$$

Moreover, $\Pi_{\mathcal{S}_j}\bar{\gamma}_0$ is time-invariant on $[n]$.

Proof. Let $U \in \mathbb{R}^{p\times p}$ and $(\tilde{\Sigma}_t)_{t\in[n]}$ be defined for the orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition as in Lemma 2.6.2. For all $j \in \{1,\dots,q\}$ and for all $\beta \in \mathcal{S}_j$ it follows from (2.D.4) that

$$\overline{\Delta\mathrm{Var}}(\beta) = \frac{2}{n}\sum_{t=1}^n \mathrm{Cov}(Y_t, \Pi_{\mathcal{S}_j} X_t)\beta - \beta^\top \frac{1}{n}\sum_{t=1}^n \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\beta$$

which has gradient

$$\nabla(\overline{\Delta\mathrm{Var}})(\beta) = \frac{2}{n}\sum_{t=1}^n \mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) - \frac{2}{n}\sum_{t=1}^n \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\beta.$$

The equation $\nabla(\overline{\Delta\mathrm{Var}})(\beta) = 0$ has solution equal to

$$\beta^* = \left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\right)^{\dagger} \frac{1}{n}\sum_{t=1}^{n} \mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t).$$

By Lemma 2.6.1, it holds that $\left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\right)^{\dagger} = \Pi_{\mathcal{S}_j}\left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(X_t)\right)^{-1}\Pi_{\mathcal{S}_j}$ and therefore $\beta^* \in \mathcal{S}_j$. In order to apply Lemma 2.6.1, we in particular use that U is such that, for all $t \in [n]$, $\tilde{\Sigma}_t = U^\top \Sigma_t U$ is block diagonal with diagonal blocks given, for all $j \in \{1, \ldots, q\}$, by $\tilde{\Sigma}_{t,j} = (U^{S_j})^\top \Sigma_t U^{S_j}$. This implies that $U^\top \left(\sum_{t=1}^{n} \Sigma_t\right) U$ is also block diagonal with diagonal blocks $(U^{S_j})^\top \left(\sum_{t=1}^{n} \Sigma_t\right) U^{S_j}$. Moreover, its inverse is block diagonal and, by the properties of orthogonal matrices, its diagonal blocks are $(U^{S_j})^\top \left(\sum_{t=1}^{n} \Sigma_t\right)^{-1} U^{S_j}$. It now remains to show that this is also the only solution in $\mathcal{S}_j$. All other solutions in $\mathbb{R}^p$ are given, for an arbitrary vector $w \in \mathbb{R}^p$, by

$$\begin{aligned}
&\beta^* + \left(I_p - \left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\right)^{\dagger} \left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\right)\right) w \\
=&\beta^* + \left(I_p - \Pi_{\mathcal{S}_j}\right) w.
\end{aligned}$$

The equality follows from the following computation

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\right)^{\dagger} \left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\right) \\
&= \Pi_{\mathcal{S}_j}\left(\sum_{t=1}^{n} \mathrm{Var}(X_t)\right)^{-1} \Pi_{\mathcal{S}_j}\Pi_{\mathcal{S}_j}\left(\sum_{t=1}^{n} \mathrm{Var}(X_t)\right)\Pi_{\mathcal{S}_j} \\
&= U^{S_j}(U^{S_j})^\top \left(\sum_{t=1}^{n} \Sigma_t\right)^{-1} U^{S_j}(U^{S_j})^\top \left(\sum_{t=1}^{n} \Sigma_t\right) U^{S_j}(U^{S_j})^\top \\
&= U^{S_j}\left(\sum_{t=1}^{n} \tilde{\Sigma}_t\right)^{-1} \left(\sum_{t=1}^{n} \tilde{\Sigma}_t\right)(U^{S_j})^\top \\
&= \Pi_{\mathcal{S}_j}.
\end{aligned}$$

The first equality follows again from Lemma 2.6.1. For all $w \in \mathcal{S}_j$, $\beta^* + (I_p - \Pi_{\mathcal{S}_j})w$ equals β^*. For all $w \notin \mathcal{S}_j$, $\beta^* + (I_p - \Pi_{\mathcal{S}_j})w$ is not in $\mathcal{S}_j$.

For all $\mathcal{S}_j$ opt-invariant on $[n]$ and for all $t \in [n]$ it holds that

$$\begin{aligned}
&\arg\max_{\beta \in \mathcal{S}_j} \overline{\Delta\mathrm{Var}}(\beta) \\
&= \arg\max_{\beta \in \mathcal{S}_j} \frac{1}{n}\sum_{s=1}^{n} \Delta\mathrm{Var}_s(\beta)
\end{aligned}$$

$$
\begin{aligned}
&= \arg\max_{\beta \in \mathcal{S}_j} \Delta\mathrm{Var}_t(\beta) \\
&= \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \,\mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) \\
&= \Pi_{\mathcal{S}_j} \gamma_{0,t},
\end{aligned}
$$

where we used the definition of opt-invariance on $[n]$ for the second equality, Lemma 2.2.5 for the third equality and Lemma 2.2.4 for the fourth equality. Since the result holds for all $t \in [n]$, it also follows that

$$\Pi_{\mathcal{S}_j} \gamma_{0,t} = \Pi_{\mathcal{S}_j} \bar{\gamma}_0.$$

We now need to prove for all $t \in [n]$ that $\mathrm{Cov}(Y_t - X_t^\top \Pi_{\mathcal{S}_j} \bar{\gamma}_0, X_t^\top \Pi_{\mathcal{S}_j} \bar{\gamma}_0) = 0$. To see this, fix $t \in [n]$. Then

$$
\begin{aligned}
\mathrm{Cov}(Y_t - X_t^\top \Pi_{\mathcal{S}_j} \bar{\gamma}_0, X_t^\top \Pi_{\mathcal{S}_j} \bar{\gamma}_0) &= \mathrm{Cov}(X_t^\top (\gamma_{0,t} - \Pi_{\mathcal{S}_j} \gamma_{0,t}), X_t^\top \Pi_{\mathcal{S}_j} \gamma_{0,t}) \\
&= \mathrm{Cov}(X_t^\top (\sum_{i=1}^{q} \Pi_{\mathcal{S}_i} \gamma_{0,t} - \Pi_{\mathcal{S}_j} \gamma_{0,t}), X_t^\top \Pi_{\mathcal{S}_j} \gamma_{0,t}) \\
&= \mathrm{Cov}(X_t^\top \sum_{i \in \{1,\dots,q\}: i \neq j} \Pi_{\mathcal{S}_i} \gamma_{0,t}, X_t^\top \Pi_{\mathcal{S}_j} \gamma_{0,t}) \\
&= \sum_{i \in \{1,\dots,q\}: i \neq j} \gamma_{0,t}^\top \,\mathrm{Cov}(\Pi_{\mathcal{S}_i} X_t, \Pi_{\mathcal{S}_j} X_t) \gamma_{0,t} \\
&= 0.
\end{aligned}
$$

The last equality follows from the definition of an orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition.

□

Lemma 2.6.8. *Let $\{\mathcal{S}_j\}_{j=1}^q$ be an irreducible orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition and let $\mathcal{S}^{\mathrm{inv}}$ be the corresponding invariant subspace. Moreover, let $\overline{\mathrm{Var}}(X) := \frac{1}{n} \sum_{t=1}^n \mathrm{Var}(X_t)$ and $\overline{\mathrm{Cov}}(X, Y) := \frac{1}{n} \sum_{t=1}^n \mathrm{Cov}(X_t, Y_t)$. Finally, let U^{inv} be the submatrix of an arbitrary irreducible joint block diagonalizer U corresponding to the irreducible orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition whose columns span $\mathcal{S}^{\mathrm{inv}}$. Then,*

$$\beta^{\mathrm{inv}} = U^{\mathrm{inv}} ((U^{\mathrm{inv}})^\top \overline{\mathrm{Var}}(X) U^{\mathrm{inv}})^{-1} (U^{\mathrm{inv}})^\top \overline{\mathrm{Cov}}(X, Y).$$

Proof. Expanding the definition of β^{inv}, we obtain that

$$
\begin{aligned}
\beta^{\mathrm{inv}} &= \arg\max_{\beta \in \mathcal{S}^{\mathrm{inv}}} \overline{\Delta\mathrm{Var}}(\beta) \\
&= \arg\max_{\beta \in \mathcal{S}^{\mathrm{inv}}} \frac{1}{n} \sum_{t=1}^{n} \Delta\mathrm{Var}_t(\beta) \\
&= \arg\max_{\beta \in \mathcal{S}^{\mathrm{inv}}} \frac{1}{n} \sum_{s=1}^{n} (2\,\mathrm{Cov}(Y_t, \Pi_{\mathcal{S}^{\mathrm{inv}}} X_t)\beta - \beta^\top \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t)\beta)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t)\right)^{\dagger} \frac{1}{n}\sum_{t=1}^{n} \mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t, Y_t) \\
&= \Pi_{\mathcal{S}^{\mathrm{inv}}} \overline{\mathrm{Var}}(X)^{-1} \Pi_{\mathcal{S}^{\mathrm{inv}}} \overline{\mathrm{Cov}}(X, Y) \\
&= U^{\mathrm{inv}} ((U^{\mathrm{inv}})^{\top} \overline{\mathrm{Var}}(X) U^{\mathrm{inv}})^{-1} (U^{\mathrm{inv}})^{\top} \overline{\mathrm{Cov}}(X, Y).
\end{aligned}
$$

The fourth equality follows from Lemma 2.6.7. The fifth equality follows by Lemma 2.6.1 (see the proof of Lemma 2.6.7). In the last equality we used that $(U^{\mathrm{inv}})^{\top}\overline{\mathrm{Var}}(X)^{-1}U^{\mathrm{inv}} = ((U^{\mathrm{inv}})^{\top}\overline{\mathrm{Var}}(X)U^{\mathrm{inv}})^{-1}$, which follows from the properties of orthogonal matrices and the block diagonal structure of $U^{\top}\overline{\mathrm{Var}}(X)U$: these imply that $(U^{\top}\overline{\mathrm{Var}}(X)U)^{-1} = U^{\top}\overline{\mathrm{Var}}(X)^{-1}U$ and

$$
\begin{aligned}
&\begin{bmatrix} ((U^{S_1})^{\top}\overline{\mathrm{Var}}(X)U^{S_1})^{-1} & & \\ & \ddots & \\ & & ((U^{S_q})^{\top}\overline{\mathrm{Var}}(X)U^{S_q})^{-1} \end{bmatrix} \\
&= \begin{bmatrix} (U^{S_1})^{\top}\overline{\mathrm{Var}}(X)^{-1}U^{S_1} & & \\ & \ddots & \\ & & (U^{S_q})^{\top}\overline{\mathrm{Var}}(X)^{-1}U^{S_q} \end{bmatrix}.
\end{aligned}
$$

□

2.E. Proofs

2.E.1 Proof of Theorem 2.2.7

Proof. Let $\{\mathcal{S}_j\}_{j=1}^{q}$ be an irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition and define $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ as in (2.2.11). By Assumption 2, $\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ form an orthogonal and $(X_t)_{t\in\mathbb{N}}$-decorrelating partition and $\mathcal{S}^{\mathrm{inv}}$ is opt-invariant on $\mathbb{N}$. Therefore, using Lemma 2.2.4 and Lemma 2.2.5, we get for all $t \in \mathbb{N}$ that

$$
\gamma_{0,t} = \arg\max_{\beta\in\mathcal{S}^{\mathrm{inv}}} \Delta\mathrm{Var}_t(\beta) + \arg\max_{\beta\in\mathcal{S}^{\mathrm{res}}} \Delta\mathrm{Var}_t(\beta).
$$

Furthermore, $\mathcal{S}^{\mathrm{inv}}$ opt-invariant on $\mathbb{N}$ implies that the first term, $\arg\max_{\beta\in\mathcal{S}^{\mathrm{inv}}} \Delta\mathrm{Var}_t(\beta)$, does not depend on t and hence it holds for all $t \in \mathbb{N}$ that

$$
\gamma_{0,t} = \arg\max_{\beta\in\mathcal{S}^{\mathrm{inv}}} \overline{\Delta\mathrm{Var}}(\beta) + \arg\max_{\beta\in\mathcal{S}^{\mathrm{res}}} \Delta\mathrm{Var}_t(\beta).
$$

Moreover, Assumption 1 ensures that the invariant and residual subspaces can be uniquely identified from an arbitrary irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition, and therefore all the above results do not depend on the considered partition. This concludes the proof of Theorem 2.2.7. □

2.E.2 Proof of Proposition 2.2.8

Proof. (i) For all $j \in \{1, \dots, q_{\max}^U\}$, let U^{S_j} denote the submatrix of U formed by the columns indexed by S_j. Then, the orthogonal projection matrix onto the subspace $\mathcal{S}_j$ can be expressed as $\Pi_{\mathcal{S}_j} = U^{S_j}(U^{S_j})^\top$. It follows for all $t \in [n]$ that

$$\begin{aligned}\operatorname{Cov}(\Pi_{\mathcal{S}_i} X_t, \Pi_{\mathcal{S}_j} X_t) &= \Pi_{\mathcal{S}_i} \Sigma_t \Pi_{\mathcal{S}_j} \\ &= U^{S_i}(U^{S_i})^\top \Sigma_t U^{S_j}(U^{S_j})^\top \\ &= 0,\end{aligned}$$

where the last equality holds since $(U^{S_i})^\top \Sigma_t U^{S_j}$ is the (i,j)-th (off-diagonal) block of $\tilde{\Sigma}_t := U^\top \Sigma_t U$, which is zero due to the block diagonal structure of $\tilde{\Sigma}_t$. Irreducibility of the orthogonal partition follows from the irreducibility of the joint block diagonal decomposition.

(ii) The statement follows from Lemma 2.6.2, taking $\mathcal{N} = [n]$. Irreducibility of the joint block diagonalizer follows from the irreducibility of the orthogonal partition. □

2.E.3 Proof of Proposition 2.2.10

Proof. (i) First, observe that, by Lemma 2.6.4, $\{\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}\}$ is an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition and that, by Lemma 2.6.5, $\mathcal{S}^{\text{inv}}$ is opt-invariant on $[n]$. Then, by definition of β^{inv} and by Lemma 2.6.7 we get for all $t \in [n]$ that

$$\beta^{\text{inv}} = \arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \overline{\Delta\text{Var}}(\beta) = \Pi_{\mathcal{S}^{\text{inv}}} \gamma_{0,t} = \Pi_{\mathcal{S}^{\text{inv}}} \bar{\gamma}_0.$$

(ii) We need to prove for all $t \in [n]$ that $\operatorname{Cov}(Y_t - X_t^\top \beta^{\text{inv}}, X_t^\top \beta^{\text{inv}}) = 0$. To see this, fix $t \in [n]$. Then

$$\begin{aligned}\operatorname{Cov}(Y_t - X_t^\top \beta^{\text{inv}}, X_t^\top \beta^{\text{inv}}) &= \operatorname{Cov}(X_t^\top(\gamma_{0,t} - \beta^{\text{inv}}), X_t^\top \beta^{\text{inv}}) \\ &= \operatorname{Cov}(X_t^\top(\Pi_{\mathcal{S}^{\text{inv}}}\gamma_{0,t} + \Pi_{\mathcal{S}^{\text{res}}}\gamma_{0,t} - \beta^{\text{inv}}), X_t^\top(\Pi_{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}})) \\ &= \operatorname{Cov}(X_t^\top \Pi_{\mathcal{S}^{\text{res}}}\gamma_{0,t}, X_t^\top \Pi_{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}}) \\ &= \gamma_{0,t}^\top \operatorname{Cov}(\Pi_{\mathcal{S}^{\text{res}}} X_t, \Pi_{\mathcal{S}^{\text{inv}}} X_t)\beta^{\text{inv}} \\ &= 0,\end{aligned}$$

where the third equality uses $\beta^{\text{inv}} = \Pi_{\mathcal{S}^{\text{inv}}}\gamma_{0,t}$ by Proposition 2.2.10 (i) and the last equality follows from the fact that $\mathcal{S}^{\text{inv}}$, $\mathcal{S}^{\text{res}}$ are an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition by Lemma 2.6.4.

(iii) Let $\mathcal{B}^n$ denote the set of all time-invariant parameters over $[n]$. By assumption, we have that $\mathcal{B}^n \subseteq \mathcal{S}^{\text{inv}}$. Moreover, by point (ii), we have that $\beta^{\text{inv}} \in \mathcal{B}^n$. Therefore, by definition of β^{inv} we obtain that $\beta^{\text{inv}} = \arg\max_{\beta\in\mathcal{B}^n} \overline{\Delta\text{Var}}(\beta)$.

This completes the proof of Proposition 2.2.10. □

2.E.4 Proof of Theorem 2.3.1

Proof. Under Assumptions 1 and 2 and using the definitions of β^{inv} and δ_t^{res}, we can write the explained variance of $\beta \in \mathbb{R}^p$ at time $t \in \mathbb{N}$ for the true time varying parameter $\gamma_{0,t} \in \mathbb{R}^p$ as

$$\begin{aligned}
\Delta\text{Var}_t(\beta) &= 2\gamma_{0,t}^\top \text{Var}(X_t)\beta - \beta^\top \text{Var}(X_t)\beta \\
&= 2(\beta^{\text{inv}} + \delta_t^{\text{res}})^\top \text{Var}((\Pi_{\mathcal{S}^{\text{inv}}} + \Pi_{\mathcal{S}^{\text{res}}})X_t)\beta - \beta^\top \text{Var}(X_t)\beta \\
&= 2(\beta^{\text{inv}})^\top \text{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_t)\Pi_{\mathcal{S}^{\text{inv}}}\beta + 2(\delta_t^{\text{res}})^\top \text{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_t)\Pi_{\mathcal{S}^{\text{res}}}\beta - \beta^\top \text{Var}(X_t)\beta.
\end{aligned}$$

Using this expansion, we get, since $\delta_t^{\text{res}} = \gamma_{0,t} - \beta^{\text{inv}}$, for all $\beta \notin \mathcal{S}^{\text{inv}}$ that

$$\inf_{\substack{\gamma_{0,t} \in \mathbb{R}^p: \\ \gamma_{0,t} - \beta^{\text{inv}} \in \mathcal{S}^{\text{res}}}} \Delta\text{Var}_t(\beta) = -\infty.$$

Therefore,

$$\begin{aligned}
\arg\max_{\beta \in \mathbb{R}^p} \inf_{\substack{\gamma_{0,t} \in \mathbb{R}^p: \\ \gamma_{0,t} - \beta^{\text{inv}} \in \mathcal{S}^{\text{res}}}} \Delta\text{Var}_t(\beta) &= \arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \inf_{\substack{\gamma_{0,t} \in \mathbb{R}^p: \\ \gamma_{0,t} - \beta^{\text{inv}} \in \mathcal{S}^{\text{res}}}} \Delta\text{Var}_t(\beta) \\
&= \arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \Delta\text{Var}_t(\beta).
\end{aligned}$$

Since by assumption it holds that $\mathcal{S}^{\text{inv}}$ is opt-invariant on $\mathbb{N}$, it further holds that for all $t \in \mathbb{N}$

$$\arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \Delta\text{Var}_t(\beta) = \arg\max_{\beta \in \mathcal{S}^{\text{inv}}} \overline{\Delta\text{Var}(\beta)}.$$

The claim follows from the definition of β^{inv}. □

2.E.5 Proof of Theorem 2.4.1

Proof. We assume without loss of generality that the observed predictors X_t have zero mean in $t \in \mathcal{I}^{\text{ad}} \cup t^*$. Alternatively, as mentioned in Section 2.4.1.2, a constant term could be added to X_t to account for the mean. We also observe that $\hat{\gamma}_{t^*}^{\text{OLS}}$ and $\hat{\delta}_{t^*}^{\text{res}}$ are both unbiased estimators for γ_{0,t^*} and $\delta_{t^*}^{\text{res}}$, respectively (this can be checked using standard OLS analysis). We now start by computing the out-of-sample MSPE for $\hat{\gamma}_{t^*}^{\text{ISD}}$.

$$\begin{aligned}
\text{MSPE}(\hat{\gamma}_{t^*}^{\text{ISD}}) &= \mathbb{E}[(X_{t^*}^\top(\hat{\gamma}_{t^*}^{\text{ISD}} - \gamma_{0,t^*}))^2] \\
&= \mathbb{E}[(X_{t^*}^\top(\Pi_{\mathcal{S}^{\text{inv}}} + \Pi_{\mathcal{S}^{\text{res}}})(\hat{\beta}^{\text{inv}} + \hat{\delta}_{t^*}^{\text{res}} - \beta^{\text{inv}} - \delta_{t^*}^{\text{res}}))^2] \\
&= \text{trace}(\mathbb{E}[(\hat{\beta}^{\text{inv}} - \beta^{\text{inv}})(\hat{\beta}^{\text{inv}} - \beta^{\text{inv}})^\top] \,\text{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*})) \\
&\quad + \text{trace}(\mathbb{E}[(\hat{\delta}_{t^*}^{\text{res}} - \delta_{t^*}^{\text{res}})(\hat{\delta}_{t^*}^{\text{res}} - \delta_{t^*}^{\text{res}})^\top] \,\text{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})) \\
&= \text{trace}(\mathbb{E}[\text{Var}(\hat{\beta}^{\text{inv}} \mid \mathbf{X})] \,\text{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*})) \\
&\quad + \text{trace}(\mathbb{E}[\text{Var}(\hat{\delta}_{t^*}^{\text{res}} \mid \mathbf{X}^{\text{ad}})] \,\text{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*}))
\end{aligned}$$

$$= \operatorname{trace}(\mathbb{E}[\operatorname{Var}(\hat{\beta}^{\text{inv}} \mid \mathbf{X})] \operatorname{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*})) + \tfrac{\sigma_{\text{ad}}^2}{m} \operatorname{trace}(\Pi_{\mathcal{S}^{\text{res}}} \mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}] \Pi_{\mathcal{S}^{\text{res}}} \operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})),$$

where we have used that $\operatorname{Var}(\hat{\delta}_{t^*}^{\text{res}} \mid \mathbf{X}^{\text{ad}}) = \sigma_{\text{ad}}^2 \Pi_{\mathcal{S}^{\text{res}}}((\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1} \Pi_{\mathcal{S}^{\text{res}}}$. We further observe that

$$\operatorname{Var}(\hat{\beta}^{\text{inv}} \mid \mathbf{X}) = \Pi_{\mathcal{S}^{\text{inv}}} (\mathbf{X}^\top \mathbf{X})^{-1} \Pi_{\mathcal{S}^{\text{inv}}} \mathbf{X}^\top \operatorname{Var}(\epsilon) \mathbf{X} \Pi_{\mathcal{S}^{\text{inv}}} (\mathbf{X}^\top \mathbf{X})^{-1} \Pi_{\mathcal{S}^{\text{inv}}},$$

where $\operatorname{Var}(\epsilon)$ is a $n \times n$ diagonal matrix whose diagonal elements are the error variances at each observed time steps, $(\operatorname{Var}(\epsilon_t))_{t \in [n]}$. Let $\sigma_{\epsilon,\min}^2 := \min_{t \in [n]} \operatorname{Var}(\epsilon_t)$ and $\sigma_{\epsilon,\max}^2 \geq \max_{t \in [n]} \operatorname{Var}(\epsilon_t)$. Then,

$$\sigma_{\epsilon,\min}^2 \Pi_{\mathcal{S}^{\text{inv}}} (\mathbf{X}^\top \mathbf{X})^{-1} \Pi_{\mathcal{S}^{\text{inv}}} \preceq \operatorname{Var}(\hat{\beta}^{\text{inv}} \mid \mathbf{X}) \preceq \sigma_{\epsilon,\max}^2 \Pi_{\mathcal{S}^{\text{inv}}} (\mathbf{X}^\top \mathbf{X})^{-1} \Pi_{\mathcal{S}^{\text{inv}}},$$

