Invariant Subspace Decomposition

Margherita Lazzaretto   MALA@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

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

Niklas Pfister[*]   NP@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Editor: Aapo Hyvarinen

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 employed 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.

Keywords: invariance, distribution shift, time adaptation, zero-shot learning, joint block diagonalization.

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.

---

[*]. NP is now at Lakera in Zürich but most of this work was done while NP was at the University of Copenhagen.

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 means 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

$$Y_t = X_t^\top \gamma_{0,t} + \epsilon_t \tag{1}$$

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. 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 $\Delta \mathrm{Var}_t(f) := \mathrm{Var}(Y_t) - \mathrm{Var}(Y_t - f(X_t))$. 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 (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

$$\gamma_{0,t} = \underset{\beta \in \mathbb{R}^p}{\arg \max} \, \Delta \mathrm{Var}_t(\beta). \tag{2}$$

Key to ISD is the decomposition of the explained variance maximization into a time-invariant and a time-dependent part: we show in Theorem 1 that, under some assumptions, for all $t \in \mathbb{N}$, $\gamma_{0,t}$ can be expressed as

$$\gamma_{0,t} = \underbrace{\underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg \max} \, \overline{\Delta \mathrm{Var}}(\beta)}_{=: \beta^{\mathrm{inv}}} + \underbrace{\underset{\beta \in \mathcal{S}^{\mathrm{res}}}{\arg \max} \, \Delta \mathrm{Var}_t(\beta)}_{=: \delta_t^{\mathrm{res}}}, \tag{3}$$

3

where $\overline{\Delta\mathrm{Var}}(\beta) := \frac{1}{n}\sum_{t=1}^{n}\Delta\mathrm{Var}_t(\beta)$, 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 $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ 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, $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$. This in particular implies that the two sub-problems in (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 $Y_t - X_t^{\top}\beta^{\mathrm{inv}}$; the sum $\beta^{\mathrm{inv}} + \delta_t^{\mathrm{res}}$ 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 $\beta^{\mathrm{inv}} + \delta_t^{\mathrm{res}}$. A two-dimensional example of $\gamma_{0,t}$ and its ISD estimates is shown in Figure 1. The same example is presented again in more detail in Example 1 in Section 2.1 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 3 that isolating an invariant component and reducing the dimensionality of the adaptation problem guarantees lower prediction error or, equivalently, higher explained variance compared to to naive methods. An example is provided for a simulated setting in Figure 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 ($\hat{\beta}^{\mathrm{mm}}$) 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 we introduce the ISD framework, starting from the identification of the invariant and residual subspaces for or-