where $\preceq$ denotes the Loewner order, and it follows from $\operatorname{diag}_n(\sigma_{\epsilon,\min}^2) \preceq \operatorname{Var}(\epsilon_t) \preceq \operatorname{diag}_n(\sigma_{\epsilon,\max}^2)$, with $\operatorname{diag}_n(\cdot)$ denoting an n-dimensional diagonal matrix with diagonal elements all equal. We have used in particular that for two symmetric matrices $A, B \in \mathbb{R}^{n \times n}$ such that $A \succeq B$ and a matrix $S \in \mathbb{R}^{n \times n}$ it holds that $S^\top A S \succeq S^\top B S$ [see, for example, Theorem 7.7.2. by Horn and Johnson, 2012]. Using this relation and Jensen's inequality, we obtain that the first term in $\operatorname{MSPE}(\hat{\gamma}_{t^*}^{\text{ISD}})$ is lower bounded by

$$\begin{aligned}
&\operatorname{trace}(\mathbb{E}[\operatorname{Var}(\hat{\beta}^{\text{inv}} \mid \mathbf{X})] \operatorname{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*})) \\
\geq &\sigma_{\epsilon,\min}^2 \operatorname{trace}(\Pi_{\mathcal{S}^{\text{inv}}} \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] \Pi_{\mathcal{S}^{\text{inv}}} \operatorname{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*})) \\
\geq &\tfrac{\sigma_{\epsilon,\min}^2}{n} \operatorname{trace}(\Pi_{\mathcal{S}^{\text{inv}}} \mathbf{\Sigma}^{-1} \Pi_{\mathcal{S}^{\text{inv}}} \Sigma_{t^*} \Pi_{\mathcal{S}^{\text{inv}}}),
\end{aligned}$$

where $\mathbf{\Sigma} := \mathbb{E}[\frac{1}{n} \mathbf{X}^\top \mathbf{X}]$. Using now that, by assumption, $\{\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}\}$ is an orthogonal and $(X_t)_{t \in \mathbb{N}}$-decorrelating partition, let $U \in \mathbb{R}^{p \times p}$ be defined for such a partition as in Lemma 2.6.2 and let $U^{\text{inv}} \in \mathbb{R}^{p \times \dim(\mathcal{S}^{\text{inv}})}$ be the submatrix of U whose columns form an orthonormal basis for $\mathcal{S}^{\text{inv}}$, such that $\Pi_{\mathcal{S}^{\text{inv}}} = U^{\text{inv}} (U^{\text{inv}})^\top$. Moreover, let $\tilde{\mathbf{\Sigma}}^{\text{inv}} := (U^{\text{inv}})^\top \mathbf{\Sigma} U^{\text{inv}}$ and $\tilde{\Sigma}_{t^*}^{\text{inv}} := (U^{\text{inv}})^\top \Sigma_{t^*} U^{\text{inv}}$ denote, as in Lemma 2.6.2, the diagonal block corresponding to $\mathcal{S}^{\text{inv}}$ of the block diagonal matrices $U^\top \mathbf{\Sigma} U$ and $U^\top \Sigma_{t^*} U$, respectively. By the properties of orthogonal matrices, $(\tilde{\mathbf{\Sigma}}^{\text{inv}})^{-1} = (U^{\text{inv}})^\top \mathbf{\Sigma}^{-1} U^{\text{inv}}$. For all $i \in \{1, \ldots, \dim(\mathcal{S}^{\text{inv}})\}$, let λ_i denote the i-th eigenvalue in decreasing order. Then, we can further express the above lower bound as

$$\begin{aligned}
&\tfrac{\sigma_{\epsilon,\min}^2}{n} \operatorname{trace}(\Pi_{\mathcal{S}^{\text{inv}}} \mathbf{\Sigma}^{-1} \Pi_{\mathcal{S}^{\text{inv}}} \Sigma_{t^*} \Pi_{\mathcal{S}^{\text{inv}}}) \\
= &\tfrac{\sigma_{\epsilon,\min}^2}{n} \operatorname{trace}(U^{\text{inv}} (U^{\text{inv}})^\top \mathbf{\Sigma}^{-1} U^{\text{inv}} (U^{\text{inv}})^\top \Sigma_{t^*} U^{\text{inv}} (U^{\text{inv}})^\top) \\
= &\tfrac{\sigma_{\epsilon,\min}^2}{n} \operatorname{trace}(U^{\text{inv}} (\tilde{\mathbf{\Sigma}}^{\text{inv}})^{-1} \tilde{\Sigma}_{t^*}^{\text{inv}} (U^{\text{inv}})^\top) \\
\geq &\sigma_{\epsilon,\min}^2 \tfrac{\dim(\mathcal{S}^{\text{inv}})}{n} \tfrac{\lambda_{\min}(\tilde{\Sigma}_{t^*}^{\text{inv}})}{\lambda_{\max}(\tilde{\mathbf{\Sigma}}^{\text{inv}})} \\
= &\sigma_{\epsilon,\min}^2 \tfrac{\dim(\mathcal{S}^{\text{inv}})}{n} c_{\text{inv}},
\end{aligned}$$

where we define $c_{\text{inv}} \coloneqq \lambda_{\min}(\tilde{\Sigma}^{\text{inv}}_{t^*})/\lambda_{\max}(\tilde{\mathbf{\Sigma}}^{\text{inv}})$. The same term in $\text{MSPE}(\hat{\gamma}^{\text{ISD}}_{t^*})$ is upper bounded by

$$\begin{aligned}
&\operatorname{trace}(\mathbb{E}[\operatorname{Var}(\hat{\beta}^{\text{inv}} \mid \mathbf{X})] \operatorname{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*})) \\
\leq &\sigma^2_{\epsilon,\max} \operatorname{trace}(\Pi_{\mathcal{S}^{\text{inv}}} \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] \Pi_{\mathcal{S}^{\text{inv}}} \operatorname{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*})) \\
\leq &\sigma^2_{\epsilon,\max} \frac{\dim(\mathcal{S}^{\text{inv}})}{n} C_{\text{inv}},
\end{aligned}$$

where $C_{\text{inv}} \coloneqq 1 + \operatorname{trace}(\Pi_{\mathcal{S}^{\text{inv}}} (\mathbb{E}[(\frac{1}{n}\mathbf{X}^\top \mathbf{X})^{-1}] - \Sigma^{-1}_{t^*}) \Pi_{\mathcal{S}^{\text{inv}}} \operatorname{Var}(\Pi_{\mathcal{S}^{\text{inv}}} X_{t^*}))$. Proceeding in a similar way, we can express the second term in $\text{MSPE}(\hat{\gamma}^{\text{ISD}}_{t^*})$ as

$$\frac{\sigma^2_{\text{ad}}}{m} \operatorname{trace}(\Pi_{\mathcal{S}^{\text{res}}} \mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}] \Pi_{\mathcal{S}^{\text{res}}} \operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})) = \frac{\sigma^2_{\text{ad}}}{m} h_{\text{res}}(m),$$

where $h_{\text{res}}(m) \coloneqq \operatorname{trace}(\Pi_{\mathcal{S}^{\text{res}}} \mathbb{E}[(\frac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}] \Pi_{\mathcal{S}^{\text{res}}} \operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*}))$. Since we have assumed that for all $t \in \mathcal{I}^{\text{ad}} \cup \{t^*\}$ the distribution of X_t does not change, by Jensen's inequality and by the fact that $\Pi_{\mathcal{S}^{\text{res}}} \Sigma^{-1}_{t^*} \Pi_{\mathcal{S}^{\text{res}}} \Sigma_{t^*} \Pi_{\mathcal{S}^{\text{res}}} = \Pi_{\mathcal{S}^{\text{res}}}$ (see Lemma 2.6.4 and (2.D.6)) we have that $h_{\text{res}}(m) \geq \dim(\mathcal{S}^{\text{res}})$, and $\lim_{m\to\infty} h_{\text{res}}(m) = \dim(\mathcal{S}^{\text{res}})$. Moreover, we can find an upper bound for $h_{\text{res}}(m)$ in the following way.

$$\begin{aligned}
h_{\text{res}}(m) &= \operatorname{trace}(\Pi_{\mathcal{S}^{\text{res}}} \mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}] \Pi_{\mathcal{S}^{\text{res}}} \operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})) \\
&= \operatorname{trace}(\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}] \operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})) \\
&\leq \operatorname{trace}(\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}]) \operatorname{trace}(\operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})) \\
&\leq \|\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}]\|_{\text{op}} \operatorname{trace}(\operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})) \\
&\leq \mathbb{E}[\|(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}\|_{\text{op}}] \operatorname{trace}(\operatorname{Var}(\Pi_{\mathcal{S}^{\text{res}}} X_{t^*})) \\
&\leq c^{-1} \lambda_{\max}(\Sigma_{t^*}) \dim(\mathcal{S}^{\text{res}}).
\end{aligned}$$

Summarizing, we obtain that

$$\text{MSPE}(\hat{\gamma}^{\text{ISD}}_{t^*}) \geq \sigma^2_{\epsilon,\min} \frac{\dim(\mathcal{S}^{\text{inv}})}{n} c_{\text{inv}} + \sigma^2_{\text{ad}} \frac{\dim(\mathcal{S}^{\text{res}})}{m}$$

and

$$\text{MSPE}(\hat{\gamma}^{\text{ISD}}_{t^*}) \leq \sigma^2_{\epsilon,\max} \frac{\dim(\mathcal{S}^{\text{inv}})}{n} C_{\text{inv}} + \sigma^2_{\text{ad}} \frac{\dim(\mathcal{S}^{\text{res}})}{m} C_{\text{res}},$$

where $C_{\text{res}} \coloneqq c^{-1} \lambda_{\max}(\Sigma_{t^*})$ is the constant introduced in the theorem statement.

We now compute the MSPE of $\hat{\gamma}^{\text{OLS}}_{t^*}$.

$$\begin{aligned}
\text{MSPE}(\hat{\gamma}^{\text{OLS}}_{t^*}) &= \mathbb{E}[(X^\top_{t^*}(\hat{\gamma}^{\text{OLS}}_{t^*} - \gamma_{0,t^*}))^2] \\
&= \operatorname{trace}(\mathbb{E}[(\hat{\gamma}^{\text{OLS}}_{t^*} - \gamma_{0,t^*})(\hat{\gamma}^{\text{OLS}}_{t^*} - \gamma_{0,t^*})^\top X_{t^*} X^\top_{t^*}]) \\
&= \operatorname{trace}(\mathbb{E}[(\hat{\gamma}^{\text{OLS}}_{t^*} - \gamma_{0,t^*})(\hat{\gamma}^{\text{OLS}}_{t^*} - \gamma_{0,t^*})^\top] \operatorname{Var}(X_{t^*})) \\
&= \operatorname{trace}(\mathbb{E}[\operatorname{Var}(\hat{\gamma}^{\text{OLS}}_{t^*} \mid \mathbf{X}^{\text{ad}})] \Sigma_{t^*})) \\
&= \sigma^2_{\text{ad}} \operatorname{trace}(\mathbb{E}[((\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}] \Sigma_{t^*})) \\
&= \tfrac{\sigma^2_{\text{ad}}}{m} \operatorname{trace}(\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\text{ad}})^\top \mathbf{X}^{\text{ad}})^{-1}] \Sigma_{t^*})).
\end{aligned}$$

We can further express $\mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{OLS}})$ as

$$\begin{aligned}
&\tfrac{\sigma_{\mathrm{ad}}^2}{m}\,\mathrm{trace}(\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}]\Sigma_{t^*})\\
=&\tfrac{\sigma_{\mathrm{ad}}^2}{m}\,\mathrm{trace}((\Pi_{\mathcal{S}^{\mathrm{inv}}}+\Pi_{\mathcal{S}^{\mathrm{res}}})\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}](\Pi_{\mathcal{S}^{\mathrm{inv}}}+\Pi_{\mathcal{S}^{\mathrm{res}}})\,\mathrm{Var}((\Pi_{\mathcal{S}^{\mathrm{inv}}}+\Pi_{\mathcal{S}^{\mathrm{res}}})X_{t^*}))\\
=&\tfrac{\sigma_{\mathrm{ad}}^2}{m}\,\mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}}\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}]\Pi_{\mathcal{S}^{\mathrm{inv}}}\,\mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}}X_{t^*}))\\
&+\tfrac{\sigma_{\mathrm{ad}}^2}{m}\,\mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{res}}}\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}]\Pi_{\mathcal{S}^{\mathrm{res}}}\,\mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}}X_{t^*})).
\end{aligned}$$

In particular, the second term in the above sum also appears in $\mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{ISD}})$. Taking now the difference between the MSPE of $\hat{\gamma}_{t^*}^{\mathrm{OLS}}$ and $\hat{\gamma}_{t^*}^{\mathrm{ISD}}$, we obtain that

$$\begin{aligned}
\mathrm{MSPE}(\gamma_{t^*}^{\mathrm{OLS}})-\mathrm{MSPE}(\gamma_{t^*}^{\mathrm{ISD}}) = &\tfrac{\sigma_{\mathrm{ad}}^2}{m}\,\mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}}\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}]\Pi_{\mathcal{S}^{\mathrm{inv}}}\,\mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}}X_{t^*}))\\
&-\mathrm{trace}(\mathbb{E}[\mathrm{Var}(\hat{\beta}^{\mathrm{inv}}\mid\mathbf{X})]\,\mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}}X_{t^*}))
\end{aligned}$$

We have already obtained an upper bound for the second term in the difference, namely $\sigma_{\epsilon,\max}^2\frac{\dim(\mathcal{S}^{\mathrm{inv}})}{n}C_{\mathrm{inv}}$. For the first term, we can make the same considerations made above for $\frac{\sigma_{\mathrm{ad}}^2}{m}\,\mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{res}}}\mathbb{E}[(\frac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}]\Pi_{\mathcal{S}^{\mathrm{res}}}\,\mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}}X_{t^*}))$ and $h_{\mathrm{res}}(m)$. In particular,

$$\tfrac{\sigma_{\mathrm{ad}}^2}{m}\,\mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}}\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}]\Pi_{\mathcal{S}^{\mathrm{inv}}}\,\mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}}X_{t^*}))=\tfrac{\sigma_{\mathrm{ad}}^2}{m}h_{\mathrm{ad,inv}}(m)$$

where

$$h_{\mathrm{ad,inv}}(m)\coloneqq\mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}}\mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top\mathbf{X}^{\mathrm{ad}})^{-1}]\Pi_{\mathcal{S}^{\mathrm{inv}}}\,\mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}}X_{t^*}))$$

satisfies $h_{\mathrm{ad,inv}}(m)\geq\dim(\mathcal{S}^{\mathrm{inv}})$ and $h_{\mathrm{ad,inv}}(m)\leq c^{-1}\lambda_{\max}(\Sigma_{t^*})\dim(\mathcal{S}^{\mathrm{inv}})$ (as for $h_{\mathrm{res}}(m)$, we observe that $\lim_{m\to\infty}h_{\mathrm{ad,inv}}(m)=\dim(\mathcal{S}^{\mathrm{inv}})$). Therefore, we obtain that

$$\mathrm{MSPE}(\gamma_{t^*}^{\mathrm{OLS}})-\mathrm{MSPE}(\gamma_{t^*}^{\mathrm{ISD}})\geq\sigma_{\mathrm{ad}}^2\tfrac{\dim(\mathcal{S}^{\mathrm{inv}})}{m}-\sigma_{\epsilon,\max}^2\tfrac{\dim(\mathcal{S}^{\mathrm{inv}})}{n}C_{\mathrm{inv}}$$

and

$$\mathrm{MSPE}(\gamma_{t^*}^{\mathrm{OLS}})-\mathrm{MSPE}(\gamma_{t^*}^{\mathrm{ISD}})\leq\sigma_{\mathrm{ad}}^2\tfrac{\dim(\mathcal{S}^{\mathrm{inv}})}{m}C_{\mathrm{ad,inv}}-\sigma_{\epsilon,\min}^2\tfrac{\dim(\mathcal{S}^{\mathrm{inv}})}{n}c_{\mathrm{inv}},$$

where $C_{\mathrm{ad,inv}}\coloneqq c^{-1}\lambda_{\max}(\Sigma_{t^*})$. In particular, this difference is always positive if n is sufficiently large. Finally, we can observe that the expected explained variance (2.4.24) for $\hat{\gamma}$, where $\hat{\gamma}=\hat{\gamma}_{t^*}^{\mathrm{ISD}}$ or $\hat{\gamma}=\hat{\gamma}_{t^*}^{\mathrm{OLS}}$, is

$$\begin{aligned}
&\mathbb{E}[\Delta\mathrm{Var}_{t^*}(\hat{\gamma})]\\
&\quad=\mathbb{E}[\mathrm{Var}(Y_{t^*})-\mathrm{Var}(Y_{t^*}-X_{t^*}^\top\hat{\gamma}\mid\hat{\gamma})]\\
&\quad=\mathrm{Var}(X_{t^*}^\top\gamma_{0,t^*}+\epsilon_{t^*})-\mathbb{E}[\mathrm{Var}(X_{t^*}^\top(\gamma_{0,t^*}-\hat{\gamma})+\epsilon_{t^*}\mid\hat{\gamma})]\\
&\quad=\gamma_{0,t^*}^\top\,\mathrm{Var}(X_{t^*})\gamma_{0,t^*}+\sigma_{\epsilon^*}^2-\mathbb{E}[(\gamma_{0,t^*}-\hat{\gamma})^\top\,\mathrm{Var}(X_{t^*})(\gamma_{0,t^*}-\hat{\gamma})]-\sigma_{\epsilon^*}^2\\
&\quad=\gamma_{0,t^*}^\top\,\mathrm{Var}(X_{t^*})\gamma_{0,t^*}-\mathbb{E}[\mathrm{trace}((\gamma_{0,t^*}-\hat{\gamma})(\gamma_{0,t^*}-\hat{\gamma})^\top\,\mathrm{Var}(X_{t^*}))]\\
&\quad=\gamma_{0,t^*}^\top\Sigma_{t^*}\gamma_{0,t^*}-\mathrm{MSPE}(\hat{\gamma})
\end{aligned}$$

and therefore the same inequalities found for the MSPEs difference equivalently hold for $\mathbb{E}[\Delta\mathrm{Var}_{t^*}(\hat{\gamma}_{t^*}^{\mathrm{ISD}})] - \mathbb{E}[\Delta\mathrm{Var}_{t^*}(\hat{\gamma}_{t^*}^{\mathrm{OLS}})]$. □

2.E.6 Proof of Proposition 2.6.1

Proof. (i) By definition of a (non-orthogonal) joint block diagonalizer, it holds for all $t \in [n]$ that the matrix $\widetilde{\Sigma}_t \coloneqq U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\widetilde{\Sigma}_{t,j} \coloneqq (U^{S_j})^\top \Sigma_t U^{S_j}$, $j \in \{1, \dots, q\}$. Define now the matrix $W \coloneqq U^{-\top}$, and observe that the following relations hold

$$\begin{aligned}
&\Sigma_t = W \widetilde{\Sigma}_t W^\top, \quad \Sigma_t^{-1} = U \widetilde{\Sigma}_t^{-1} U^\top \quad \text{and} \\
&\widetilde{\Sigma}_t^{-1} = W^\top \Sigma_t^{-1} W \quad \text{block diagonal with blocks } \widetilde{\Sigma}_{t,j}^{-1} \coloneqq (W^{S_j})^\top \Sigma_t^{-1} W^{S_j}.
\end{aligned}$$

The last relation is obtained by observing that

$$\tilde{\Sigma}_t^{-1} = \begin{bmatrix} \tilde{\Sigma}_{t,1}^{-1} & & \\ & \ddots & \\ & & \tilde{\Sigma}_{t,q}^{-1} \end{bmatrix} = \begin{bmatrix} (W^{S_1})^\top \\ \vdots \\ (W^{S_q})^\top \end{bmatrix} \Sigma_t^{-1} \begin{bmatrix} W^{S_1} & \dots & W^{S_q} \end{bmatrix}.$$

Moreover, $\tilde{\Sigma}_{t,j}^{-1} = \mathrm{Var}((U^{S_j})^\top X_t)^{-1}$. We observe that, because $W^\top U = I_p$, it holds for all $i, j \in \{1, \dots, q\}$ that $(W^{S_j})^\top U^{S_i} = 0$, that is, the space spanned by the columns of W^{S_j} is orthogonal to the space spanned by the columns of U^{S_i}. Moreover, the matrix $U^{S_j}(W^{S_j})^\top$ is an oblique projection matrix onto $\mathcal{S}_j$ along $\bigoplus_{i \in \{1,\dots,q\} \setminus \{j\}} \mathcal{S}_i$. Fix $j \in \{1, \dots, q\}$. Then, using these relations we obtain that

$$\begin{aligned}
P_{\mathcal{S}_j | \mathcal{S}_{-j}} \gamma_{0,t} &= U^{S_j} (W^{S_j})^\top \Sigma_t^{-1} \mathrm{Cov}(X_t, Y_t) \\
&= U^{S_j} (W^{S_j})^\top \Sigma_t^{-1} W U^\top \mathrm{Cov}(X_t, Y_t) \\
&= U^{S_j} (W^{S_j})^\top \Sigma_t^{-1} \begin{bmatrix} W^{S_1} & \dots & W^{S_q} \end{bmatrix} \begin{bmatrix} (U^{S_1})^\top \\ \vdots \\ (U^{S_q})^\top \end{bmatrix} \mathrm{Cov}(X_t, Y_t) \\
&= U^{S_j} (W^{S_j})^\top \Sigma_t^{-1} W^{S_j} (U^{S_j})^\top \mathrm{Cov}(X_t, Y_t) \\
&= U^{S_j} \tilde{\Sigma}_{t,j}^{-1} \mathrm{Cov}((U^{S_j})^\top X_t, Y_t) \\
&= U^{S_j} \mathrm{Var}((U^{S_j})^\top X_t)^{-1} \mathrm{Cov}((U^{S_j})^\top X_t, Y_t).
\end{aligned}$$

The fourth equality follows from the block diagonal structure of $W^\top \Sigma_t^{-1} W$, which implies that, for all $j \neq 1$, $(W^{S_1})^\top \Sigma_t^{-1} W^{S_j} = 0$.

(ii) By definition of β^{inv} and using point (i) of this proposition and that $\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ form a $(X_t)_{t \in [n]}$-decorrelating partition, we obtain that, for all $t \in [n]$,

$$\beta^{\mathrm{inv}} = U^{\mathrm{inv}} \mathrm{Var}((U^{S^{\mathrm{inv}}})^\top X_t)^{-1} (U^{\mathrm{inv}})^\top \mathrm{Cov}(X_t, Y_t).$$

We now observe that $\tilde{\beta}^{\mathrm{inv}} \coloneqq \mathrm{Var}((U^{S^{\mathrm{inv}}})^\top X_t)^{-1} (U^{\mathrm{inv}})^\top \mathrm{Cov}(X_t, Y_t)$ is the OLS solution

of regressing Y_t onto $(U^{\text{inv}})^\top X_t$ (as we assume all variables have zero mean). Moreover, this quantity is constant by proj-invariance of $\mathcal{S}^{\text{inv}}$. In particular, this means that $\tilde{\beta}^{\text{inv}} = \arg\min_{\beta \in \mathbb{R}^{\dim(\mathcal{S}^{\text{inv}})}} \mathbb{E}[\frac{1}{n}\sum_{t=1}^{n}(Y_t - X_t^\top U^{\text{inv}}\beta)^2]$, which implies that

$$\tilde{\beta}^{\text{inv}} = ((U^{\text{inv}})^\top \overline{\text{Var}}(X) U^{\text{inv}})^{-1} (U^{\text{inv}})^\top \overline{\text{Cov}}(X, Y).$$

Therefore, $\beta^{\text{inv}} = U^{\text{inv}}((U^{\text{inv}})^\top \overline{\text{Var}}(X) U^{\text{inv}})^{-1} (U^{\text{inv}})^\top \overline{\text{Cov}}(X, Y)$.

(iii) The claim follows from the definition of δ_t^{res}, from the fact that $\{\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}\}$ forms a $(X_t)_{t\in[n]}$-decorrelating partition and from point (i) of this proposition. □

3 Invariance-based dynamic regret minimization

Margherita Lazzaretto, Jonas Peters and Niklas Pfister

Abstract

We consider stochastic non-stationary linear bandits where the linear parameter connecting contexts to the reward changes over time. Existing algorithms in this setting localize the policy by gradually discarding or down-weighting past data, effectively shrinking the time horizon over which learning can occur. However, in many settings historical data may still carry partial information about the reward model. We propose to leverage such data while adapting to changes, by assuming the reward model decomposes into stationary and non-stationary components. Based on this assumption, we introduce ISD-linUCB, an algorithm that uses past data to learn invariances in the reward model and subsequently exploits them to improve online performance. We show both theoretically and empirically that leveraging invariance reduces the problem dimensionality, yielding significant regret improvements in fast-changing environments when sufficient historical data is available.

3.1 Introduction

A stochastic contextual bandit models an online decision-making process in which an agent, over $T \in \mathbb{N}$ rounds, sequentially selects actions based on contextual information [see, e.g., Lattimore and Szepesvári, 2020]. The agent's objective is to learn a decision policy that selects actions in a way that maximizes a cumulative reward, by balancing exploration of new actions and exploitation of acquired knowledge. Bandit algorithms provide the simplest framework for decision problems under uncertainty, where the actions of the agent do not affect the environment, often serving as a starting point for more complex models such as reinforcement learning. The design of stochastic contextual bandit algorithms relies on assumptions that specify the class of possible reward distributions over which the agent learns. We consider the case of linear time-varying reward functions, where the expected reward is assumed to depend linearly on some p-dimensional context-action feature.

Regret is the main performance measure of bandit algorithms. It corresponds to the difference between the cumulative reward of an optimal (oracle) sequence of actions and

the action selected by the bandit algorithm. The analysis of linear bandit algorithms involves studying finite sample upper and lower bounds for the regret in terms of the time horizon T and the dimensionality p. The standard setup assumes that the environment in which the agent acts is stationary, and is widely studied in the literature [see, for example, Lattimore and Szepesvári, 2020, a detailed review of the bandit literature is provided in Appendix 3.A.]. Dani et al. [2008] show a lower bound for the regret of the stationary stochastic linear bandit problem of $\Omega(p\sqrt{T})$. Different algorithms have been proposed in the literature achieving an upper bound that match this lower bound up to logarithmic factors, e.g., based on strategies using upper confidence bound (UCB) [Auer, 2002, Dani et al., 2008, Abbasi-yadkori et al., 2011]. Recent works relax the stationarity assumption. In such settings, the evaluation of non-stationary bandit algorithms additionally takes into account a variation budget B_T, measuring how much the environment changes through the T rounds. When the underlying reward function changes through time, learning and updating a reliable policy can be challenging since the algorithm needs to continually explore the context-action space to detect and adapt to changes. Cheung et al. [2019] deal with non-stationarity by using a sliding window regularized least squares estimator, Russac et al. [2019] weight, instead, past observations by a discounting factor; Zhao et al. [2020] propose to periodically restart a standard linear bandit algorithm, where the restarting interval depends on B_T. All these algorithms achieve the same regret upper bound in terms of p, T and B_T given by $\tilde{O}(p^{\frac{7}{8}}T^{\frac{3}{4}}B_T^{\frac{1}{4}})$. The approach of Zhao et al. [2020] implies in particular that the excess regret compared to the stationary case arises from assuming a fixed linear parameter, and incurring an additional loss that depends on its variation between restarts.

With the overall goal of adapting more rapidly to environment changes, we investigate whether parts of the non-stationary reward function remain invariant through all rounds, so that the data collected further in the past can be exploited rather than being discounted or fully discarded. Invariant information allows, for example, to learn policies that select worst-case optimal actions [e.g., Saengkyongam et al., 2023]. Since purely invariant policies can be sub-optimal at specific time points, we are interested in learning an invariant policy that can be updated online. To do so, we rely on the invariant subspace decomposition (ISD) framework proposed by Lazzaretto et al. [2025] for regression in linear non-stationary settings. ISD splits the learning of the time-varying parameter into two lower-dimensional components, one of which is time-invariant of dimension $p^{\text{inv}} < p$: this allows us to use all available data for the invariant component estimation and thus to reduce prediction error. A schematic representation of the proposed algorithm is shown in Figure 3.1.1.

Our contributions are as follows.

(i) We propose a practical novel linear contextual bandit algorithm (ISD-linUCB) that reduces online adaptation to a lower-dimensional residual subspace by exploiting the ISD framework to estimate the invariant component from historical data.

(ii) We establish regret bounds scaling with the residual dimension $(p - p^{\text{inv}})$ rather than p, yielding significant improvements in rapidly changing environments when sufficient historical data is available.

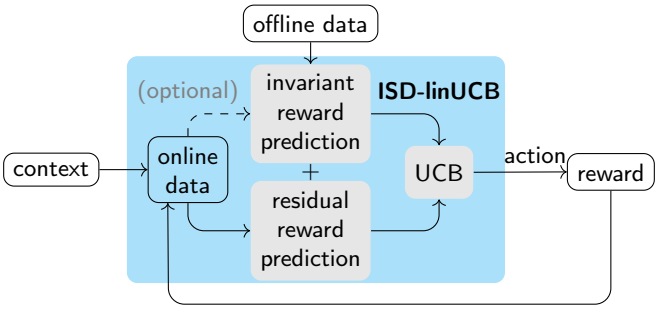


Figure 3.1.1: ISD-linUCB exploits historical data to improve reward predictions used by the UCB policy.

3.1.1 Notation

For all $n \in \mathbb{N}$, we define $[n] \coloneqq \{1, \dots, n\}$ and $[-n] \coloneqq \{-n, \dots, -1\}$. For all linear subspaces $\mathcal{S} \subseteq \mathbb{R}^p$, we denote by $\Pi^{\mathcal{S}} \in \mathbb{R}^{p \times p}$ the orthogonal projection matrix onto $\mathcal{S}$ and by $\mathcal{S}^\perp$ the orthogonal complement of $\mathcal{S}$ in $\mathbb{R}^p$. For all $x \in \mathbb{R}^p$ and for all positive semi-definite matrices $A \in \mathbb{R}^{p \times p}$, we define the norm $\|x\|_A \coloneqq \sqrt{x^\top A x}$. We further denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the minimum and maximum eigenvalues of A, respectively. For all $p \in \mathbb{N}$, we denote by I_p the p-dimensional identity matrix. We write $\tilde{O}(\cdot)$ for asymptotic bounds up to polylogarithmic factors, i.e., $b_n = \tilde{O}(a_n)$ if there exist constants $c > 0$ and $k \geq 0$ such that $b_n \leq c a_n \log(n)^k$ for sufficiently large n.

3.2 Problem setting

In a linear stochastic contextual bandit, an agent sequentially observes context $X_t \in \mathcal{X} \subseteq \mathbb{R}^p$ for rounds $t \in [T]$, where each X_t is drawn independently from previous contexts and has distribution $\mathbb{P}_{X_t}$. Given X_t, the agent selects an action $a_t \in \mathcal{A} = [K]$ and receives a noisy reward $R_t^{a_t}$. The reward for all $a \in \mathcal{A}$ is assumed to satisfy

$$R_t^a = \varphi(X_t, a)^\top \gamma_{0,t} + \epsilon_t \tag{3.2.1}$$

where $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^p$ is a known context-action feature map, ϵ_t is a conditionally zero mean σ-sub-Gaussian noise variable with respect to the σ-algebra $\mathcal{F}_t = \sigma((\varphi(X_\tau, a_\tau), R_\tau^{a_\tau})_{\tau \in [t-1]}, \varphi(X_t, a_t))$, and $\gamma_{0,t} \in \mathbb{R}^p$ is an unknown linear parameter. As in the standard linear bandit setup [e.g., Lattimore and Szepesvári, 2020], we assume that the context-action features and the linear parameter are bounded, and define $L \coloneqq \max_{t \in \mathbb{N}} \sup_{a \in \mathcal{A}} \|\varphi(X_t, a)\|_2$ and $M \coloneqq \max_{t \in \mathbb{N}} \|\gamma_{0,t}\|_2$.

Linear bandit algorithms sequentially estimate the linear parameter $\gamma_{0,t}$ and use it to select the optimal action given the observed context X_t. The quality of the algorithm

can then be assessed by considering the *dynamic regret* defined by

$$\mathrm{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T}(R_t^{a_t^*} - R_t^{a_t})\right] \tag{3.2.2}$$

where $a_t^* := \arg\max_{a\in\mathcal{A}} \varphi(X_t, a)^\top \gamma_{0,t}$ is the unknown optimal action. We call the regret at time t, $\mathrm{reg}_t := \mathbb{E}(R_t^{a_t^*} - R_t^{a_t})$, *instantaneous regret*.

We consider a non-stationary setup in which both the parameter $\gamma_{0,t}$ and the context distribution $\mathbb{P}_{X_t}$ may vary across rounds. To be able to effectively estimate $\gamma_{0,t}$ in the non-stationary setting, we make two assumptions: (1) We assume $\gamma_{0,t}$ can be decomposed into a varying and an invariant part and (2) we assume that the varying part changes slowly so it can be seen as approximately constant within short time-intervals. To formalize (1), we introduce Assumption 1 below, which is based on the invariant subspace decomposition framework proposed by Lazzaretto et al. [2025] adapted to the linear bandit setup.

Assumption 1 (Invariant subspace decomposition (ISD))**.** *There exists a non-degenerate* invariant subspace decomposition *of* $\mathbb{R}^p$ *for the time-varying linear bandit model* (3.2.1)*, that is, a partition of* $\mathbb{R}^p$ *into two linear subspaces* $(\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}})$ *such that* $\dim(\mathcal{S}^{\mathrm{inv}}) := p^{\mathrm{inv}} < p$ *and* $\dim(\mathcal{S}^{\mathrm{res}}) := p^{\mathrm{res}} = p - p^{\mathrm{inv}}$*, satisfying*

i) $\mathcal{S}^{\mathrm{inv}} = (\mathcal{S}^{\mathrm{res}})^\perp$*;*

ii) $\forall t \in \mathbb{N}$, $\forall x \in \mathcal{X}$, $\forall a \in \mathcal{A}$*:* $\mathrm{Cov}(\Pi^{\mathcal{S}^{\mathrm{inv}}}\varphi(x,a), \Pi^{\mathcal{S}^{\mathrm{res}}}\varphi(x,a)) = 0$*;*

iii) $\forall t \in \mathbb{N}$, $\exists \beta^{\mathrm{inv}} \in \mathcal{S}^{\mathrm{inv}}$, $\exists \delta_t^{\mathrm{res}} \in \mathcal{S}^{\mathrm{res}}$*:* $\gamma_{0,t} = \beta^{\mathrm{inv}} + \delta_t^{\mathrm{res}}$*;*

iv) $\forall$ *partitions* $(\tilde{\mathcal{S}}^{\mathrm{inv}}, \tilde{\mathcal{S}}^{\mathrm{res}})$ *of* $\mathbb{R}^p$ *satisfying (i)—(iii):* $\dim(\tilde{\mathcal{S}}^{\mathrm{inv}}) \leq \dim(\mathcal{S}^{\mathrm{inv}})$.

The spaces $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ are called *invariant* and *residual subspace*, and β^{inv} and δ_t^{res} are called *invariant* and *residual component* of $\gamma_{0,t}$, respectively. Lazzaretto et al. [2025] show that both the invariant and the residual component can be expressed as the least squares solution to the regression problem in the corresponding subspace.

Using Assumption 1, we can now formalize the bandit setting. We decompose the problem into an offline phase followed by a shorter online phase.

Setting 3.2.1. *We have access to* $T_0 \in \mathbb{N}$ *offline observations* $(\varphi(X_t, a_t), a_t, R_t^{a_t})_{t\in[-T_0]}$*, collected by a bandit agent interacting with a changing environment satisfying model* (3.2.1)*. The online time horizon* $T \in \mathbb{N}$ *is such that, for all* $t \in [T]$*, the parameter* $\gamma_{0,t}$ *in model* (3.2.1) *is constant and Assumption 1 holds.*

The constant $\gamma_{0,t}$ assumption simplifies the analysis and is motivated by settings of other non-stationary bandit algorithms, where T may represent a single epoch in restarting algorithms [Zhao et al., 2020] or a sliding window [Russac et al., 2019]. In practice, our proposed algorithm sequentially moves previously seen observations to the offline data. This framing also facilitates comparison with stationary linear bandit algorithms.

While for stationary bandits the time horizon T tends to dominate the bounds, for non-stationary bandits the p term becomes more dominant because time horizons are

short due to the shifting distributions. By leveraging the offline data, we are able to reduce the upper bound from $\tilde{O}(p\sqrt{T})$ to $\tilde{O}(p^{\text{res}}\sqrt{T})$ where p^{res} is the dimension of the residual subspace (in Appendix 3.D we show that under Assumption 1 the lower bound for the regret of a linear bandit algorithm is $\Omega(p^{\text{res}}\sqrt{T})$).

We begin by recalling some concepts for stationary linear bandits relevant for our work in Section 3.2.1. We then introduce our method (Section 3.3) and analyze it in two steps: first assuming the decomposition $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ is known (Section 3.4.1.2), then extending the analysis to the case where it is estimated from data (Section 3.4.2).

3.2.1 Regret analysis for stationary linear bandits

In a linear contextual bandit algorithm, the exploration is normally based on the uncertainty in the linear parameter estimation. More specifically, let $\hat{\gamma}_t$ denote an estimate of $\gamma_{0,t}$ obtained as the solution to ridge regression with regularization parameter $\lambda > 0$, using observations $(X_\tau, R_\tau^{a_\tau})_{\tau=1}^{t-1}$. The instantaneous regret is usually upper bounded with high probability by a term proportional to the product of the estimation error on the linear parameter, $\|\hat{\gamma}_t - \gamma_{0,t}\|_{\hat{\Sigma}_{t-1}}$, and the context-action features norm, $\|\varphi(X_t, a_t)\|_{\hat{\Sigma}_{t-1}^{-1}}$, normalized by $\hat{\Sigma}_{t-1}$ and its inverse, respectively, where for all $t \in [T]$ we define the regularized sample covariance matrix

$$\hat{\Sigma}_t := \lambda I_p + \sum_{\tau=1}^{t} \varphi(X_\tau, a_\tau)\varphi(X_\tau, a_\tau)^\top. \tag{3.2.3}$$

This is the case, for example, both for the upper confidence bound based algorithm by Abbasi-yadkori et al. [2011] (LinUCB) and for the Thompson sampling based algorithm by Agrawal and Goyal [2013] (LinTS). Using the explicit expression for the linear parameter estimator and the triangle inequality, it holds that

$$\|\hat{\gamma}_t - \gamma_{0,t}\|_{\hat{\Sigma}_{t-1}} \leq \|\sum_{\tau=1}^{t-1} \varphi(X_\tau, a_\tau)^\top \epsilon_\tau\|_{\hat{\Sigma}_{t-1}^{-1}} + \sqrt{\lambda}\|\gamma_{0,t}\|_2. \tag{3.2.4}$$

A result that plays a key role in bounding the linear parameter estimation error is given by the following lemma by Abbasi-yadkori et al. [2011].

Lemma 3.2.2 (Abbasi-yadkori et al. [2011], Theorem 1)**.** *Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\epsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that ϵ_t is F_t-measurable and sub-Gaussian conditionally on F_{t-1} with parameter $\sigma > 0$. Let $\{\varphi(X_t, a_t)\}_{t=1}^{\infty}$ be an $\mathbb{R}^p$-valued stochastic process such that $\varphi(X_t, a_t)$ is F_{t-1}-measurable. Let $\lambda \in \mathbb{R}$ be a strictly positive constant. Then, for all $\eta \in (0, 1)$, it holds with probability at least $1 - \eta$ that, for all $t \in \mathbb{N}$,*

$$\|\sum_{\tau=1}^{t} \varphi(X_\tau, a_\tau)\epsilon_\tau\|_{\hat{\Sigma}_t^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{1}{\eta}\sqrt{\frac{\det(\hat{\Sigma}_t)}{\det(\lambda I_p)}}\right).$$

Moreover, if, for all $t \in [T]$, $\|\varphi(X_t, a_t)\|_2 \leq L$, *then*

$$\log\left(\frac{\det(\hat{\Sigma}_t)}{\det(\lambda I_p)}\right) \leq p \log\left(1 + \frac{tL^2}{\lambda p}\right).$$

Lemma 3.2.2 implies that $\|\hat{\gamma}_t - \gamma_{0,t}\|_{\hat{\Sigma}_{t-1}}$ is $\tilde{O}(\sqrt{p})$, where the dependence on p is determined by the dimensionality of the matrix $\hat{\Sigma}_t$. The sum over the time horizon of the squared norm of the context-action features, $\sum_{t=1}^{T} \|\varphi(X_t, a_t)\|^2_{\hat{\Sigma}_{t-1}^{-1}}$, is also shown to be $\tilde{O}(p)$, depending again on the context-action features dimensionality. This term appears under square root when bounding the cumulative regret, leading to an additional $\tilde{O}(\sqrt{p})$ term in the final bound. Our goal is to show that, in Setting 3.2.1, we can reduce the regret to depend on p^{res} rather than p, up to some term that becomes negligible for large enough T_0. In the following, we focus on the LinUCB analysis, but a similar reasoning could apply for other linear bandit algorithms based on least squares estimation.

3.3 ISD-linUCB algorithm

Assuming Setting 3.2.1, we propose an algorithm that is split into an offline phase for the estimation of $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ and β^{inv}, followed by an online bandit phase adapting to the non-stationarity.

The offline phase relies on the T_0 historical observations from Setting 3.2.1. To estimate $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$, Lazzaretto et al. [2025] estimate an orthonormal matrix U that jointly block diagonalizes the covariance matrices $(\text{Var}(\varphi(X_t, a_t)))_{t\in[-T_0]}$. U can be partitioned into two submatrices $U^{\text{inv}} \in \mathbb{R}^{p\times p^{\text{inv}}}$ and $U^{\text{res}} \in \mathbb{R}^{p\times p^{\text{res}}}$ whose columns form a basis for $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$, respectively. Then, the orthogonal projection matrices onto $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ are $\Pi^{\mathcal{S}^{\text{inv}}} = U^{\text{inv}}(U^{\text{inv}})^\top$ and $\Pi^{\mathcal{S}^{\text{res}}} = U^{\text{res}}(U^{\text{res}})^\top$. Under Assumption 1, for all $t \in [T]$ we can rewrite (3.2.1) as

$$R_t^{a_t} = \varphi(X_t, a_t)^\top \Pi^{\mathcal{S}^{\text{inv}}} \beta^{\text{inv}} + \varphi(X_t, a_t)^\top \Pi^{\mathcal{S}^{\text{res}}} \delta_t^{\text{res}} + \epsilon_t, \tag{3.3.5}$$

where $\text{Cov}(\Pi^{\mathcal{S}^{\text{inv}}}\varphi(X_t, a_t), \Pi^{\mathcal{S}^{\text{res}}}\varphi(X_t, a_t)) = 0$. The subspace decomposition therefore allows to estimate β^{inv} and δ_t^{res} separately. For any index set $\mathcal{T} \subseteq [-T_0] \cup [T]$, we define the sample covariance matrix

$$\hat{\Sigma}_{\mathcal{T}} := \sum_{\tau\in\mathcal{T}} \varphi(X_\tau, a_\tau)\varphi(X_\tau, a_\tau)^\top.$$

We make the following assumption to ensure that $\hat{\Sigma}_{[-T_0]}$, is strictly positive definite.

Assumption 2. *The policy used to collect the* T_0 *observations is such that* $\exists \lambda_0 > 0$ *such that* $\lambda_{\min}(\frac{1}{T_0}\hat{\Sigma}_{[-T_0]}) \geq \lambda_0$ *almost surely.*

This means in particular that the policy used in the collection of the offline data has explored the context-action feature space sufficiently well in all directions. This is also

required to be able to estimate the invariant subspace decomposition in the first place [see Lazzaretto et al., 2025].

We denote an estimate of U by $\hat{U} = [\hat{U}^{\text{inv}}, \hat{U}^{\text{res}}]$. Under Assumption 2, we can estimate the invariant component β^{inv} as the OLS solution in the invariant subspace, using the estimated $\hat{U}^{\text{inv}}$ and the offline observations, that is,

$$\hat{\beta}^{\text{inv}} := \hat{U}^{\text{inv}}(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1} \sum_{t\in[-T_0]} (\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t) R_t^{a_t}, \tag{3.3.6}$$

where $\tilde{\Sigma}^{\text{inv}}_{[-T_0]} := (\hat{U}^{\text{inv}})^\top \hat{\Sigma}_{[-T_0]} \hat{U}^{\text{inv}}$. We define a confidence set around $\hat{\beta}^{\text{inv}}$ by

$$\hat{\mathcal{C}}^{\beta} := \{\beta \in \hat{\mathcal{S}}^{\text{inv}} \mid \|\hat{\beta}^{\text{inv}} - \beta\|^2_{\hat{\Sigma}_{[-T_0]}} \leq \hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)\}, \tag{3.3.7}$$

where $\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)$ is defined in (3.C.15) in Appendix 3.C.3 such that the set contains $\hat{\Pi}^{\mathcal{S}^{\text{inv}}} \beta^{\text{inv}}$ with probability at least $1 - \eta$. As new observations become available, the estimate $\hat{\beta}^{\text{inv}}$ and the confidence set in in (3.3.7) may be recomputed online.

At round $t \in [T]$ of the online ISD-linUCB algorithm, we estimate the residual component as

$$\hat{\delta}^{\text{res}}_t := \hat{U}^{\text{res}}(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1} \sum_{\tau=1}^{t-1} (\hat{U}^{\text{res}})^\top \varphi_\tau (R_\tau^{a_\tau} - \varphi_\tau^\top \hat{\beta}^{\text{inv}})$$

where $\varphi_\tau := \varphi(X_\tau, a_\tau)$ and for all $t \in [T]$, $\tilde{\Sigma}^{\text{res}}_t := (\hat{U}^{\text{res}})^\top \hat{\Sigma}_t \hat{U}^{\text{res}}$ and $\hat{\Sigma}_t$ is defined as in (3.2.3). We then define a confidence set around $\hat{\delta}^{\text{res}}_t$ by

$$\hat{\mathcal{C}}^{\delta}_t := \{\delta \in \hat{\mathcal{S}}^{\text{res}} \mid \|\hat{\delta}^{\text{res}}_t - \delta\|^2_{\hat{\Sigma}_{t-1}} \leq \hat{\rho}^{\text{res}}_t(\eta, L, M)\},$$

where $\hat{\rho}^{\text{res}}_t(\eta, L, M)$ is defined in (3.C.20) in Appendix 3.C.3 such that the set contains $\hat{\Pi}^{\mathcal{S}^{\text{res}}} \delta^{\text{res}}_t$ with probability at least $1-\eta$, and choose an action based on $\hat{\mathcal{C}}^{\beta} \oplus \hat{\mathcal{C}}^{\delta}_t$. Assumption 1 ensures that there is no bias introduced due to omitting the context-action features in the residual subspace when estimating β^{inv} (and vice versa for δ^{res}_t) and therefore that $\hat{\mathcal{C}}^{\beta}$ and $\hat{\mathcal{C}}^{\delta}_t$ have the desired coverage.

The full ISD-linUCB algorithm for a single online step, with and without updating the invariant component, is provided in Algorithm 1. Since Setting 3.2.1 assumes $\gamma_{0,t}$ is fixed in $[T]$, ISD-linUCB acts in $\mathcal{S}^{\text{res}}$ as a standard LinUCB algorithm. When this assumption fails, Algorithm 1 can be modified to act in $\mathcal{S}^{\text{res}}$ as existing non-stationary algorithms (e.g., using a sliding window), to improve their performance in terms of dimensionality.

3.4 Regret analysis

To motivate our approach, in Section 3.4.1 we first analyze the regret of a simplified algorithm having oracle knowledge of the subspace decomposition $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$. We then discuss the complete regret analysis in Section 3.4.3.

Algorithm 1 ISD-linUCB (iteration at time t)

Input: $(\varphi(X_\tau, a_\tau), a_\tau, R_\tau^{a_\tau})_{\tau\in[t-1]\cup[-T_0]}$, X_t

Parameters: $\mathcal{A}$, λ, η, L, M, `recompute`

$\mathbf{X}_t \leftarrow [\varphi(X_1, a_1), \dots, \varphi(X_{t-1}, a_{t-1})]^\top$

$\mathbf{R}_t \leftarrow [R_1^{a_1}, \dots, R_{t-1}^{a_{t-1}}]^\top$

if `recompute` or $t = 1$ **then**

$\mathcal{T} \leftarrow [-T_0] \cup [t-1]$

$\overline{\mathbf{X}} \leftarrow [\varphi(X_{-T_0}, a_{-T_0}), \dots, \varphi(X_{t-1}, a_{t-1})]^\top$

compute $\hat{U}^{\mathrm{res}}$ and $\hat{U}^{\mathrm{inv}}$ using joint block diagonalization

$\overline{\mathbf{X}}^{\mathrm{inv}} \leftarrow \overline{\mathbf{X}}\hat{U}^{\mathrm{inv}}$

$\hat{\beta}^{\mathrm{inv}} \leftarrow \hat{U}^{\mathrm{inv}}\big((\overline{\mathbf{X}}^{\mathrm{inv}})^\top\overline{\mathbf{X}}^{\mathrm{inv}}\big)^{-1}(\overline{\mathbf{X}}^{\mathrm{inv}})^\top\overline{\mathbf{R}}$

$\hat{\mathcal{C}}^\beta \leftarrow \{\beta \in \hat{\mathcal{S}}^{\mathrm{inv}} \mid \|\hat{\beta}^{\mathrm{inv}} - \beta\|^2_{\hat{\Sigma}_\mathcal{T}} \le \hat{\rho}^{\mathrm{inv}}_{|\mathcal{T}|}(\eta, L, M)\}$

else

load $\hat{\beta}^{\mathrm{inv}}$, $\hat{U}^{\mathrm{res}}$ and $\hat{\mathcal{C}}^\beta$ from previous iteration

$\mathbf{R}_t^{\mathrm{res}} \leftarrow \mathbf{R}_t - \mathbf{X}_t\hat{\beta}^{\mathrm{inv}}$

$\mathbf{X}_t^{\mathrm{res}} \leftarrow \mathbf{X}_t\hat{U}^{\mathrm{res}}$

$\hat{\delta}_t^{\mathrm{res}} \leftarrow \hat{U}^{\mathrm{res}}\big(\lambda I_{p^{\mathrm{res}}} + (\mathbf{X}_t^{\mathrm{res}})^\top\mathbf{X}_t^{\mathrm{res}}\big)^{-1}(\mathbf{X}_t^{\mathrm{res}})^\top\mathbf{R}_t^{\mathrm{res}}$

$\hat{\mathcal{C}}_t^\delta \leftarrow \{\delta \in \hat{\mathcal{S}}^{\mathrm{res}} \mid \|\hat{\delta}_t^{\mathrm{res}} - \delta\|^2_{\hat{\Sigma}_{t-1}} \le \hat{\rho}_t^{\mathrm{res}}(\eta, L, M)\}$

$a_t \leftarrow \arg\max_{a\in\mathcal{A}} \max_{\gamma\in\hat{\mathcal{C}}^\beta\oplus\hat{\mathcal{C}}_t^\delta} \varphi(X_t, a)^\top\gamma$

return a_t

3.4.1 Motivation: oracle ISD-linUCB

We study Algorithm 1 assuming the matrix $U = [U^{\mathrm{inv}}, U^{\mathrm{res}}]$ is known. For simplicity, we keep the notation introduced in the previous section to denote the same quantities with known U. We define $(\bar{\beta}_t, \bar{\delta}_t) \coloneqq \arg\max_{\beta\in\hat{\mathcal{C}}^\beta, \delta\in\hat{\mathcal{C}}^\delta} \varphi(X_t, a_t)^\top(\beta + \delta)$ the parameters at which the upper confidence bound for the chosen action at time t is achieved.

Using the decomposition of the reward from (3.3.5) and the standard LinUCB analysis [see, for example, Lattimore and Szepesvári, 2020, Section 19.3], we obtain that at time $t \in [T]$, with high probability, the instantaneous regret is upper bounded by

$$\mathrm{reg}_t \le \underbrace{\varphi(X_t, a_t)^\top U^{\mathrm{inv}}(U^{\mathrm{inv}})^\top(\bar{\beta}_t - \beta^{\mathrm{inv}})}_{\mathrm{reg}_t^{\mathrm{inv}}} + \underbrace{\varphi(X_t, a_t)^\top U^{\mathrm{res}}(U^{\mathrm{res}})^\top(\bar{\delta}_t - \delta_t^{\mathrm{res}})}_{\mathrm{reg}_t^{\mathrm{res}}}. \tag{3.4.8}$$

Intuitively, for larger T_0 the confidence set $\hat{\mathcal{C}}^\beta$ shrinks and $\bar{\beta}_t - \beta^{\mathrm{inv}}$ gets close to zero. Therefore, for $T_0 \gg T$, $\mathrm{reg}_t^{\mathrm{inv}}$ becomes negligible with respect to $\mathrm{reg}_t^{\mathrm{res}}$. Removing the uncertainty on the invariant component implies that, in the regret analysis, the dependence on the dimension p of the bandit parameter is reduced to the dimension

p^{res} of the residual parameter. We first study the regret assuming to also have oracle knowledge of the invariant component β^{inv} (Section 3.4.1.1). Then, we include in the regret analysis the estimation of the invariant component (Section 3.4.1.2).

3.4.1.1 Regret analysis for oracle $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$, β^{inv}

If we know β^{inv}, then $\hat{\mathcal{C}}^{\beta} = \{\beta^{\text{inv}}\}$ and $\text{reg}_t^{\text{inv}} = 0$. Therefore, Algorithm 1 only performs the exploration in the residual subspace. As shown in Theorem 3.4.1, its regret only depends on the uncertainty on $\hat{\delta}_t^{\text{res}}$, which lies on a p^{res}-dimensional space.

Theorem 3.4.1. *Consider Setting 3.2.1, assume $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ and β^{inv} are known and consider the oracle version of Algorithm 1 that uses β^{inv}, U^{inv} and U^{res} instead of $\hat{\beta}^{\text{inv}}$, $\hat{U}^{\text{inv}}$ and $\hat{U}^{\text{res}}$. Then, the regret of this oracle algorithm over a time horizon T with $\eta = 1/T$ is $\tilde{O}(p^{\text{res}}\sqrt{T})$.*

The proof of Theorem 3.4.1 uses that, under oracle knowledge, $\text{reg}_t = \text{reg}_t^{\text{res}}$, where the dimension of $(U^{\text{res}})^\top \varphi(X_t, a_t)$ and of $(U^{\text{res}})^\top(\bar{\delta}_t - \delta_t^{\text{res}})$ is p^{res}. It then follows similar steps as the regret analysis for the standard LinUCB algorithm introduced in Section 3.2.1. A detailed proof is provided in Appendix 3.C.1.

3.4.1.2 Estimating the invariant component from offline data

When also considering the uncertainty on the invariant component estimated using T_0 observations, we can upper bound the regret of Algorithm 1 as follows.

Theorem 3.4.2. *Consider Setting 3.2.1, assume Assumption 2 holds and that $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ is known, and consider the oracle version of Algorithm 1 that uses U^{inv} and U^{res} instead of $\hat{U}^{\text{inv}}$ and $\hat{U}^{\text{res}}$. Then, the regret of this oracle algorithm over a time horizon T with $\eta = 1/T$ is $\tilde{O}\left(\sqrt{T}\left(p^{\text{res}} + \left(\sqrt{\frac{p^{\text{inv}}}{\lambda_0}} + \frac{1}{\lambda_0}\right)\sqrt{\frac{T}{T_0}}\right)\right)$.*

Theorem 3.4.2 implies that, whenever T_0 is sufficiently larger than T, we have an advantage in using offline data to estimate invariant information in the regret function over only relying on online data. Indeed, the larger T_0 is, the closer we get to the oracle bound shown in Section 3.4.1.1 of $\tilde{O}(p^{\text{res}}\sqrt{T})$.

Proof sketch of Theorem 3.4.2. To bound $\text{reg}_t^{\text{inv}}$ we choose $\hat{\rho}_{T_0}^{\text{inv}}(\eta, L, M)$ such that $\hat{\mathcal{C}}^{\beta}$ contains β^{inv} with probability at least $1 - \eta$. By Assumption 2, we have that $\hat{\Sigma}_{[-T_0]} \succeq \lambda_0 T_0 I_p$ (with $\succeq$ denoting the Loewner order between two matrices) which implies that $\tilde{\Sigma}^{\text{inv}}_{[-T_0]} \succeq \lambda_0 T_0 I_{p^{\text{inv}}}$. It follows that, for all $\beta \in \hat{\mathcal{C}}^{\beta}$,

$$\begin{aligned}
\|\hat{\beta}^{\text{inv}} - \beta\|^2_{\hat{\Sigma}_{[-T_0]}} &= \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{\frac{1}{2}}(U^{\text{inv}})^\top(\hat{\beta}^{\text{inv}} - \beta)\|_2^2 \\
&\geq \|\sqrt{\lambda_0 T_0}(U^{\text{inv}})^\top(\hat{\beta}^{\text{inv}} - \beta)\|_2^2 \\
&= \lambda_0 T_0 \|(U^{\text{inv}})^\top(\hat{\beta}^{\text{inv}} - \beta)\|_2^2. \qquad (3.4.9)
\end{aligned}$$

Finally, this implies that

$$\|(U^{\text{inv}})^\top(\hat{\beta}^{\text{inv}} - \beta)\|_2^2 \leq \frac{1}{\lambda_0 T_0}\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M).$$

Therefore, with probability at least $1-\eta$,

$$\text{reg}_t^{\text{inv}} \leq 2\|(U^{\text{inv}})^\top\varphi(X_t, a_t)\|_2\sqrt{\frac{\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)}{\lambda_0 T_0}}.$$

By Lemma 3.2.2 and Assumption 2, we can define $\sqrt{\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)}$ to be $\tilde{O}(\sqrt{p^{\text{inv}}} + \sqrt{\frac{1}{\lambda_0}})$. The analysis of $\text{reg}_t^{\text{res}}$ is similar to Theorem 3.4.1, with the addition of a term due to the introduction of $\hat{\beta}^{\text{inv}}$ which is $\tilde{O}(\sqrt{T\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)/(\lambda_0 T_0)})$. This implies, summing over $t \in [T]$, that the cumulative regret is $\tilde{O}\left(\sqrt{T}\left(p^{\text{res}} + \sqrt{\frac{p^{\text{inv}}T}{\lambda_0 T_0}} + \frac{1}{\lambda_0}\sqrt{\frac{T}{T_0}}\right)\right)$. For the full proof, see Appendix 3.C.2. □

3.4.2 Accounting for errors in the subspace decomposition

When estimating $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ using the available T_0 observations, we need to consider two sources of errors. Lazzaretto et al. [2025] propose to estimate the subspaces via joint block diagonalization, in our case of the sample covariance matrices of the context-action features through time. The procedure first finds an irreducible joint decomposition of the features into orthogonal lower dimensional subspaces such that the projections of the features onto the subspaces are pairwise uncorrelated. Then, it groups together subspaces in which the linear relationship between the reward and the features is invariant, defining $\mathcal{S}^{\text{inv}}$, and the remaining ones in which the linear relationship is time-varying, defining $\mathcal{S}^{\text{res}}$.

The first source of error can occur when some directions in $\mathbb{R}^p$ are wrongly identified as invariant or time-varying. Labeling directions as time-varying when they are invariant is less problematic, as it leads to a suboptimal but not incorrect algorithm: in this case, p^{inv} is underestimated, preventing the decomposition from achieving its maximum benefit. Overestimating p^{inv}, i.e., assuming invariance of time-varying directions, implies instead that in such directions the algorithm uses the pooled estimated parameter from the T_0 historical observations to predict the reward, while the true parameter may have changed. Then, intuitively, ISD-linUCB suffers an additional loss scaling with how different the true parameter is in $[T]$ compared to its average in $[-T_0]$ in the wrongly labeled subspace.

The second source of error is the estimation error due to finite sample approximation. To estimate the matrix U we first estimate $(\text{Var}(\varphi(X_t, a_t)))_{t\in[-T_0]}$; in practice, it is sufficient to split $[-T_0]$ into m windows and compute $\hat{\Sigma}_{\mathcal{M}_i}$, defined as in (3.3) with $\mathcal{M}_i := \{-\frac{i}{m}T_0, \dots, -\frac{i-1}{m}T_0 - 1\}$, $i \in \{1, \dots, m\}$. Then, $\hat{U}$ is an estimated joint block diagonalizer for $(\hat{\Sigma}_{\mathcal{M}_i})_{i\in\{1,\dots,m\}}$. The matrix $\hat{U}$ then enters in the estimation of both β^{inv} and δ_t^{res}. Let $\hat{\gamma}^{\text{ISD}} = \hat{\beta}^{\text{inv}} + \hat{\delta}_t^{\text{res}}$ be an estimator for $\gamma_{0,t}$ obtained using the ISD

framework. For all $t \in [T]$, we can express the estimation error as

$$\begin{aligned}\hat{\gamma}_t^{\mathrm{ISD}} - \gamma_{0,t} &= \hat{\beta}^{\mathrm{inv}} - \beta^{\mathrm{inv}} + \hat{\delta}_t^{\mathrm{res}} - \delta_t^{\mathrm{res}} \\ &= (\hat{\beta}^{\mathrm{inv}} - \hat{\Pi}^{\mathcal{S}^{\mathrm{inv}}}\beta^{\mathrm{inv}}) + (\hat{\Pi}^{\mathcal{S}^{\mathrm{inv}}} - \Pi^{\mathcal{S}^{\mathrm{inv}}})\beta^{\mathrm{inv}} \\ &\quad + (\hat{\delta}_t^{\mathrm{res}} - \hat{\Pi}^{\mathcal{S}^{\mathrm{res}}}\delta_t^{\mathrm{res}}) + (\hat{\Pi}^{\mathcal{S}^{\mathrm{res}}} - \Pi^{\mathcal{S}^{\mathrm{res}}})\delta_t^{\mathrm{res}}. \end{aligned} \tag{3.4.10}$$

The first and third term correspond to the estimation error for the invariant and residual component on the estimated subspaces. The second and fourth term only depend on the subspaces estimation error. When bounding the regret of Algorithm 1, we assume that the subspace decomposition is correctly estimated and only take into account the finite sample error in estimating $(\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}})$, starting from the decomposition in (3.4.10).

3.4.3 Complete regret analysis

To analyze the terms in (3.4.10) we need to quantify the subspace estimation error. A common way to do this is to consider the distance between the true and the estimated subspace, which can be described through the notion of principal angles. Let $U^{\mathcal{S}} \in \mathbb{R}^{p\times k}$ be the submatrix of U whose columns span the subspace $\mathcal{S}$ of $\mathbb{R}^p$, and let $\hat{U}^{\mathcal{S}}$ be its estimate obtained through $\hat{U}$, whose columns span $\hat{\mathcal{S}}$. The principal angles between $\mathcal{S}$ and $\hat{\mathcal{S}}$ are characterized as the inverse cosine of the nonzero singular values of the matrix $(\hat{U}^{\mathcal{S}})^\top U^{\mathcal{S}}$ (intuitively, the cosine of the angle between two vectors is given by their normalized inner product). Let $\Theta^{\mathcal{S}}$ denote the diagonal matrix with the principal angles between $\mathcal{S}$ and $\hat{\mathcal{S}}$ on the diagonal. Then, we measure the distance between $\mathcal{S}$ and $\hat{\mathcal{S}}$ by $\|\sin\Theta^{\mathcal{S}}\|_{\mathrm{op}}$, i.e., the largest principal angle between $\mathcal{S}$ and $\hat{\mathcal{S}}$ [see for example Theorem 4.5 by Stewart and Sun, 1990], which also equals $\|\hat{\Pi}^{\mathcal{S}} - \Pi^{\mathcal{S}}\|_{\mathrm{op}}$ [Stewart and Sun, 1990, Corollary 4.6]. We denote by $\Delta\Pi$ such distance for the invariant and residual subspaces (this must be the same for both subspaces, see Stewart and Sun [1990, Preliminaries—Corollary 5.4]), and introduce the following assumption for the subspace decomposition error.

Assumption 3. *For all $\eta \in (0,1)$, it holds with probability at least $1-\eta$ that $\Delta\Pi$ is $O(\sqrt{\log(p/\eta)/T_0})$.*

Assumption 3 can be justified using the Davis-Kahan theorem [see, for example, Yu et al., 2015] and by concentration results for sample covariance matrices [see, for example, Vershynin, 2018, Sections 5.4 and 5.6]. In more detail, the Davis-Kahan theorem provides an upper bound for $\|\sin\Theta^{\mathcal{S}}\|_{\mathrm{op}}$ in terms of estimation error on a matrix $M \in \mathbb{R}^{p\times p}$ such that the columns of U are eigenvectors for M. In our case, U jointly block diagonalizes $(\mathrm{Var}(\varphi(X_t, a_t)))_{t\in[-T_0]}$, and for all $t \in [-T_0]$ the span of the columns of U^{inv} and U^{res} coincide with the union of subsets of eigenspaces of $\mathrm{Var}(\varphi(X_t, a_t))$. We estimate $\hat{U}$ through the matrices $(\hat{\Sigma}_{\mathcal{M}_i})_{i=1,\dots,m}$, each computed using T_0/m observations. As $\mathrm{Var}(\varphi(X_t, a_t))$ is not constant with t, we cannot use exact concentration results; with Assumption 3, we assume that $\Delta\Pi$ scales as the finite-sample error for estimating a fixed p-dimensional covariance matrix from T_0 observations (we verify it empirically in Appendix 3.B.1).

The sum of the second and fourth term in (3.4.10) equals $\Delta\Pi\gamma_{0,t}$. Under Assumption 3, this quantity is, with probability at least $1-\eta$, $O(\sqrt{\log(p/\eta)/T_0})$. Moreover, under Assumption 3, we can show the following high probability bound for the estimation error on $\hat{\beta}^{\text{inv}}$ in the estimated invariant subspace.

Lemma 3.4.3. *Consider Setting 3.2.1 and assume Assumptions 2 and 3 hold. Then, for all $\eta \in (0,1)$ it holds with probability at least $1-\eta$ that $\|\hat{\beta}^{\text{inv}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}}\|_{\hat{\Sigma}_{[-T_0]}}$ is*

$$O\left(\sqrt{p^{\text{inv}}\log(\tfrac{1}{p^{\text{inv}}\lambda_0}) + \log(\tfrac{1}{\eta})}\right) + O\left(\sqrt{p^{\text{inv}}\log(\tfrac{p}{\eta})}\right) + O\left(\sqrt{\tfrac{1}{\lambda_0}\log(\tfrac{p}{\eta})}\right).$$

Proof sketch. We can decompose the error into three terms as follows

$$\begin{aligned}\|\hat{\beta}^{\text{inv}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}}\|_{\hat{\Sigma}_{[-T_0]}} \leq & \|(\hat{U}^{\text{inv}})^\top \hat{\Sigma}_{[-T_0]}(\Pi^{\mathcal{S}^{\text{inv}}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}})\beta^{\text{inv}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\ & + \|(\hat{U}^{\text{inv}})^\top \sum_{t=-T_0}^{-1} \varphi(X_t,a_t)\varphi(X_t,a_t)^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\ & + \|(\hat{U}^{\text{inv}})^\top \sum_{t=-T_0}^{-1} \varphi(X_t,a_t)\epsilon_t\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}}.\end{aligned}$$

The last term can be analyzed as in the oracle case (Theorem 3.4.2). The first two terms are introduced due to the misalignment between the true and the estimated subspaces. Under Assumption 3, for all $\eta \in (0,1)$, with probability at least $1-\eta$, the first is $O(\sqrt{p^{\text{inv}}\log(\frac{p}{\eta})})$ and the second is $O(\sqrt{\frac{1}{\lambda_0}\log(\frac{p}{\eta})})$. For the full proof, see Appendix 3.C.3. □

We can further obtain an upper bound for $\|\hat{\beta}^{\text{inv}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}}\|_2$ by multiplying all the factors in Lemma 3.4.3 by $\frac{1}{\sqrt{\lambda_0 T_0}}$ (see (3.4.9)).

For the residual component, we can similarly bound $\|\hat{\delta}_t^{\text{res}} - \hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}}$ using Assumption 3 as follows.

Lemma 3.4.4. *Consider Setting 3.2.1 and assume Assumptions 2 and 3 hold. Then, for all $t \in [T]$ and for all $\eta \in (0,1)$ it holds with probability at least $1-\eta$ that $\|\hat{\delta}_t^{\text{res}} - \hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}}$ is*

$$O\left(\sqrt{p^{\text{res}}\log(1+\tfrac{t}{p^{\text{res}}}) + \log(\tfrac{1}{\eta})}\right) + O\left(\sqrt{p^{\text{res}}t}\left(\sqrt{\tfrac{\log(p/\eta)}{T_0}} + \|\hat{\beta}^{\text{inv}} - \beta^{\text{inv}}\|_2\right)\right).$$

The first term is the same that appears in the case of oracle subspaces. The second term is instead due to the error in the estimation of the subspaces.

Proof sketch. We can upper bound $\|\hat{\delta}_t^{\text{res}} - \hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}}$ by the sum of the following components

$$\begin{aligned}\|\hat{\delta}_t^{\text{res}} - \hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}} \leq &\|\sum_{\tau=1}^{t-1}(\hat{U}^{\text{res}})^\top\varphi_\tau\varphi_\tau^\top(\Pi^{\mathcal{S}^{\text{res}}} - \hat{\Pi}^{\mathcal{S}^{\text{res}}})\delta_t^{\text{res}}\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\\ &+\|\sum_{\tau=1}^{t-1}(\hat{U}^{\text{res}})^\top\varphi_\tau\varphi_\tau^\top(\beta^{\text{inv}} - \hat{\beta}^{\text{inv}})\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\\ &+\|\sum_{\tau=1}^{t-1}(\hat{U}^{\text{res}})^\top\varphi_\tau\epsilon_\tau\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} + \sqrt{\lambda}\|\delta_t^{\text{res}}\|_2\end{aligned}$$

where $\varphi_\tau := \varphi(X_\tau, a_\tau)$. The analysis of the last two terms is the same as in Theorem 3.4.1 and results in such terms being, for all $\eta \in (0,1)$, with probability at least $1-\eta$, $O\left(\sqrt{\log(\frac{1}{\eta}) + p^{\text{res}}\log(\frac{t}{p^{\text{res}}})}\right)$. The first two terms are introduced due to the subspace estimation error. Under Assumption 3, the first term is, for all $\eta \in (0,1)$, with probability at least $1-\eta$, $O\left(\sqrt{\frac{p^{\text{res}}t}{T_0}\log(\frac{p}{\eta})}\right)$ and the second is $O\left(\sqrt{p^{\text{res}}t}\|\hat{\beta}^{\text{inv}} - \beta^{\text{inv}}\|_2\right)$. For the full proof, see Appendix 3.C.3. □

Lemma 3.4.3 and Lemma 3.4.4 imply the following regret bound for the ISD-linUCB algorithm.

Theorem 3.4.5. *Consider Setting 3.2.1 and assume Assumptions 2 and 3 hold. If $\hat{\mathcal{C}}^\beta$ is defined to contain $\hat{\Pi}^{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}}$ with probability at least $1-4/T_0$ and $\hat{\mathcal{C}}_t^\delta$ is defined to contain $\hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta_t^{\text{res}}$ with probability at least $1-(\frac{2}{T}+\frac{4}{T_0})$, then, with probability at least $1-(\frac{2}{T}+\frac{9}{T_0})$ the regret of ISD-linUCB is*

$$\tilde{O}\left(\sqrt{T}\left(p^{\text{res}} + p^{\text{res}}\sqrt{\tfrac{T}{\lambda_0 T_0}}\left(\sqrt{p^{\text{inv}}} + \sqrt{\max\{\lambda_0, \tfrac{1}{\lambda_0}\}}\right)\right)\right).$$

Theorem 3.4.5 shows in particular that, for T_0 sufficiently larger than T (e.g., $T_0 = \Omega(T^{1+\varepsilon})$ for some $\epsilon > 0$), the regret bound for ISD-linUCB is dominated by $p^{\text{res}}\sqrt{T}$.

3.5 Simulation experiments

To support our theoretical results, we present several simulation experiments. We first apply the ISD-linUCB algorithm with oracle subspace knowledge in Section 3.5.1, and show that the regret indeed scales as $p^{\text{res}}\sqrt{T}$, supporting the result that if T_0 grows faster than T, the regret indeed scales as $p^{\text{res}}\sqrt{T}$. We then include the estimation of the subspace decomposition in the algorithm (Section 3.5.2), and show that for increasing T_0 its performance gets closer to the one of ISD-linUCB with oracle knowledge on subspaces. In all simulations, we provide a comparison with the standard LinUCB algorithm since, in Setting 3.2.1 we assume $\gamma_{0,t}$ to be fixed in [T] (in this setting, the performance of other non-stationary algorithms is comparable to the one of LinUCB, see Appendix 3.B.2). The

code for the presented experiments is available at https://github.com/mlazzaretto/ISD-linUCB.git.

3.5.1 Oracle subspace decomposition

We consider a setting with $T_0 = 2000$, $T = 100$ and $|\mathcal{A}| = 5$ and with known $U^{\text{inv}}, U^{\text{res}}$. The matrix U is sampled as a random orthonormal matrix, and we generate the covariance matrix of context-action features as $U\tilde{V}_t U^\top$, where $\tilde{V}_t$ is a block diagonal matrix with two blocks of dimensions $p^{\text{inv}}, p^{\text{res}}$. The entries of $(U^{\text{inv}})^\top \beta^{\text{inv}}$ are sampled uniformly in $(0.5, 1.5)$. The same is done for the initial values of $(U^{\text{res}})^\top \delta_t^{\text{res}}$, to which for $t \in [-T_0]$ we add $-1.5(t/T0)\sin^2(0.25it/T0 + i)$, with $i \in \{1, \dots, p^{\text{res}}\}$ being the entry index. The data in $[-T_0]$, used to estimate β^{inv}, is the output of a bandit algorithm using a policy that chooses the actions uniformly at random. For $t \in [T]$, the entries of $(U^{\text{res}})^\top \delta_t^{\text{res}}$ are sampled uniformly in $(0.5, 1.5)$. We set $\lambda = 0.1$ and we use $\eta = 1/T$. We then run two experiments (in both, $\hat{\beta}^{\text{inv}}$ is not updated online).

First, for a fixed dimension $p = 10$ of the context-action features we consider values of $p^{\text{res}} \in \{2, 4, 6, 8\}$. Figure 3.5.2 shows that the regret of ISD-linUCB (with oracle subspace information) grows sublinearly in T (left) and (approximately) linearly in p^{res} (right), empirically supporting the result obtained in Theorems 3.4.2. To further support these results, we run a second experiment where we consider context-action features of dimension p varying between 3 and 10, while keeping the dimension of $\mathcal{S}^{\text{res}}$ fixed to $p^{\text{res}} = 2$. The results are shown in Figure 3.5.3, which compares the cumulative regret to the one of the standard LinUCB algorithm. While the latter increases linearly with p, the regret of ISD-linUCB remains approximately constant as p^{res} is fixed.

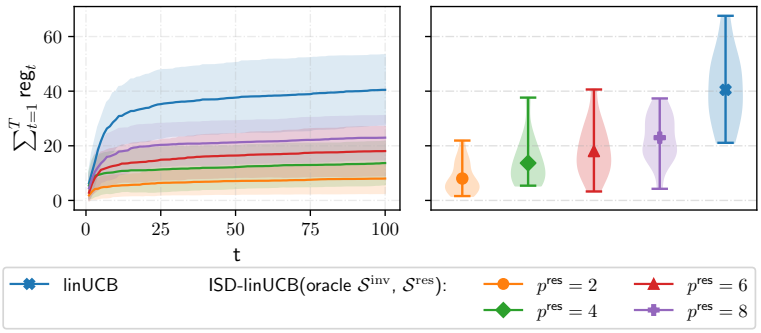


Figure 3.5.2: Regret of ISD-linUCB with oracle $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ over $T = 100$ rounds for $p^{\text{res}} \in \{2, 4, 6, 8\}$. For each p^{res} the experiment is repeated 20 times. The left plot shows the average performance and the standard deviation over the 20 repetitions, the right plot shows the distribution of the regret over the 20 repetitions.

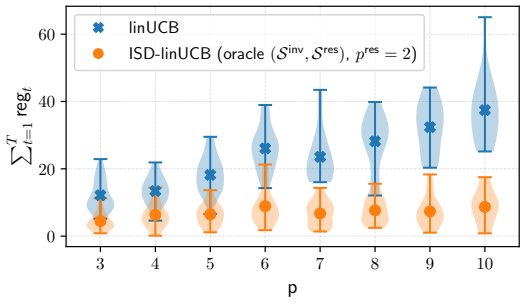


Figure 3.5.3: Cumulative regret of standard LinUCB and ISD-linUCB with oracle $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ for $T = 100$ and increasing values of context-action feature dimension p. For ISD-linUCB, the invariant component β^{inv} is estimated using $T_0 = 2000$ observations. p^{inv} varies from 3 to 10, while the p^{res} is fixed to 2. For each p the experiment is repeated 20 times.

3.5.2 Estimated subspace decomposition

We consider the same data-generating process as in Section 3.5.1, but now include the subspace estimation using the T_0 observations. We set $T = 500$ and consider $T_0 \in \{1000, 3500, 8000\}$. Moreover, we fix the dimensions of the subspaces to $p^{\text{inv}} = 7$ and $p^{\text{res}} = 3$. Figure 3.5.4 shows the cumulative regret for increasing T_0 for the standard LinUCB algorithm (unaffected by T_0) and for the algorithms studied in Sections 3.4.1.1, 3.4.1.2 and 3.4.2, supporting the presented theoretical results. Indeed, for increasing T_0, the regret of the ISD-linUCB algorithm estimating $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ and β^{inv} using the T_0 observations gets closer to the one of an algorithm with oracle knowledge. Moreover, the regret of the algorithm using the estimated ISD is lower than the standard LinUCB one for all considered values of T_0.

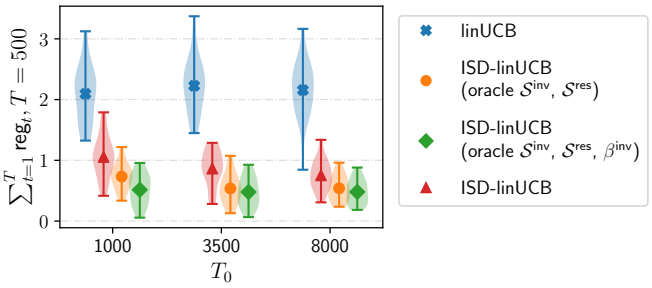


Figure 3.5.4: Cumulative regret for $T_0 \in \{1000, 3500, 8000\}$ (20 repetitions) for ISD-linUCB, in comparison with the same algorithm having oracle information and with the standard linUCB algorithm (unaffected by T_0). For increasing T_0, the regret of ISD-linUCB gets closer to the one of the oracle version.

3.6 Conclusions

Exploiting invariances in non-stationary environments allows linear contextual bandit algorithms to adapt more efficiently to changes in the reward model. We propose ISD-linUCB, a novel algorithm that learns such invariances from offline bandit data by decomposing the context-action feature space into two orthogonal subspaces: in one of the two subspaces, which has dimension p^{inv}, the relationship between reward and context-action features remains stationary. When sufficient offline data are available, the regret of ISD-linUCB scales with $(p - p^{\text{inv}})$ rather than p, leading to a substantial performance gain in environments that are subject to rapid changes.

Supplement to 'Invariance-based dynamic regret minimization'

3.A. Further related works

Linear contextual bandits The stochastic linear bandit problem is extensively studied in the literature. It is commonly assumed [see, for example, Lattimore and Szepesvári, 2020] that there exists an unknown linear parameter of dimension p that is shared among all context-action features: this allows in particular to deal with arbitrarily large action spaces, and to obtain regret bounds that do not depend on the cardinality K of the action space. Two standard algorithms proposed in the literature to solve this problem are LinUCB (linear upper confidence bound), and LinTS (linear Thompson sampling). In UCB-based algorithms, at all rounds the agent chooses the action maximizing an upper confidence bound on the estimated regret. Li et al. [2010] consider finite action spaces and build such confidence bound directly on the estimated regret for each possible action. Chu et al. [2011] show a regret bound for this version of LinUCB of $\tilde{O}(\sqrt{pKT})$. Abbasi-yadkori et al. [2011] propose a more general UCB algorithm that allows for infinitely large action spaces, by constructing a confidence set for the estimated linear parameter rather than for the reward, and show a regret bound of $\tilde{O}(p\sqrt{T})$. Linear bandits based on Thompson sampling [Agrawal and Goyal, 2013] at all rounds sample from the rewards posterior distribution for all possible arms, and then choose the action maximizing the sampled reward. LinTS is shown to achieve a regret of $\tilde{O}(p^{\frac{3}{2}}\sqrt{T})$.

In these algorithms, no assumptions (besides boundedness of the context-action features and of the bandit parameter) are made on the nature of the observed contexts. Papini et al. [2021] review different diversity conditions on the context-action features space that allow to improve the regret bounds, leading to either logarithmic or constant regret in T. One example is when the variance of the optimal context-action features is a positive definite matrix, meaning that all directions in $\mathbb{R}^p$ are potentially optimal for some context.

Linear latent contextual bandits To deal with heterogeneity in the reward function, Kausik et al. [2025] consider a latent linear contextual bandits framework, where a latent linear parameter $\theta \in \mathbb{R}^{p_l}$, with $p_l < p$, is shared across all heterogeneous observations. They assume the reward is of the form $R_t^{a_t} = \varphi(X_t, a_t)^\top \gamma_0 + \epsilon_t$, with $\gamma_0 = U\theta$ and $U \in \mathbb{R}^{p \times p_l}$ is a unitary matrix. They propose to learn the matrix U from offline data, and use this to obtain a regret bound that depends on the dimension p_l of the latent parameter rather than the context dimension p.

Bilaj et al. [2024] assume that the agent sequentially interacts with different tasks such that, for each task, the parameter γ_0 is independently drawn from the same distribution. They assume there exists an orthogonal projection matrix P with rank p_l such that $\mathbb{E}_{\gamma_0}[\|(I-P)(\gamma_0 - \mathbb{E}[\gamma_0])\|^2] \ll \mathbb{E}_{\gamma_0}[\|P(\gamma_0 - \mathbb{E}[\gamma_0])\|^2] \leq \mathbb{E}_{\gamma_0}[\|\gamma_0 - \mathbb{E}[\gamma_0]\|^2]$, namely the variance of the distribution of γ_0 is very low along $p - p_l$ orthogonal directions. They show a regret bound that depends on p_l rather than p.

Qin et al. [2022] consider a multi-task sequential linear bandit model where the agent plays a sequence of S bandit tasks (drawn from m different environments) for N rounds each. They model the reward as $R_t^{a_t} = \varphi(X_t, a_t)^\top \gamma_{s(t)} + \epsilon_t$ and assume that there exists a shared latent parameter $\theta \in \mathbb{R}^{p_l}$ such that for all $s \in [S]$ there exists a matrix $U_s \in \mathbb{R}^{p \times p_l}$ such that $\gamma_s = U_s \theta$. They propose an algorithm that sequentially learns the latent representation and achieves a regret of $\tilde{O}(pp_l\sqrt{mSN} + p_l S\sqrt{N})$.

Trella et al. [2025] consider a non-stationary reward model governed by an underlying latent variable evolving as an AR process of order ℓ. They show a regret bound of $\tilde{O}(\ell\sqrt{pT})$.

Non-stationary linear bandits Cheung et al. [2019] and Russac et al. [2019] consider the linear bandit problem under non-stationarity of the environment, and in particular allow for the linear bandit parameter to change through time, so that the reward model is of the form $R_t^a = \varphi(X_t, a)^\top \gamma_{0,t} + \epsilon_t$. They define the variation budget B_T as a constant such that $\sum_{t=1}^{T-1} \|\gamma_{0,t+1} - \gamma_{0,t}\|_2 \leq B_T$. Cheung et al. [2019] propose an algorithm based on a sliding window regularized least squares estimator, Russac et al. [2019] propose instead to weight past observations by a discounting factor. Both achieve a regret bound of $\tilde{O}(p^{\frac{7}{8}} T^{\frac{3}{4}} B_T^{\frac{1}{4}})$. As shown by Zhao et al. [2020], this same regret is achieved by periodically restarting the standard LinUCB algorithm, where the optimal number of rounds to be played before each restart depends on B_T. This means that the excess regret compared to the stationary case is due to applying the standard LinUCB algorithm assuming that the linear parameter is fixed, and incurring an additional loss that depends on how much the parameter is changing in the time period between two restarts. Cheung et al. [2019] also show a lower bound for this setting of $\Omega(p^{\frac{2}{3}} T^{\frac{2}{3}} B_T^{\frac{1}{3}})$.

3.B. Additional experiments

3.B.1 Projection error

The key role of Assumption 3 is to ensure that the projection error on $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ decreases with the sample size T_0 used to estimate the matrix U (and thus the projection matrices onto the subspaces). Within the same experiment described in Section 3.5.2, we evaluate such error, i.e., $\|\hat{\Pi}^{\mathcal{S}^{\text{inv}}} - \Pi^{\mathcal{S}^{\text{inv}}}\|_{\text{op}}$, and show (Figure 3.B.1) that it indeed decreases approximately as $\sqrt{T_0}$, as assumed in Assumption 3.

3.B.2 Other non-stationary algorithms

Within the same setting of the second experiment described in Section 3.5.1, we add an additional comparison with two non-stationary algorithms: one using a sliding window (SW-linUCB) by Cheung et al. [2019] and one using a discounting factor (D-linUCB) by Russac et al. [2019]. In this setting, the linear parameter $\gamma_{0,t}$ is fixed in the time horizon $[T]$, so the dimension of the window for SW-linUCB is set to T and the discounting factor for D-linUCB is set to 0.999, and Figure 3.B.2 shows that these two algorithms

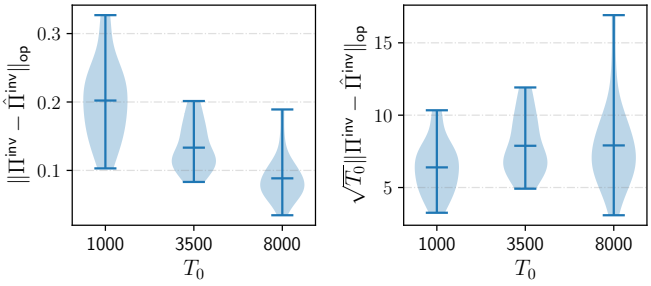


Figure 3.B.1: Projection error for $T_0 \in \{1000, 3500, 8000\}$. The plot on the left shows the same values multiplied by $\sqrt{T_0}$, confirming our assumption.

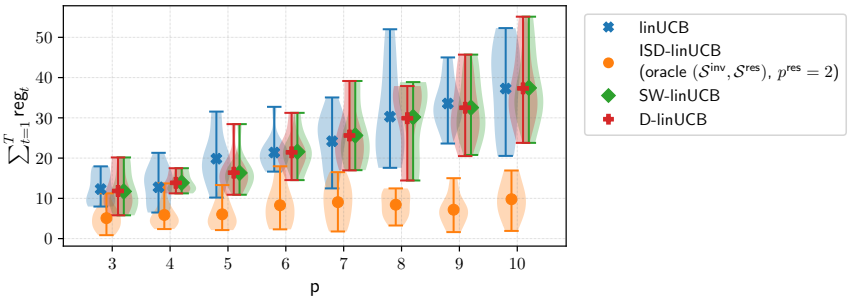


Figure 3.B.2: Cumulative regret of standard LinUCB, SW-linUCB, D-linUCB and ISD-linUCB with oracle $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ for $T = 100$ and increasing values of context-action feature dimension p. For ISD-linUCB, the invariant component β^{inv} is estimated using $T_0 = 2000$ observations. p^{inv} varies from 3 to 10, while the p^{res} is fixed to 2.

indeed have a comparable performance to LinUCB. In particular, we keep $\gamma_{0,t}$ fixed because our focus is on the reduction of the dimensionality (for the non-stationary part of the problem). When this assumption fails, we could still integrate the ISD framework within existing non-stationary algorithms to improve their performance in terms of dimensionality, as we do in the presented experiments with LinUCB.

The same happens for the simulation experiments presented in Section 3.5.2, as shown in Figure 3.B.3.

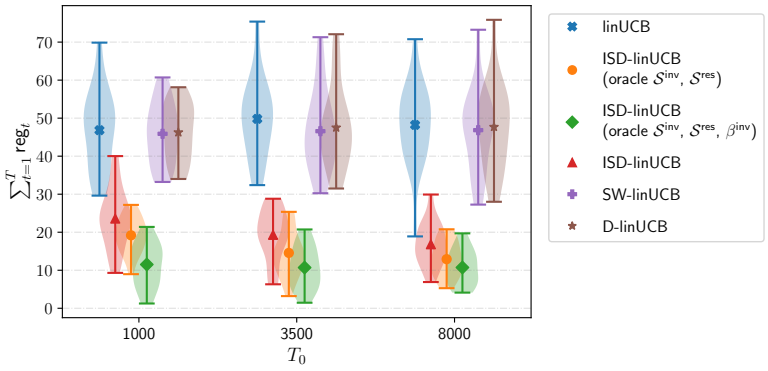


Figure 3.B.3: Cumulative regret for $T_0 \in \{1000, 3500, 8000\}$ for ISD-linUCB, in comparison with the same algorithm having oracle information and with the standard linUCB alogorithm, with SW-linUCB and D-linUCB (unaffected by T_0). For each T_0 the experiment is repeated 20 times.

3.C. Regret analysis: proofs

3.C.1 Oracle $\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}}, \beta^{\text{inv}}$

Proof of Theorem 3.4.1. We start by proving that the estimated confidence set $\hat{\mathcal{C}}_t^{\delta}$ around $\hat{\delta}_t^{\text{res}}$ contains δ_t^{res} with high probability.

Let $\tilde{\Sigma}_t^{\text{res}}$ and $\hat{\Sigma}_t$ be defined as in Section 3.3 and (3.2.3), respectively, and let $\tilde{\delta}_t^{\text{res}} := \left(\lambda I_{p^{\text{res}}} + (\mathbf{X}^{\text{res}})_t^\top \mathbf{X}_t^{\text{res}}\right)^{-1} (\mathbf{X}_t^{\text{res}})^\top \mathbf{R}_t^{\text{res}}$. For all $t \in [T]$, we have that

$$
\begin{aligned}
\|\hat{\delta}_t^{\text{res}} - \delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}} &= \|\tilde{\delta}_t^{\text{res}} - (U^{\text{res}})^\top \delta_t^{\text{res}}\|_{\tilde{\Sigma}_{t-1}^{\text{res}}} \\
&= \|(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1} \sum_{\tau=1}^{t-1} (U^{\text{res}})^\top \varphi(X_\tau, a_\tau)(\varphi(X_\tau, a_\tau)^\top \delta_\tau^{\text{res}} + \epsilon_\tau) - (U^{\text{res}})^\top \delta_t^{\text{res}}\|_{\tilde{\Sigma}_{t-1}^{\text{res}}} \\
&= \|(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1} \sum_{\tau=1}^{t-1} (U^{\text{res}})^\top \varphi(X_\tau, a_\tau)\epsilon_\tau - \lambda(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{\tilde{\Sigma}_{t-1}^{\text{res}}} \\
&= \|\sum_{\tau=1}^{t-1} (U^{\text{res}})^\top \varphi(X_\tau, a_\tau)\epsilon_\tau - \lambda(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} \\
&\leq \|\sum_{\tau=1}^{t-1} (U^{\text{res}})^\top \varphi(X_\tau, a_\tau)\epsilon_\tau\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} + \|\lambda(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} \\
&\leq \|\sum_{\tau=1}^{t-1} (U^{\text{res}})^\top \varphi(X_\tau, a_\tau)\epsilon_\tau\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} + \sqrt{\lambda}\|\delta_t^{\text{res}}\|_2. \qquad (3.C.1)
\end{aligned}
$$

From Lemma 3.2.2, we obtain that, for all $\eta \in (0,1)$, with probability at least $1-\eta$,

$$\|\hat{\delta}_t^{\text{res}} - \delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}} \leq \sigma \sqrt{2\log\left(\frac{1}{\eta}\right) + \log\left(\frac{\det(\tilde{\Sigma}_{t-1}^{\text{res}})}{\det(\lambda I_{p^{\text{res}}})}\right)} + \sqrt{\lambda}\|\delta_t^{\text{res}}\|_2.$$

Assuming that L and M where chosen sufficiently large in Algorithm 1 such that they indeed bound the context-action features and the linear parameter respectively, we get that, with probability at least $1-\eta$,

$$\|\hat{\delta}_t^{\text{res}} - \delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}} \leq \sigma \sqrt{2\log\left(\frac{1}{\eta}\right) + p^{\text{res}}\log\left(1 + \frac{tL^2}{\lambda p^{\text{res}}}\right)} + \sqrt{\lambda}M =: \sqrt{\hat{\rho}_t^{\text{res}}(\eta, L, M)}.$$

Assumption 1 allows us to estimate $\gamma_{0,t}$ by separately estimating β^{inv} and δ_t^{res} on the invariant and residual subspace respectively. This implies that the full regret consists of the sum of regret on the invariant and residual subspace, as follows

$$\begin{aligned}
\text{reg}_t &= (\varphi(X_t, a_t^*) - \varphi(X_t, a_t))^\top \gamma_{0,t} \\
&\leq \varphi(X_t, a_t)^\top (\bar{\gamma}_t - \gamma_{0,t}) \\
&= \underbrace{\varphi(X_t, a_t)^\top (\bar{\beta}_t - \beta^{\text{inv}})}_{\text{reg}_t^{\text{inv}}} + \underbrace{\varphi(X_t, a_t)^\top (\bar{\delta}_t - \delta_t^{\text{res}})}_{\text{reg}_t^{\text{res}}} \\
&= \text{reg}_t^{\text{res}},
\end{aligned} \tag{3.C.2}$$

where $\bar{\gamma}_t = \arg\max_{\gamma \in \hat{\mathcal{C}}^\beta \oplus \hat{\mathcal{C}}_t^\delta} \varphi(X_t, a_t)^\top \gamma$, $(\bar{\beta}_t, \bar{\delta}_t) \coloneqq \arg\max_{\beta \in \hat{\mathcal{C}}^\beta, \delta \in \hat{\mathcal{C}}_t^\delta} \varphi(X_t, a_t)^\top (\beta + \delta)$ and we have that by construction $\bar{\gamma}_t = \bar{\beta}_t + \bar{\delta}_t$. The first inequality holds because, by definition, $\bar{\gamma}_t$ maximizes the upper confidence bound at time t. The last equality follows from the assumption that we have oracle knowledge of β^{inv}, therefore $\hat{\mathcal{C}}^\beta = \{\beta^{\text{inv}}\}$ and $\text{reg}_t^{\text{inv}} = 0$ and we only need to consider the regret on the residual subspace. Then, for all $t \in [T]$, the residual instantaneous regret is such that

$$\begin{aligned}
\text{reg}_t^{\text{res}} &= \varphi(X_t, a_t)^\top (\bar{\delta}_t - \delta_t^{\text{res}}) \\
&\leq \|(U^{\text{res}})^\top \varphi(X_t, a_t)\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} \|(U^{\text{res}})^\top (\bar{\delta}_t - \delta_t^{\text{res}})\|_{\tilde{\Sigma}_{t-1}^{\text{res}}} \\
&\leq 2\|(U^{\text{res}})^\top \varphi(X_t, a_t)\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} \sqrt{\hat{\rho}_t^{\text{res}}(\eta, L, M)},
\end{aligned} \tag{3.C.3}$$

The first inequality follows from Cauchy-Schwarz. The last inequality is obtained by adding and subtracting $\hat{\delta}_t^{\text{res}}$ inside the second norm and using that $\bar{\delta}_t \in \hat{\mathcal{C}}_t^\delta$ by definition and that, with probability at least $1-\eta$, $\delta_t^{\text{res}} \in \hat{\mathcal{C}}_t^\delta$. Moreover, we have that for all $t \in [T]$, $\text{reg}_t \leq 2LM$ and for $\eta = 1/T$, with $T > 2$, that $\hat{\rho}_t^{\text{res}}(\eta, L, M) \geq 1$. Together with (3.C.3), it implies that

$$\text{reg}_t^{\text{res}} \leq \min\{2LM, 2\sqrt{\hat{\rho}_t^{\text{res}}(\eta, L, M)}\|(U^{\text{res}})^\top \varphi(X_t, a_t)\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\}$$

$$\leq 2\max\{LM,1\}\sqrt{\hat{\rho}_t^{\text{res}}(\eta,L,M)}\min\{1,\|(U^{\text{res}})^\top\varphi(X_t,a_t)\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\}$$
$$\leq C^{\text{res}}\sqrt{\hat{\rho}_t^{\text{res}}(\eta,L,M)}\min\{1,\|(U^{\text{res}})^\top\varphi(X_t,a_t)\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\}$$

where $C^{\text{res}} \coloneqq 2\max\{LM,1\}$. Using that, for all $x \geq 0$, $\min\{1,x\} \leq 2\log(1+x)$, we can upper bound the cumulative full regret as

$$\begin{aligned}\text{Reg}_T &= \sum_{t\in[T]} \text{reg}_T^{\text{res}} \leq \sqrt{T\sum_{t\in[T]}(\text{reg}_t^{\text{res}})^2} \\ &\leq C^{\text{res}}\sqrt{T\rho_T^{\text{res}}(\eta,L,M)\sum_{t\in[T]}\log(1+\|(U^{\text{res}})^\top\varphi(X_t,a_t)\|^2_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}})}. \end{aligned} \tag{3.C.4}$$

As a final step, we use the following Lemma by Lattimore and Szepesvári [2020], with notation adapted to our problem.

Lemma 3.6.1 (Lattimore and Szepesvári [2020], Lemma 19.4)**.** *Consider the same assumptions of Theorem 3.4.1. Let $\lambda > 0$ and $L < \infty$. Then,*

$$\sum_{t\in[T]}\log(1+\|(U^{\text{res}})^\top\varphi(X_t,a_t)\|^2_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}) \leq p^{\text{res}}\log\left(1+\frac{TL^2}{\lambda p^{\text{res}}}\right).$$

Lemma 3.6.1 implies that

$$\text{Reg}_T \leq C^{\text{res}}\sqrt{Tp^{\text{res}}\log\left(1+\frac{TL^2}{\lambda p^{\text{res}}}\right)}\left(\sigma\sqrt{2\log\left(\frac{1}{\eta}\right)+p^{\text{res}}\log\left(1+\frac{TL^2}{\lambda p^{\text{res}}}\right)}+\sqrt{\lambda}M\right).$$

Finally, choosing $\eta = 1/T$ implies that Reg_T is $\tilde{O}(p^{\text{res}}\sqrt{T})$. □

3.C.2 Oracle subspaces

Proof of Theorem 3.4.2. We start by showing that, under oracle knowledge of the subspaces, the radius $\sqrt{\hat{\rho}_{T_0}^{\text{inv}}(\eta,L,M)}$ for the confidence set $\hat{\mathcal{C}}^\beta$ introduced in (3.3.7) can be defined to be $\tilde{O}(\sqrt{p^{\text{inv}}}+\sqrt{\frac{1}{\lambda_0}})$. To do so, we consider the estimation error on $\hat{\beta}^{\text{inv}}$ defined in (3.3.6), where $\hat{U}^{\text{inv}}$ is replaced by U^{inv}.

$$\begin{aligned}&\|\hat{\beta}^{\text{inv}}-\beta\|_{\hat{\Sigma}_{[-T_0]}} \\ &= \|U^{\text{inv}}(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}\sum_{t=-T_0}^{-1}(U^{\text{inv}})^\top\varphi(X_t,a_t)R_t^{a_t} - U^{\text{inv}}(U^{\text{inv}})^\top\beta^{\text{inv}}\|_{\hat{\Sigma}_{[-T_0]}} \\ &= \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}\sum_{t=-T_0}^{-1}(U^{\text{inv}})^\top\varphi(X_t,a_t)R_t^{a_t} - (U^{\text{inv}})^\top\beta^{\text{inv}}\|_{\tilde{\Sigma}^{\text{inv}}_{[-T_0]}}\end{aligned} \tag{3.C.5}$$

$$
\begin{aligned}
&= \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1} \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t)(\varphi(X_t, a_t)^\top(\beta^{\text{inv}} + \delta_t^{\text{res}}) + \epsilon_t) - (U^{\text{inv}})^\top \beta^{\text{inv}}\|_{\tilde{\Sigma}^{\text{inv}}_{[-T_0]}} \\
&= \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1} \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t)(\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}} + \epsilon_t)\|_{\tilde{\Sigma}^{\text{inv}}_{[-T_0]}} \\
&\leq \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1} \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{\tilde{\Sigma}^{\text{inv}}_{[-T_0]}} \\
&\quad + \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1} \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t)\epsilon_t\|_{\tilde{\Sigma}^{\text{inv}}_{[-T_0]}} \\
&= \| \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \qquad (3.C.6) \\
&\quad + \|(\tilde{\mathbf{X}}^{\text{inv}}_{T_0})^\top \boldsymbol{\epsilon}_{T_0}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \qquad (3.C.7)
\end{aligned}
$$

where $\tilde{\Sigma}^{\text{inv}}_{[-T_0]} \coloneqq (U^{\text{inv}})^\top \hat{\Sigma}_{[-T_0]} U^{\text{inv}}$, $\tilde{\mathbf{X}}^{\text{inv}}_{T_0} \coloneqq [(U^{\text{inv}})^\top \varphi(X_{-T_0}, a_{-T_0}), \dots, (U^{\text{inv}})^\top \varphi(X_{-1}, a_{-1})]^\top \in \mathbb{R}^{T_0 \times p^{\text{inv}}}$ and $\boldsymbol{\epsilon}_{T_0} \coloneqq [\epsilon_{-T_0}, \dots, \epsilon_{-1}]^\top \in \mathbb{R}^{T_0}$. For the term in (3.C.6), it holds that

$$
\begin{aligned}
&\| \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\
\leq &\|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-\frac{1}{2}}\|_{\text{op}} \| \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_2 \\
\leq &\sqrt{\frac{1}{\lambda_0 T_0}} \| \sum_{t=-T_0}^{-1} \alpha_t\|_2 \qquad (3.C.8)
\end{aligned}
$$

where $\alpha_t \coloneqq (U^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}} \in \mathbb{R}^{p^{\text{inv}}}$. We bound $\| \sum_{t=-T_0}^{-1} \alpha_t\|_2$ using the matrix Hoeffding inequality [see, for example, Tropp, 2012, Theorem 1.3, which we report below].

Lemma 3.6.2 (Matrix Hoeffding inequality)**.** *Let $\{\mathbf{M}_k\}_{k\in[K]}$ a finite sequence of independent, random, self-adjoint matrices of dimension p and let $\{\mathbf{A}_k\}_{k\in[K]}$ be a sequence of fixed self-adjoint matrices of dimension p. Assume that, for all $k \in [K]$, $\mathbb{E}[\mathbf{M}_k] = 0$ and $\mathbf{M}_k^2 \preceq \mathbf{A}_k$ almost surely. Then, for all $\xi > 0$, it holds that*

$$
\mathbb{P}\left(\lambda_{\max}\left(\sum_{k\in[K]} \mathbf{M}_k\right) \geq \xi\right) \leq p \exp\left(-\frac{\xi^2}{8\sigma^2}\right) \quad \textit{where} \quad \sigma^2 \coloneqq \left\|\sum_{k\in[K]} \mathbf{A}_k^2\right\|.
$$

By definition of ISD, it holds that, for all $t \in [-T_0]$, $\mathbb{E}[\alpha_t] = 0$. Using Cauchy-Schwarz, we also have that $\|\alpha_t\|_2 \leq L^2 M$. Moreover, we can construct the self-adjoint dilation

[see Tropp, 2012, Section 2.6] for α_t as

$$\mathscr{S}(\alpha_t) \coloneqq \begin{bmatrix} 0 & \alpha_t \\ \alpha_t^\top & 0 \end{bmatrix} \in \mathbb{R}^{(p^{\text{inv}}+1)\times(p^{\text{inv}}+1)},$$

which is such that $\|\alpha_t\|_2 = \|\mathscr{S}(\alpha_t)\|_{\text{op}} = \lambda_{\max}(\mathscr{S}(\alpha_t))$ and

$$\mathscr{S}(\alpha_t)^2 \coloneqq \begin{bmatrix} \alpha_t\alpha_t^\top & 0 \\ 0 & \alpha_t^\top\alpha_t \end{bmatrix}.$$

Hoeffding inequality can be now applied to the sequence of self-adjoint matrices $(\mathscr{S}(\alpha_t))_{t\in[-T_0]}$. Let $\sigma_\alpha^2 \coloneqq \|\sum_{t\in[-T_0]} \mathscr{S}(\alpha_t)^2\|_{\text{op}} \leq \sum_{t\in[-T_0]} \|\mathscr{S}(\alpha_t)\|_{\text{op}}^2 \leq T_0(L^2M)^2$. Then, for all $\xi \geq 0$ it holds by Hoeffding's inequality that

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{t\in[-T_0]} \mathscr{S}(\alpha_t)\right) \geq \xi\right) \leq (p^{\text{inv}}+1)e^{-\frac{\xi^2}{8\sigma_\alpha^2}}.$$

Using the definition of $\mathscr{S}(\alpha_t)$ and defining $\eta \coloneqq (p^{\text{inv}}+1)e^{-\frac{\xi^2}{8\sigma_\alpha^2}}$, the above is equivalent to

$$\mathbb{P}\left(\|\sum_{t\in[-T_0]} \alpha_t\|_2 \leq 2L^2M\sqrt{2T_0\log\left(\frac{p^{\text{inv}}+1}{\eta}\right)}\right) \geq 1-\eta.$$

Therefore, with probability at least $1-\eta$, $\sqrt{\frac{1}{\lambda_0 T_0}}\|\sum_{t=-T_0}^{-1} \alpha_t\|_2$ is $O\left(\sqrt{\frac{1}{\lambda_0}\log\left(\frac{p^{\text{inv}}}{\eta}\right)}\right)$.

We now need to bound the term $\|(\tilde{\mathbf{X}}_{T_0}^{\text{inv}})^\top \boldsymbol{\epsilon}_{T_0}\|_{(\tilde{\Sigma}_{[-T_0]}^{\text{inv}})^{-1}}$ in (3.C.7). By Assumption 2, we have that $\hat{\Sigma}_{[-T_0]} \succeq \lambda_0 T_0 I_p$, which implies that $\tilde{\Sigma}_{[-T_0]}^{\text{inv}} \succeq \lambda_0 T_0 I_{p^{\text{inv}}}$ and therefore $\tilde{\Sigma}_{[-T_0]}^{\text{inv}} \succeq \frac{1}{2}(\lambda_0 T_0 I_{p^{\text{inv}}} + \tilde{\Sigma}_{[-T_0]}^{\text{inv}})$. This implies that,

$$\|(\tilde{\mathbf{X}}_{T_0}^{\text{inv}})^\top \boldsymbol{\epsilon}_{T_0}\|_{(\tilde{\Sigma}_{[-T_0]}^{\text{inv}})^{-1}} \leq \sqrt{2}\|(\tilde{\mathbf{X}}_{T_0}^{\text{inv}})^\top \boldsymbol{\epsilon}_{T_0}\|_{(\lambda_0 T_0 I_{p^{\text{inv}}} + \tilde{\Sigma}_{T_0}^{\text{inv}})^{-1}}.$$

Lemma 3.2.2 implies that, for all $\eta \in (0,1)$, with probability at least $1-\eta$,

$$\|(\tilde{\mathbf{X}}_{T_0}^{\text{inv}})^\top \boldsymbol{\epsilon}_{T_0}\|_{(\lambda_0 T_0 I_{p^{\text{inv}}} + \tilde{\Sigma}_{[-T_0]}^{\text{inv}})^{-1}} \leq \sigma\sqrt{\log\left(\frac{1}{\eta}\right) + \frac{1}{2}\log\left(\frac{\det(\tilde{\Sigma}_{[-T_0]}^{\text{inv}})}{\det(\lambda_0 T_0 I_{p^{\text{inv}}})}\right)}.$$

By the arithmetic mean—geometric mean inequality, it holds that

$$\det(\tilde{\Sigma}_{[-T_0]}^{\text{inv}}) \leq \left(\frac{1}{p^{\text{inv}}}\operatorname{trace}(\tilde{\Sigma}_{[-T_0]}^{\text{inv}})\right)^{p^{\text{inv}}} \leq \left(\frac{T_0L^2}{p^{\text{inv}}}\right)^{p^{\text{inv}}}$$

where the last inequality follows from $\operatorname{trace}(\tilde{\Sigma}_{[-T_0]}^{\text{inv}}) = \operatorname{trace}((\tilde{\mathbf{X}}_{T_0}^{\text{inv}})^\top \tilde{\mathbf{X}}_{T_0}^{\text{inv}}) \leq T_0L^2$. More-

over, $\det(\lambda_0 T_0 I_{p^{\text{inv}}}) = (\lambda_0 T_0)^{p^{\text{inv}}}$. Therefore, with probability at least $1-\eta$,

$$\|(\tilde{\mathbf{X}}_{T_0}^{\text{inv}})^\top \boldsymbol{\epsilon}_{T_0}\|_{(\lambda_0 T_0 I_{p^{\text{inv}}} + \tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \leq \sigma \sqrt{\log\left(\frac{1}{\eta}\right) + \frac{p^{\text{inv}}}{2} \log\left(\frac{T_0 L^2}{p^{\text{inv}} \lambda_0 T_0}\right)} \tag{3.C.9}$$

and, with probability at least $1-2\eta$,

$$\begin{aligned}
&\|\hat{\beta}^{\text{inv}} - \beta\|_{\hat{\Sigma}_{[-T_0]}} \\
\leq &\sigma \sqrt{2\log\left(\frac{1}{\eta}\right) + p^{\text{inv}} \log\left(\frac{L^2}{p^{\text{inv}} \lambda_0}\right)} + 2L^2 M \sqrt{\frac{2}{\lambda_0} \log\left(\frac{p^{\text{inv}}+1}{\eta}\right)} =: \hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M).
\end{aligned} \tag{3.C.10}$$

Using Assumption 2, we further obtain that, with probability at least $1-2\eta$,

$$\|\hat{\beta}^{\text{inv}} - \beta\|_2 \leq \sqrt{\frac{\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)}{\lambda_0 T_0}}. \tag{3.C.11}$$

Consider now Algorithm 1. For all $a \in \mathcal{A}$, let $\text{UCB}_t^{\beta,\delta}(a) \coloneqq \max_{\beta \in \hat{\mathcal{C}}^\beta, \delta \in \hat{\mathcal{C}}_t^\delta} \varphi(X_t, a)^\top (\beta + \delta)$ and let $(\bar{\beta}_t, \bar{\delta}_t) \coloneqq \arg\max_{\beta \in \hat{\mathcal{C}}^\beta, \delta \in \hat{\mathcal{C}}_t^\delta} \varphi(X_t, a_t)^\top (\beta + \delta)$. Then, as in (3.C.2), for all $t \in [T]$, the instantaneous regret is such that $\text{reg}_t \leq \text{reg}_t^{\text{inv}} + \text{reg}_t^{\text{res}}$. We upper bound the $\text{reg}_t^{\text{inv}}$ as follows

$$\begin{aligned}
\text{reg}_t^{\text{inv}} &\coloneqq ((U^{\text{inv}})^\top \varphi(X_t, a_t))^\top (U^{\text{inv}})^\top (\bar{\beta}_t - \beta^{\text{inv}}) \\
&\leq \|(U^{\text{inv}})^\top \varphi(X_t, a_t)\|_2 \|(U^{\text{inv}})^\top (\bar{\beta}_t - \beta^{\text{inv}})\|_2 \\
&\leq \frac{2L}{\sqrt{\lambda_0 T_0}} \left(\sigma \sqrt{2\log\left(\frac{1}{\eta}\right) + p^{\text{inv}} \log\left(\frac{L^2}{p^{\text{inv}} \lambda_0}\right)} + 2L^2 M \sqrt{\frac{2}{\lambda_0} \log\left(\frac{p^{\text{inv}}+1}{\eta}\right)} \right)
\end{aligned}$$

where the first inequality holds with probability $1-2\eta$ and follows from the definition of $\hat{\mathcal{C}}^\beta$ and from Cauchy-Schwarz inequality.

For $\text{reg}_t^{\text{res}}$, the analysis from Theorem 3.4.1 remains valid, with an additional term to be considered due to using the estimated invariant component instead of the oracle one. In particular, following the same steps as in (3.C.1), we have that, for all $t \in [T]$,

$$\begin{aligned}
\|\hat{\delta}_t^{\text{res}} - \delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}} \leq &\|\sum_{\tau=1}^{t-1} (U^{\text{res}})^\top \varphi(X_\tau, a_\tau) \epsilon_\tau\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}} + \sqrt{\lambda}\|\delta_t^{\text{res}}\|_2 \\
&+ \|\sum_{\tau=1}^{t-1} (U^{\text{res}})^\top \varphi(X_\tau, a_\tau) \varphi(X_\tau, a_\tau)^\top (\beta^{\text{inv}} - \hat{\beta}^{\text{inv}})\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}.
\end{aligned}$$

We analyzed the first two terms in Theorem 3.4.1. For the last term, we have that

$$\begin{aligned}&\|\sum_{\tau=1}^{t-1}(U^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(\beta^{\text{inv}}-\hat{\beta}^{\text{inv}})\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}\\ \leq&\|(\tilde{\Sigma}^{\text{res}}_{t-1})^{-\frac{1}{2}}\|_{\text{op}}\|\sum_{\tau=1}^{t-1}(U^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top U^{\text{inv}}\|_{\text{op}}\|(U^{\text{inv}})^\top(\beta^{\text{inv}}-\hat{\beta}^{\text{inv}})\|_2.\end{aligned}$$

Using matrix Hoeffding inequality (see Lemma 3.6.2) and following the same steps used to bound (3.C.6) that, for all $\eta\in(0,1)$, with probability at least $1-\eta$,

$$\|\sum_{\tau=1}^{t-1}(U^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top U^{\text{inv}}\|_{\text{op}}\leq 2L^2\sqrt{2(t-1)\log\left(\frac{p+1}{\eta}\right)}.$$

Therefore, with probability at least $1-3\eta$,

$$\begin{aligned}&\|\sum_{\tau=1}^{t-1}(U^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(\beta^{\text{inv}}-\hat{\beta}^{\text{inv}})\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}\\ \leq&\frac{1}{\sqrt{\lambda}}2L^2\sqrt{2(t-1)\log\left(\frac{p+1}{\eta}\right)}\sqrt{\frac{\hat{\rho}^{\text{inv}}_{T_0}(\eta,L,M)}{\lambda_0 T_0}}.\end{aligned}$$

The right-hand side of the inequality can be added to $\sqrt{\hat{\rho}^{\text{res}}_t(\eta,L,M)}$ obtained in the proof of Theorem 3.4.1, so that, with probability at least $1-4\eta$

$$\begin{aligned}\sum_{t\in[T]}\text{reg}^{\text{res}}_t\leq C^{\text{res}}&\sqrt{Tp^{\text{res}}\log\left(1+\frac{TL^2}{\lambda p^{\text{res}}}\right)}\\ &\times\left(\sigma\sqrt{2\log\left(\frac{1}{\eta}\right)+p^{\text{res}}\log\left(1+\frac{TL^2}{\lambda p^{\text{res}}}\right)}\right.\\ &\left.+\sqrt{\lambda}M2L^2\sqrt{\frac{2T\hat{\rho}^{\text{inv}}_{T_0}(\eta,L,M)}{\lambda\lambda_0 T_0}\log\left(\frac{p+1}{\eta}\right)}\right).\end{aligned}$$

Finally, this implies that the cumulative regret can be upper bounded, for all $\eta\in(0,\frac{1}{5})$ with probability at least $1-5\eta$ as

$$\begin{aligned}\text{Reg}_T=\sum_{t\in[T]}\text{reg}_t&=\sum_{t\in[T]}\text{reg}^{\text{inv}}+\sum_{t\in[T]}\text{reg}^{\text{res}}\\ &\leq\frac{2LT}{\sqrt{\lambda_0 T_0}}\left(\sigma\sqrt{2\log\left(\frac{1}{\eta}\right)+p^{\text{inv}}\log\left(\frac{L^2}{p^{\text{inv}}\lambda_0}\right)}+2L^2M\sqrt{\frac{2}{\lambda_0}\log\left(\frac{p^{\text{inv}}+1}{\eta}\right)}\right)\end{aligned}$$

$$+ C^{\mathrm{res}} \sqrt{T p^{\mathrm{res}} \log\left(1 + \frac{TL^2}{\lambda p^{\mathrm{res}}}\right)}$$
$$\times \left(\sigma \sqrt{2 \log\left(\frac{1}{\eta}\right) + p^{\mathrm{res}} \log\left(1 + \frac{TL^2}{\lambda p^{\mathrm{res}}}\right)} \right.$$
$$\left. + \sqrt{\lambda} M + 2L^2 \sqrt{\frac{2T \hat{\rho}^{\mathrm{inv}}_{T_0}(\eta, L, M)}{\lambda \lambda_0 T_0} \log\left(\frac{p+1}{\eta}\right)} \right).$$

Choosing $\eta' = 1/T$, where $\eta' \coloneqq 5\eta$, implies that Reg_T is, with probability at least $1 - \eta'$ $\tilde{O}\left(\sqrt{T}\left(p^{\mathrm{res}} + \sqrt{\frac{p^{\mathrm{inv}} T}{\lambda_0 T_0}} + \frac{1}{\lambda_0}\sqrt{\frac{T}{T_0}}\right)\right)$. □

3.C.3 Accounting for subspace decomposition errors

Proof of Lemma 3.4.3. We want to bound the $\hat{\Sigma}_{[-T_0]}$-norm of the estimation error for the invariant component in the estimated invariant subspace $\hat{\mathcal{S}}^{\mathrm{inv}}$, i.e., $\|\hat{\beta}^{\mathrm{inv}} - \hat{\Pi}^{\mathcal{S}^{\mathrm{inv}}} \beta^{\mathrm{inv}}\|_{\hat{\Sigma}_{[-T_0]}}$. Let $\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]} \coloneqq (\hat{U}^{\mathrm{inv}})^\top \hat{\Sigma}_{[-T_0]} \hat{U}^{\mathrm{inv}}$, $\mathbf{X}_{T_0} \coloneqq [\varphi(X_{-T_0}, a_{-T_0}), \dots, \varphi(X_{-1}, a_{-1})]^\top \in \mathbb{R}^{T_0 \times p}$, $\mathbf{R}_{T_0} \coloneqq [R^{a_{-T_0}}_{-T_0}, \dots, R^{a_{-1}}_{-1}]^\top \in \mathbb{R}^{T_0}$ and $\boldsymbol{\epsilon}_{T_0} \coloneqq [\epsilon_{-T_0}, \dots, \epsilon_{-1}]^\top \in \mathbb{R}^{T_0}$. Then, we have that

$$\begin{aligned}
\|\hat{\beta}^{\mathrm{inv}} - \hat{\Pi}^{\mathcal{S}^{\mathrm{inv}}} \beta^{\mathrm{inv}}\|_{\hat{\Sigma}_{[-T_0]}} &= \|\hat{U}^{\mathrm{inv}}((\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]})^{-1} \sum_{t=-T_0}^{-1} (\hat{U}^{\mathrm{inv}})^\top \varphi(X_t, a_t) R^{a_t}_t - (\hat{U}^{\mathrm{inv}})^\top \beta^{\mathrm{inv}})\|_{\hat{\Sigma}_{[-T_0]}} \\
&\leq \|(\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]})^{-1} (\hat{U}^{\mathrm{inv}})^\top \mathbf{X}_{T_0}^\top \mathbf{X}_{T_0} \beta^{\mathrm{inv}} - (\hat{U}^{\mathrm{inv}})^\top \beta^{\mathrm{inv}}\|_{\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]}} \\
&\quad + \|(\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]})^{-1} \sum_{t=-T_0}^{-1} (\hat{U}^{\mathrm{inv}})^\top \varphi(X_t, a_t) \varphi(X_t, a_t)^\top \delta^{\mathrm{res}}_t\|_{\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]}} \\
&\quad + \|(\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]})^{-1} (\hat{U}^{\mathrm{inv}})^\top \mathbf{X}_{T_0}^\top \boldsymbol{\epsilon}_{T_0}\|_{\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]}} \\
&= \|(\hat{U}^{\mathrm{inv}})^\top \mathbf{X}_{T_0}^\top \mathbf{X}_{T_0} (U^{\mathrm{inv}}(U^{\mathrm{inv}})^\top - \hat{U}^{\mathrm{inv}}(\hat{U}^{\mathrm{inv}})^\top) \beta^{\mathrm{inv}}\|_{(\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]})^{-1}} \qquad (3.\mathrm{C}.12) \\
&\quad + \|\sum_{t=-T_0}^{-1} (\hat{U}^{\mathrm{inv}})^\top \varphi(X_t, a_t) \varphi(X_t, a_t)^\top \delta^{\mathrm{res}}_t\|_{(\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]})^{-1}} \qquad (3.\mathrm{C}.13) \\
&\quad + \|(\hat{U}^{\mathrm{inv}})^\top \mathbf{X}_{T_0}^\top \boldsymbol{\epsilon}_{T_0}\|_{(\tilde{\Sigma}^{\mathrm{inv}}_{[-T_0]})^{-1}}. \qquad (3.\mathrm{C}.14)
\end{aligned}$$

The term in (3.C.14) can be upper bounded exactly as in Section 3.C.2 (see Equation (3.C.9)), since the presence of $\hat{U}^{\mathrm{inv}}$ in place of U^{inv} does not influence its analysis. The terms in (3.C.12) and (3.C.13) appear because of the misalignment between the true and the estimated invariant subspace: the former represents how much of the invariant parameter we are not able to estimate due to the difference between $\hat{\mathcal{S}}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{inv}}$, the latter quantifies how much of the true residual parameter enters in the estimation of the

invariant component due to the intersection between $\hat{\mathcal{S}}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$. For (3.C.12), we have that

$$\begin{aligned}
&\|(\hat{U}^{\text{inv}})^\top \mathbf{X}_{T_0}^\top \mathbf{X}_{T_0}(\Pi^{\mathcal{S}^{\text{inv}}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}})\beta^{\text{inv}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\
=&\|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-\frac{1}{2}}(\hat{U}^{\text{inv}})^\top \mathbf{X}_{T_0}^\top \mathbf{X}_{T_0}(\Pi^{\mathcal{S}^{\text{inv}}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}})\beta^{\text{inv}}\|_2 \\
\leq&\sqrt{\sum_{t\in[-T_0]} \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-\frac{1}{2}}(\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t)\|_2^2}\sqrt{\sum_{t\in[-T_0]} (\varphi(X_t, a_t)^\top(\Pi^{\mathcal{S}^{\text{inv}}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}})\beta^{\text{inv}})^2} \\
\leq&\sqrt{p^{\text{inv}}}\sqrt{T_0 L^2 M^2 (\Delta\Pi)^2} =: \hat{\rho}_1^{\text{inv}}.
\end{aligned}$$

The first inequality uses Cauchy-Schwarz inequality. The second inequality follows from the fact that

$$\begin{aligned}
&\sum_{t\in[-T_0]} \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-\frac{1}{2}}(\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t)\|_2^2 \\
=&\sum_{t\in[-T_0]} \varphi(X_t, a_t)^\top \hat{U}^{\text{inv}}(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}(\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t) \\
=&\operatorname{trace}\left(\sum_{t\in[-T_0]} (\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top \hat{U}^{\text{inv}}(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}\right) \\
=&\operatorname{trace}\left(I_{p^{\text{inv}}}\right) \\
=&\, p^{\text{inv}},
\end{aligned}$$

and by singling out the norm of the individual factors under the second square root. Finally, under Assumption 3 we obtain that, for all $\eta \in (0,1)$, it holds with probability at least $1-\eta$ that $\|(\hat{U}^{\text{inv}})^\top \mathbf{X}_{T_0}^\top \mathbf{X}_{T_0}(\Pi^{\mathcal{S}^{\text{inv}}} - \hat{\Pi}^{\mathcal{S}^{\text{inv}}})\beta^{\text{inv}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}}$ is $O(\sqrt{p^{\text{inv}} \log(\frac{p}{\eta})})$.

We now need to upper bound the term in (3.C.13), that is

$$\begin{aligned}
&\|\sum_{t=-T_0}^{-1} (\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\
=&\|\sum_{t=-T_0}^{-1} (\hat{U}^{\text{inv}})^\top (U^{\text{inv}}(U^{\text{inv}})^\top + U^{\text{res}}(U^{\text{res}})^\top)\varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\
\leq&\|\sum_{t=-T_0}^{-1} (\hat{U}^{\text{inv}})^\top U^{\text{inv}}(U^{\text{inv}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\
+&\|\sum_{t=-T_0}^{-1} (\hat{U}^{\text{inv}})^\top U^{\text{res}}(U^{\text{res}})^\top \varphi(X_t, a_t)\varphi(X_t, a_t)^\top U^{\text{res}}(U^{\text{res}})^\top \delta_t^{\text{res}}\|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}}
\end{aligned}$$

For the second term, it holds that

$$\begin{aligned}
&\| \sum_{t=-T_0}^{-1} (\hat{U}^{\text{inv}})^\top U^{\text{res}} (U^{\text{res}})^\top \varphi(X_t, a_t) \varphi(X_t, a_t)^\top U^{\text{res}} (U^{\text{res}})^\top \delta_t^{\text{res}} \|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\
\leq & \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-\frac{1}{2}}\|_{\text{op}} \|(\hat{U}^{\text{inv}})^\top U^{\text{res}}\|_{\text{op}} T_0 L^2 M \\
\leq & \sqrt{\frac{1}{\lambda_0 T_0}} \Delta\Pi T_0 L^2 M \\
= & \sqrt{\frac{T_0}{\lambda_0}} \Delta\Pi L^2 M =: \hat{\rho}_2^{\text{inv}}.
\end{aligned}$$

Under Assumption 3, for all $\eta \in (0,1)$, this quantity is, with probability at least $1-\eta$, $O\left(\sqrt{\frac{\log(p/\eta)}{\lambda_0}}\right)$. For the first term, we obtain

$$\begin{aligned}
&\| \sum_{t=-T_0}^{-1} (\hat{U}^{\text{inv}})^\top U^{\text{inv}} (U^{\text{inv}})^\top \varphi(X_t, a_t) \varphi(X_t, a_t)^\top U^{\text{res}} (U^{\text{res}})^\top \delta_t^{\text{res}} \|_{(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-1}} \\
\leq & \|(\tilde{\Sigma}^{\text{inv}}_{[-T_0]})^{-\frac{1}{2}}\|_{\text{op}} \|(\hat{U}^{\text{inv}})^\top U^{\text{inv}}\|_{\text{op}} \| \sum_{t=-T_0}^{-1} (U^{\text{inv}})^\top \varphi(X_t, a_t) \varphi(X_t, a_t)^\top U^{\text{res}} (U^{\text{res}})^\top \delta_t^{\text{res}} \|_2 \\
\leq & \sqrt{\frac{1}{\lambda_0 T_0}} \| \sum_{t=-T_0}^{-1} \alpha_t \|_2
\end{aligned}$$

where $\alpha_t \coloneqq (U^{\text{inv}})^\top \varphi(X_t, a_t) \varphi(X_t, a_t)^\top U^{\text{res}} (U^{\text{res}})^\top \delta_t^{\text{res}} \in \mathbb{R}^{p^{\text{inv}}}$ and in the last inequality we have used that $\|(\hat{U}^{\text{inv}})^\top U^{\text{inv}}\|_{\text{op}} \leq 1$ since both matrices have orthonormal columns. This is the same quantity that appears in (3.C.8) in the oracle subspaces case (together with (3.C.14), this forms the full oracle bound given in (3.C.10), which we denote here by $\hat{\rho}_3^{\text{inv}}$). We have shown (3.C.8) to be, for all $\eta \in (0,1)$, with probability at least $1-\eta$, $O\left(\sqrt{\frac{1}{\lambda_0} \log\left(\frac{p^{\text{inv}}}{\eta}\right)}\right)$. For the same choice of η, this term is dominated by $O\left(\sqrt{\frac{1}{\lambda_0} \log\left(\frac{p}{\eta}\right)}\right)$. Hence, the expression in (3.C.13) is, with probability at least $1-2\eta$, $O\left(\sqrt{\frac{1}{\lambda_0} \log\left(\frac{p^{\text{inv}}}{\eta}\right)}\right)$.

Let

$$\begin{aligned}
\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M) &\coloneqq \sum_{i=1}^{3} \hat{\rho}_i^{\text{inv}} \qquad (3.C.15) \\
&= \sigma \sqrt{2 \log\left(\frac{1}{\eta}\right) + p^{\text{inv}} \log\left(\frac{L^2}{p^{\text{inv}} \lambda_0}\right)} + 2L^2 M \sqrt{\frac{2}{\lambda_0} \log\left(\frac{p^{\text{inv}}+1}{\eta}\right)}
\end{aligned}$$

$$+\sqrt{p^{\text{inv}}T_0}\Delta\Pi LM+\sqrt{\frac{T_0}{\lambda_0}}\Delta\Pi L^2M.$$

By the union bound we have that, with probability at least $1-4\eta$,

$$\|\hat{\beta}^{\text{inv}}-\hat{\Pi}^{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}}\|_{\hat{\Sigma}_{[-T_0]}}\leq\sum_{i=1}^{3}\hat{\rho}_i^{\text{inv}}=:\hat{\rho}_{T_0}^{\text{inv}}(\eta,L,M)$$

is

$$O\left(\sqrt{\left(\log\left(\frac{1}{\eta}\right)+p^{\text{inv}}\log\left(\frac{1}{p^{\text{inv}}\lambda_0}\right)\right)}\right)+O\left(\sqrt{p^{\text{inv}}\log\left(\frac{p}{\eta}\right)}\right)+O\left(\sqrt{\frac{1}{\lambda_0}\log\left(\frac{p}{\eta}\right)}\right).$$

This concludes the proof of the lemma. □

Proof of Lemma 3.4.4. We want to bound the estimation error for the residual component in the residual subspace, that is, $\|\hat{\delta}_t^{\text{res}}-\hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}}$. Recall that we have defined $\tilde{\Sigma}_t^{\text{res}}:=(\hat{U}^{\text{res}})^\top\hat{\Sigma}_t\hat{U}^{\text{res}}$. We start by stating explicitly the expression for $\hat{\delta}_t^{\text{res}}$, obtaining that

$$\begin{aligned}
&\|\hat{\delta}_t^{\text{res}}-\hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}}\\
=&\|\hat{U}^{\text{res}}(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)(R_\tau^{a_\tau}-\varphi(X_\tau,a_\tau)^\top\hat{\beta}^{\text{inv}})-\hat{U}^{\text{res}}(\hat{U}^{\text{res}})^\top\delta_t^{\text{res}}\|_{\hat{\Sigma}_{t-1}}\\
=&\|(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)(R_\tau^{a_\tau}-\varphi(X_\tau,a_\tau)^\top\hat{\beta}^{\text{inv}})-(\hat{U}^{\text{res}})^\top\delta_t^{\text{res}}\|_{\tilde{\Sigma}_{t-1}^{\text{res}}}\\
\leq&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(U^{\text{res}}(U^{\text{res}})^\top-\hat{U}^{\text{res}}(\hat{U}^{\text{res}})^\top)\delta_t^{\text{res}}-\lambda(\hat{U}^{\text{res}})^\top\delta_t^{\text{res}}\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\\
+&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(\beta^{\text{inv}}-\hat{\beta}^{\text{inv}})\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\\
+&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\epsilon_\tau\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}\\
\leq&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(U^{\text{res}}(U^{\text{res}})^\top-\hat{U}^{\text{res}}(\hat{U}^{\text{res}})^\top)\delta_t^{\text{res}}\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} &\text{(3.C.16)}\\
+&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(\beta^{\text{inv}}-\hat{\beta}^{\text{inv}})\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}} &\text{(3.C.17)}\\
+&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\epsilon_\tau\|_{(\tilde{\Sigma}_{t-1}^{\text{res}})^{-1}}+\sqrt{\lambda}\|\delta_t^{\text{res}}\|_2. &\text{(3.C.18)}
\end{aligned}$$

The analysis of the terms in (3.C.18) follows the same steps as the one for (3.C.1) in the oracle case. In particular, from Theorem 3.4.1 we have that for all $\eta\in(0,1)$, with

probability at least $1-\eta$,

$$\begin{aligned}&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\epsilon_\tau\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}+\sqrt{\lambda}\|\delta^{\text{res}}_t\|_2\\ \leq&\sigma\sqrt{2\log\left(\frac{1}{\eta}\right)+p^{\text{res}}\log\left(1+\frac{tL^2}{\lambda p^{\text{res}}}\right)}+\sqrt{\lambda}M=:\hat{\rho}^{\text{res}}_1.\end{aligned}$$

For the term in (3.C.16), we have that

$$\begin{aligned}&\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(\Pi^{\mathcal{S}^{\text{res}}}-\hat{\Pi}^{\mathcal{S}^{\text{res}}})\delta^{\text{res}}_t\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}\\ \leq&\sqrt{\sum_{\tau\in[t-1]}\|(\tilde{\Sigma}^{\text{res}}_{t-1})^{-\frac{1}{2}}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\|_2^2}\sqrt{\sum_{\tau\in[t-1]}(\varphi(X_\tau,a_\tau)^\top(\Pi^{\mathcal{S}^{\text{res}}}-\hat{\Pi}^{\mathcal{S}^{\text{res}}})\delta^{\text{res}}_t)^2}\\ \leq&\sqrt{p^{\text{res}}}\sqrt{t\max_{\tau\in[t-1]}\|\varphi(X_\tau,a_\tau)\|_2^2\|\delta^{\text{res}}_t\|_2^2\|\Pi^{\mathcal{S}^{\text{res}}}-\hat{\Pi}^{\mathcal{S}^{\text{res}}}\|_{\text{op}}^2}\\ \leq&LM\Delta\Pi\sqrt{p^{\text{res}}t}=:\hat{\rho}^{\text{res}}_2.\end{aligned}\tag{3.C.19}$$

The first inequality follows from Cauchy-Schwarz inequality. The second inequality uses the fact that

$$\begin{aligned}&\sum_{\tau\in[t-1]}\|(\tilde{\Sigma}^{\text{res}}_{t-1})^{-\frac{1}{2}}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\|_2^2\\ =&\sum_{\tau\in[t-1]}\varphi(X_\tau,a_\tau)^\top\hat{U}^{\text{res}}(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\\ =&\,\text{trace}\left(\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top\hat{U}^{\text{res}}(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}\right)\\ \leq&\,\text{trace}\left(\left(\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top\hat{U}^{\text{res}}+\lambda I_{p^{\text{res}}}\right)(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}\right)\\ =&\,\text{trace}(I_{p^{\text{res}}})\\ =&p^{\text{res}}.\end{aligned}$$

Under Assumption 3, for all $\eta\in(0,1)$, with probability at least $1-\eta$, (3.C.19) is $O\left(\sqrt{\frac{p^{\text{res}}t\log(p/\eta)}{T_0}}\right)$. To bound the term in (3.C.17) we use the result in Lemma 3.4.3 on the estimation error for the invariant component, since

$$\|\sum_{\tau\in[t-1]}(\hat{U}^{\text{res}})^\top\varphi(X_\tau,a_\tau)\varphi(X_\tau,a_\tau)^\top(\beta^{\text{inv}}-\hat{\beta}^{\text{inv}})\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}$$

$$
\begin{aligned}
&= \|(\tilde{\Sigma}^{\text{res}}_{t-1})^{-\frac{1}{2}} \sum_{\tau\in[t-1]} (\hat{U}^{\text{res}})^\top \varphi(X_\tau, a_\tau)\varphi(X_\tau, a_\tau)^\top (\beta^{\text{inv}} - \hat{\beta}^{\text{inv}})\|_2 \\
&\leq \sqrt{\sum_{\tau\in[t-1]} \|(\tilde{\Sigma}^{\text{res}}_{t-1})^{-\frac{1}{2}} (\hat{U}^{\text{res}})^\top \varphi(X_\tau, a_\tau)\|_2^2} \sqrt{\sum_{\tau\in[t-1]} (\varphi(X_\tau, a_\tau)^\top (\beta^{\text{inv}} - \hat{\beta}^{\text{inv}}))^2} \\
&\leq \sqrt{p^{\text{res}}} \sqrt{t \max_{\tau\in[t-1]} \|\varphi(X_\tau, a_\tau)\|_2^2 \|\beta^{\text{inv}} - \hat{\beta}^{\text{inv}}\|_2^2} \\
&\leq L\|\beta^{\text{inv}} - \hat{\beta}^{\text{inv}}\|_2 \sqrt{p^{\text{res}} t} =: \hat{\rho}^{\text{res}}_3.
\end{aligned}
$$

Taking the union bound, we have that, with probability at least $1 - 2\eta$,

$$
\|\hat{\delta}^{\text{res}}_t - \hat{\Pi}^{\mathcal{S}^{\text{res}}} \delta^{\text{res}}_t\|_{\hat{\Sigma}_{t-1}} \leq \sum_{i=1}^{3} \hat{\rho}^{\text{res}}_i =: \hat{\rho}^{\text{res}}_t(\eta, L, M). \tag{3.C.20}
$$

To conclude the proof of the lemma it is sufficient to replace η with $\eta' := \eta/2$ in all above terms. □

Proof of Theorem 3.4.5. We start by considering a decomposition of the instantaneous regret similar to the one in (3.4.8). Now, the true subspace decomposition is unknown, so $\hat{\mathcal{C}}^\beta$ and $\hat{\mathcal{C}}^\delta_t$ are confidence sets containing, with probability at least $1-\eta$, $\hat{\Pi}^{\mathcal{S}^{\text{inv}}} \beta^{\text{inv}}$ and $\hat{\Pi}^{\mathcal{S}^{\text{res}}} \delta^{\text{res}}_t$, respectively. As before, we denote by $(\bar{\beta}_t, \bar{\delta}_t) = \arg\max_{\beta\in\hat{\mathcal{C}}^\beta, \delta\in\hat{\mathcal{C}}^\delta_t} \varphi(X_t, a_t)^\top (\beta + \delta)$, and by $\bar{\gamma}_t = \bar{\beta}_t + \bar{\delta}_t$ the parameter at which the upper confidence bound for the predicted reward at time t is achieved. Moreover, recall that $\tilde{\Sigma}^{\text{res}}_t = (\hat{U}^{\text{res}})^\top \hat{\Sigma}_t \hat{U}^{\text{res}}$ and $\tilde{\Sigma}^{\text{inv}}_{[-T_0]} = (\hat{U}^{\text{inv}})^\top \hat{\Sigma}_{[-T_0]} \hat{U}^{\text{inv}}$. Then, at time $t \in [T]$, reg_t is such that

$$
\begin{aligned}
\text{reg}_t &:= (\varphi(X_t, a^*_t) - \varphi(X_t, a_t))^\top \gamma_{0,t} \\
&\leq \varphi(X_t, a_t)^\top (\bar{\gamma}_t - \gamma_{0,t}) \\
&= \varphi(X_t, a_t)^\top (\bar{\beta}_t - \beta^{\text{inv}}) + \varphi(X_t, a_t)^\top (\bar{\delta}_t - \delta^{\text{res}}_t) \\
&= \varphi(X_t, a_t)^\top (\bar{\beta}_t - \hat{\Pi}^{\mathcal{S}^{\text{inv}}} \beta^{\text{inv}} + \hat{\Pi}^{\mathcal{S}^{\text{inv}}} \beta^{\text{inv}} - \beta^{\text{inv}}) \\
&+ \varphi(X_t, a_t)^\top (\bar{\delta}_t - \hat{\Pi}^{\mathcal{S}^{\text{res}}} \delta^{\text{res}}_t + \hat{\Pi}^{\mathcal{S}^{\text{res}}} \delta^{\text{res}}_t - \delta^{\text{res}}_t) \\
&\leq \|(\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t)\|_2 \|(\hat{U}^{\text{inv}})^\top \bar{\beta}_t - (\hat{U}^{\text{inv}})^\top \beta^{\text{inv}}\|_2 \\
&+ \|(\hat{U}^{\text{res}})^\top \varphi(X_t, a_t)\|_{((\hat{U}^{\text{res}})^\top \hat{\Sigma}_{t-1} \hat{U}^{\text{res}})^{-1}} \|(\hat{U}^{\text{res}})^\top \bar{\delta}_t - (\hat{U}^{\text{res}})^\top \delta^{\text{res}}_t\|_{(\hat{U}^{\text{res}})^\top \hat{\Sigma}_{t-1} \hat{U}^{\text{res}}} \\
&+ \|\varphi(X_t, a_t)\|_2 \|\gamma_{0,t}\|_2 \Delta\Pi \\
&\leq \|(\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t)\|_2 \tfrac{1}{\sqrt{\lambda_0 T_0}} \|(\hat{U}^{\text{inv}})^\top \bar{\beta}_t - (\hat{U}^{\text{inv}})^\top \beta^{\text{inv}}\|_{\tilde{\Sigma}^{\text{inv}}_{[-T_0]}} \\
&+ \|(\hat{U}^{\text{res}})^\top \varphi(X_t, a_t)\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}} \|(\hat{U}^{\text{res}})^\top \bar{\delta}_t - (\hat{U}^{\text{res}})^\top \delta^{\text{res}}_t\|_{\tilde{\Sigma}^{\text{res}}_{t-1}} \\
&+ \|\varphi(X_t, a_t)\|_2 \|\gamma_{0,t}\|_2 \Delta\Pi \\
&\leq 2\|(\hat{U}^{\text{inv}})^\top \varphi(X_t, a_t)\|_2 \sqrt{\tfrac{\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)}{\lambda_0 T_0}} + 2\|(\hat{U}^{\text{res}})^\top \varphi(X_t, a_t)\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}} \sqrt{\hat{\rho}^{\text{res}}_t(\eta, L, M)} \\
&+ \|\varphi(X_t, a_t)\|_2 \|\gamma_{0,t}\|_2 \Delta\Pi.
\end{aligned}
$$

Lemma 3.4.3 implies that we can define $\hat{\mathcal{C}}^{\beta}$ to contain $\hat{\Pi}^{\mathcal{S}^{\text{inv}}}\beta^{\text{inv}}$ with probability at least $1-\eta$ so that $\sqrt{\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)}$ is

$$O\left(\sqrt{p^{\text{inv}}\log\left(\tfrac{1}{p^{\text{inv}}\lambda_0}\right)+\log(\tfrac{1}{\eta})}\right)+O\left(\sqrt{p^{\text{inv}}\log(\tfrac{p}{\eta})}\right)+O\left(\sqrt{\tfrac{1}{\lambda_0}\log(\tfrac{p}{\eta})}\right).$$

Lemma 3.4.4 implies instead that we can define $\hat{\mathcal{C}}^{\delta}_t$ to contain $\hat{\Pi}^{\mathcal{S}^{\text{res}}}\delta^{\text{res}}_t$ so that $\sqrt{\hat{\rho}^{\text{res}}_t(\eta, L, M)}$ is, for all $\eta \in (0, \frac{1}{2})$ with probability at least $1-2\eta$

$$O\left(\sqrt{p^{\text{res}}\log(1+\tfrac{t}{p^{\text{res}}})+\log(\tfrac{1}{\eta})}\right)+O\left(\sqrt{\tfrac{p^{\text{res}}t}{T_0}}\left(\sqrt{\log(\tfrac{p}{\eta})}+\sqrt{\tfrac{\hat{\rho}^{\text{inv}}_{T_0}(\eta,L,M)}{\lambda_0}}\right)\right). \quad (3.C.21)$$

Moreover, by Assumption 3, we have that, for all $\eta \in (0,1)$, with probability at least $1-\eta$,

$$\|\varphi(X_t, a_t)\|_2 \Delta\Pi \|\gamma_{0,t}\|_2 \leq LM\Delta\Pi$$

is $O(\sqrt{\log(p/\eta)/T_0})$. Taking now the sum over the time horizon T we obtain that

$$\begin{aligned}
\text{Reg}_T &= \sum_{t=1}^{T} \text{reg}_t \\
&\leq \sum_{t=1}^{T}\left(2\|(\hat{U}^{\text{inv}})^\top\varphi(X_t,a_t)\|_2\sqrt{\tfrac{\hat{\rho}^{\text{inv}}_{T_0}(\eta,L,M)}{\lambda_0 T_0}}+\|\varphi(X_t,a_t)\|_2\|\gamma_{0,t}\|_2\Delta\Pi\right) \\
&+ \sum_{t=1}^{T} 2\|(\hat{U}^{\text{res}})^\top\varphi(X_t,a_t)\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}\sqrt{\rho^{\text{res}}_t}.
\end{aligned}$$

We consider these two sums separately. For the first, we obtain that

$$\begin{aligned}
&\sum_{t=1}^{T}\left(2\|(\hat{U}^{\text{inv}})^\top\varphi(X_t,a_t)\|_2\sqrt{\tfrac{\hat{\rho}^{\text{inv}}_{T_0}(\eta,L,M)}{\lambda_0 T_0}}+\|\varphi(X_t,a_t)\|_2\|\gamma_{0,t}\|_2\Delta\Pi\right) \\
\leq& T\left(2L\sqrt{\tfrac{\hat{\rho}^{\text{inv}}_{T_0}(\eta,L,M)}{\lambda_0 T_0}}+LM\Delta\Pi\right).
\end{aligned}$$

From the results above, we obtain that this term is, for all $\eta \in (0, \frac{1}{5})$, with probability at least $1-5\eta$,

$$O\left(\tfrac{T}{\sqrt{T_0\lambda_0}}\sqrt{p^{\text{inv}}+\log(\tfrac{1}{\eta})}\right)+O\left(\tfrac{T}{\sqrt{T_0\lambda_0}}\sqrt{p^{\text{inv}}\log(\tfrac{p}{\eta})}\right)+O\left(\tfrac{T}{\sqrt{T_0\lambda_0}}\sqrt{\max\{\lambda_0,\tfrac{1}{\lambda_0}\}\log(\tfrac{p}{\eta})}\right).$$

For the remaining part of the cumulative regret, we have that

$$\begin{aligned}&\sum_{t=1}^{T} 2\|(\hat{U}^{\text{res}})^\top \varphi(X_t, a_t)\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}} \sqrt{\hat{\rho}^{\text{res}}_t(\eta, L, M)} \\ \leq & 2\sqrt{T\hat{\rho}^{\text{res}}_T(\eta, L, M) \sum_{t=1}^{T} \|(\hat{U}^{\text{res}})^\top \varphi(X_t, a_t)\|^2_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}}}.\end{aligned}$$

As in the oracle case (see the proof of Lemma 3.4.4), we can use (3.C.4) and Lemma 3.6.1 and obtain

$$\sum_{t=1}^{T} 2\|(\hat{U}^{\text{res}})^\top \varphi(X_t, a_t)\|_{(\tilde{\Sigma}^{\text{res}}_{t-1})^{-1}} \sqrt{\hat{\rho}^{\text{res}}_t(\eta, L, M)} \leq C^{\text{res}} \sqrt{\hat{\rho}^{\text{res}}_T(\eta, L, M) T p^{\text{res}} \log\left(1 + \tfrac{TL^2}{\lambda p^{\text{res}}}\right)}.$$

Using (3.C.21), we have that the above term is, for all $\eta \in (0, \frac{1}{2})$, with probability at least $1 - 2\eta$,

$$\begin{aligned}&O\left(p^{\text{res}} \sqrt{T \log\left(1 + \tfrac{TL^2}{\lambda p^{\text{res}}}\right)\left(\log(1 + \tfrac{T}{p^{\text{res}}}) + \log(\tfrac{1}{\eta})\right)}\right) \\ +&O\left(\tfrac{p^{\text{res}} T}{\sqrt{T_0}} \sqrt{\log\left(1 + \tfrac{TL^2}{\lambda p^{\text{res}}}\right)}\left(\sqrt{\log(\tfrac{p}{\eta})} + \sqrt{\tfrac{\hat{\rho}^{\text{inv}}_{T_0}(\eta, L, M)}{\lambda_0}}\right)\right).\end{aligned}$$

As a result, up to logarithmic terms, we obtain that the cumulative regret for the ISD-linUCB algorithm is, for all $\eta \in (0, \frac{1}{11})$, with probability at least $1 - 11\eta$

$$\tilde{O}\left(\sqrt{T}\left(p^{\text{res}} + p^{\text{res}} \sqrt{\tfrac{T}{\lambda_0 T_0}}\left(\sqrt{p^{\text{inv}}} + \sqrt{\max\{\lambda_0, \tfrac{1}{\lambda_0}\}}\right)\right)\right).$$

□

3.D. Lower bound

3.D.1 Lower bound for LinUCB

Lattimore and Szepesvári [2020] show the lower bound $\Omega(p\sqrt{T})$ for the LinUCB algorithm. In the linear model, they consider Gaussian noise with unit variance. Then, for a context-action space $\mathcal{X} \times \mathcal{A}$ and a linear parameter $\gamma \in \mathbb{R}^p$, the regret of a policy is

$$\text{Reg}_T(\mathcal{X} \times \mathcal{A}, \gamma) = T \max_{\varphi \in \mathcal{X} \times \mathcal{A}} \varphi^\top \gamma - \mathbb{E}_\gamma \left[\sum_{t=1}^{T} R_t^{a_t}\right]$$

where the expectation is taken with respect to the measure on the rewards induced by the interaction of the policy with the bandit parametrized by γ. The idea is to find

a worst case instance of the bandit problem, which is characterized by the context-action space and by the parameter γ, with cumulative regret at least of the order of $p\sqrt{T}$. Lattimore and Szepesvári [2020] show the result for two different choices of the context-action space. The first is the hypercube, where $\mathcal{X} \times \mathcal{A} = [-1, 1]^p$. The second choice is the unit sphere, namely $\mathcal{X} \times \mathcal{A} = \{x \in \mathbb{R}^p \mid \|x\|_2 \leq 1\}$. In the first case, the proof is simpler because the value of the context-action feature in one direction is not constrained by the values in the remaining directions: we focus on this proof for now. More in detail, the proof for $\mathcal{X} \times \mathcal{A} = [-1, 1]^p$ shows that there exists a parameter $\gamma \in \Gamma \coloneqq \left\{\pm\frac{1}{\sqrt{T}}\right\}^p$ such that $\mathrm{Reg}_T(\mathcal{X} \times \mathcal{A}, \gamma)$ is $\Omega(p\sqrt{T})$. The proof starts by considering the relative entropy (or KL-divergence) between the probability measures $\mathbb{P}_\gamma, \mathbb{P}_{\gamma'}$ on the rewards, induced by the interaction of a fixed policy, i.e., a fixed sequence of context-action features $\{\varphi(X_t, a_t)\}_{t=1}^T$, with two bandits parametrized by $\gamma, \gamma' \in \Gamma$ respectively (with noise $\epsilon_t \sim \mathcal{N}(0, 1)$), that is,

$$D(\mathbb{P}_\gamma, \mathbb{P}_{\gamma'}) = \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_\gamma[(\varphi(X_t, a_t)^\top (\gamma - \gamma'))^2].$$

Denoting with the subscript i the i-th component of a vector, they then define for $i \in [p]$ and $\gamma \in \Gamma$

$$q_{\gamma_i} \coloneqq \mathbb{P}_\gamma \left(\sum_{t=1}^{T} \mathbb{1}\{\mathrm{sign}(\varphi_i(X_t, a_t) \neq \mathrm{sign}(\gamma_i)\} \geq \frac{T}{2} \right),$$

that is, the probability that the number of times that the sign of the i-th component of the selected context-action feature is not optimal exceeds half of the total learning horizon. For a fixed $i \in [p]$ and $\gamma \in \Gamma$, they further fix γ' as the parameter that equals γ in all its components except for the i-th, which is set to $\gamma'_i = -\gamma_i$. Then, it holds by a property of the relative entropy (Bretangolle-Huber inequality) that

$$q_{\gamma_i} + q_{\gamma'_i} \geq \frac{1}{2} \exp\left(-D(\mathbb{P}_\gamma, \mathbb{P}_{\gamma'})\right) = \frac{1}{2} \exp\left(-\frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_\gamma[(\varphi(X_t, a_t)^\top (\gamma - \gamma'))^2] \right) \geq \frac{1}{2} \exp(-2),$$

where the last inequality follows from the definition of the spaces $\mathcal{A} \times \mathcal{X}$ and Γ. Averaging over all possible parameters $\gamma \in \Gamma$, this implies that

$$\frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \sum_{i \in [p]} q_{\gamma_i} = \frac{1}{|\Gamma|} \sum_{i \in [p]} \sum_{\gamma \in \Gamma} q_{\gamma_i} \geq \frac{1}{|\Gamma|} \sum_{i \in [p]} \frac{|\Gamma|}{2} \frac{1}{2} \exp(-2) = \frac{p}{4} \exp(-2).$$

Finally, this implies that there exists at least one parameter $\gamma \in \Gamma$ such that $\sum_{i \in [p]} q_{\gamma_i} \geq \frac{p}{4} \exp(-2)$. For this choice of γ, the regret is

$$\mathrm{Reg}_T(\mathcal{X} \times \mathcal{A}, \gamma) = \mathbb{E}_\gamma \left[\sum_{t=1}^{T} \sum_{i=1}^{p} (\mathrm{sign}(\gamma_i) - \varphi_i(X_t, a_t)) \gamma_i \right]$$

$$
\begin{aligned}
&\geq \frac{1}{\sqrt{T}} \sum_{i=1}^{p} \mathbb{E}_\gamma \left[\sum_{t=1}^{T} \mathbb{1}\{\operatorname{sign}(\varphi_i(X_t, a_t)) \neq \operatorname{sign}(\gamma_i)\} \right] \\
&\geq \frac{\sqrt{T}}{2} \sum_{i=1}^{p} \mathbb{P}_\gamma \left(\sum_{t=1}^{T} \mathbb{1}\{\operatorname{sign}(\varphi_i(X_t, a_t)) \neq \operatorname{sign}(\gamma_i)\} \geq \frac{T}{2} \right) \\
&= \frac{\sqrt{T}}{2} \sum_{i=1}^{p} q_{\gamma_i} \\
&\geq \frac{\exp -2}{8} p\sqrt{T}.
\end{aligned}
$$

The first equality follows from the fact that $\max_{\varphi \in [-1,1]^p} \varphi^\top \theta = \operatorname{sign}(\theta)^\top \theta$. The first inequality follows from the definition of the space Γ and a case-based analysis on the value of the summation terms. The second inequality uses Markov's inequality. The last equality follows from the choice of γ. This concludes the proof for the lower bound.

3.D.2 Lower bound for ISD with oracle subspaces and oracle invariant component

Consider now the case in which we have oracle access to the invariant subspace decomposition $(\mathcal{S}^{\text{inv}}, \mathcal{S}^{\text{res}})$ and the invariant parameter β^{inv}. We can modify the proof for the standard linear case above by considering a parameter space where the parameters only differ in their residual components, that is, their projection onto $\mathcal{S}^{\text{res}}$, while the invariant component remains fixed. Under this additional constraint, the effective dimension of Γ is reduced from p to p^{res}, leading to the lower bound $\Omega(p^{\text{res}}\sqrt{T})$.

More in detail, we consider again the context-action space $\mathcal{X} \times \mathcal{A} = [-1, 1]^p$. We now define the parameter space as $\Gamma \coloneqq \left\{ \gamma \in \mathbb{R}^p \mid \gamma = \beta^{\text{inv}} + \delta^{\text{res}}, (U^{\text{res}})^\top \delta^{\text{res}} \in \left\{ \pm \frac{1}{\sqrt{T}} \right\}^{p^{\text{res}}} \right\}$. Then, we have that for two parameters $\gamma, \gamma' \in \Gamma$,

$$
D(\mathbb{P}_\gamma, \mathbb{P}_{\gamma'}) = \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_\gamma [(\varphi(X_t, a_t)^\top (\delta^{\text{res}} - \delta'^{\text{res}}))^2].
$$

For all $t \in [T]$ let $\tilde{\varphi}^{\text{res}}(X_t, a_t) \coloneqq (U^{\text{res}})^\top \varphi(X_t, a_t)$ and $\tilde{\delta}^{\text{res}} \coloneqq (U^{\text{res}})^\top \delta^{\text{res}}$. As above, we can define for $i \in [p^{\text{res}}]$ and $\gamma \in \Gamma$

$$
q_{\gamma_i} \coloneqq \mathbb{P}_\gamma \left(\sum_{t=1}^{T} \mathbb{1}\{\operatorname{sign}(\tilde{\varphi}_i^{\text{res}}(X_t, a_t)) \neq \operatorname{sign}(\tilde{\delta}_i^{\text{res}})\} \geq \frac{T}{2} \right).
$$

For a fixed $i \in [p^{\text{res}}]$ and $\gamma \in \Gamma$, define $\gamma' \in \Gamma$ as the parameter obtained from γ by setting $\tilde{\delta}_i'^{\text{res}} = -\tilde{\delta}_i^{\text{res}}$ (and otherwise equal to γ). Then,

$$
q_{\gamma_i} + q_{\gamma_i'} \geq \frac{1}{2} \exp \left(-\frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_\gamma [(\varphi(X_t, a_t)^\top (\delta^{\text{res}} - \delta'^{\text{res}}))^2] \right) \geq \frac{1}{2} \exp(-2)
$$

which again follows from the definition of the spaces $\mathcal{X} \times \mathcal{A}$ and Γ. Taking the average over all possible parameters $\gamma \in \Gamma$, we obtain

$$\frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \sum_{i \in [p^{\text{res}}]} q_{\gamma_i} \geq \frac{1}{|\Gamma|} \sum_{i \in [p^{\text{res}}]} \frac{|\Gamma|}{2} \frac{1}{2} \exp(-2) = p^{\text{res}} \frac{\exp(-2)}{4},$$

meaning that there exists $\gamma \in \Gamma$ such that $\sum_{i \in [p^{\text{res}}]} q_{\gamma_i} \geq p^{\text{res}} \frac{\exp(-2)}{4}$. The last part of the proof follows exactly the same steps as the one in the standard case, leading this time to a regret that is $\Omega(p^{\text{res}} \sqrt{T})$.

Bibliography

Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 127–135, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 3438–3450. Curran Associates, Inc., 2021.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE international conference on computer vision*, pages 769–776, 2013.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

S. Bilaj, S. Dhouib, and S. Maghsudi. Meta learning in bandits within shared affine subspaces. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 523–531. PMLR, 02–04 May 2024.

K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

P. Bühlmann. Invariance, causality and robustness: 2018 Neyman lecture. *Statistical Science*, 35(3):pp. 404–426, 2020. ISSN 08834237, 21688745.

P. Bühlmann and N. Meinshausen. Magging: Maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, 104(1):126–135, 2016.

Y. Chen and P. Bühlmann. Domain adaptation under structural causal models. *Journal of Machine Learning Research*, 22(261):1–80, 2021.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Learning to optimize under non-stationarity. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1079–1087. PMLR, 16–18 Apr 2019.

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

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865, 2017.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pages 355–366, 2008.

L. De Lathauwer. Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications*, 30(3): 1033–1066, 2008.

J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*, volume 38. OUP Oxford, 2012.

J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Stat Interface*, 1(1):179, 2008.

F. Feng, B. Huang, K. Zhang, and S. Magliacane. Factored adaptation for non-stationary reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 31957–31971. Curran Associates, Inc., 2022.

C. Févotte and F. J. Theis. Pivot selection strategies in Jacobi joint block-diagonalization. In *Independent Component Analysis and Signal Separation*, pages 177–184, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), 2014. ISSN 0360-0300.

J. L. Gamella, J. Peters, and P. Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.

N. Gnecco, J. Peters, S. Engelke, and N. Pfister. Boosted control functions: Distribution generalization and invariance in confounded models. *Journal of Machine Learning Research*, 27(46):1–57, 2026.

W. Günther, U. Ninad, and J. Runge. Causal discovery for time series from multiple datasets with latent contexts. In *Uncertainty in Artificial Intelligence*, pages 766–776. PMLR, 2023.

H. W. Gutch and F. J. Theis. Uniqueness of linear factorizations into independent subspaces. *Journal of Multivariate Analysis*, 112:48–62, 2012.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(4):757–779, 1993.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, NY, 2 edition, 2009.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 2012.

Z. Hu and L. J. Hong. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

P. J. Huber. *Robust Statistics*. Wiley New York, 1981.

C. Kausik, K. Tan, and A. Tewari. Leveraging offline data in linear latent contextual bandits. In *International Conference on Machine Learning*, pages 29345–29389. PMLR, 2025.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5815–5826. PMLR, 18–24 Jul 2021.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

M. Lazzaretto, J. Peters, and N. Pfister. Invariant subspace decomposition. *Journal of Machine Learning Research*, 26(95):1–56, 2025.

M. Lazzaretto, J. Peters, and N. Pfister. Invariance-based dynamic regret minimization. *arXiv preprint arXiv:2603.03843*, 2026.

B. Li, Y. Shen, Y. Wang, W. Zhu, C. Reed, D. Li, K. Keutzer, and H. Zhao. Invariant information bottleneck for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7399–7407, 2022.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery.

Z. Li, R. Cai, Z. Yang, H. Huang, G. Chen, Y. Shen, Z. Chen, X. Song, Z. Hao, and K. Zhang. When and how: Learning identifiable latent states for nonstationary time series forecasting. *arXiv preprint arXiv:2402.12767*, 2024.

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. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10846–10856. Curran Associates, Inc., 2018.

N. Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10, 2018.

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

P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

K. Murota, Y. Kanno, M. Kojima, and S. Kojima. A numerical algorithm for block-diagonal decomposition of matrix-algebras with application to semidefinite programming. *Japan Journal of Industrial and Applied Mathematics*, 27(1):125–160, 2010.

V. Y. Nastl and M. Hardt. Do causal predictors generalize better to new domains? In *Advances in Neural Information Processing Systems*, volume 37, pages 31202–31315. Curran Associates, Inc., 2024.

D. Nion. A tensor framework for nonunitary joint block diagonalization. *IEEE Transactions on Signal Processing*, 59(10):4585–4594, 2011.

M. Papini, A. Tirinzoni, M. Restelli, A. Lazaric, and M. Pirotta. Leveraging good representations in linear contextual bandits. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8371–8380. PMLR, 18–24 Jul 2021.

J. Pearl. *Causality*. Cambridge University Press, 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.

J. Peters, D. Janzing, and B. Scholkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

N. Pfister, P. Bühlmann, and J. Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019a.

N. Pfister, S. Weichwald, P. Bühlmann, and B. Schölkopf. Robustifying independent component analysis by adjusting for group-wise stationary noise. *Journal of Machine Learning Research*, 20(147):1–50, 2019b.

N. Pfister, E. G. William, J. Peters, R. Aebersold, and P. Bühlmann. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246, 2021.

Y. Qin, T. Menara, S. Oymak, S. Ching, and F. Pasqualetti. Non-stationary representation learning in sequential linear bandits. *IEEE Open Journal of Control Systems*, 1: 41–56, 2022.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

E. Rosenfeld, P. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

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.

Y. Russac, C. Vernade, and O. Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

S. Saengkyongam, L. Henckel, N. Pfister, and J. Peters. Exploiting independent instruments: Identification and distribution generalization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 18935–18958. PMLR, 17–23 Jul 2022.

S. Saengkyongam, N. Thams, J. Peters, and N. Pfister. Invariant policy learning: A causal perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8606–8620, 2023.

S. Saengkyongam, N. Pfister, P. Klasnja, S. Murphy, and J. Peters. Effect-invariant mechanisms for policy generalization. *Journal of Machine Learning Research*, 25(34): 1–36, 2024.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012.

J. R. Schott. *Matrix analysis for statistics*. John Wiley & Sons, 2016.

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.

G. W. Stewart and J.-G. Sun. *Matrix perturbation theory*. Academic Press, 1990.

P. Stojanov, Z. Li, M. Gong, R. Cai, J. Carbonell, and K. Zhang. Domain adaptation with invariant representation learning: What transformations to learn? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24791–24803. Curran Associates, Inc., 2021.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5):985–1005, 2007.

B. Sun, J. Feng, and K. Saenko. *Correlation Alignment for Unsupervised Domain Adaptation*, pages 153–171. Springer International Publishing, Cham, 2017.

P. Tichavsky and Z. Koldovsky. Algorithms for nonorthogonal approximate joint block-diagonalization. In *2012 Proceedings of the 20th European signal processing conference (EUSIPCO)*, pages 2094–2098. IEEE, 2012.

P. Tichavsky and A. Yeredor. Fast approximate joint diagonalization incorporating weight matrices. *IEEE Transactions on Signal Processing*, 57(3):878–891, 2008.

P. Tichavskỳ, A. Yeredor, and Z. Koldovskỳ. On computation of approximate joint block-diagonalization using ordinary AJD. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 163–171. Springer, 2012.

A. L. Trella, W. Dempsey, F. Doshi-Velez, and S. A. Murphy. Non-stationary latent auto-regressive bandits. *Reinforcement Learning Journal*, 6:765–789, 2025.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

W. Yao, G. Chen, and K. Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 26492–26503. Curran Associates, Inc., 2022.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 819–827, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

K. Zhang, M. Gong, and B. Schoelkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, Feb. 2015.

H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pages 7523–7532. PMLR, 2019.

P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 746–755. PMLR, 26–28 Aug 2020.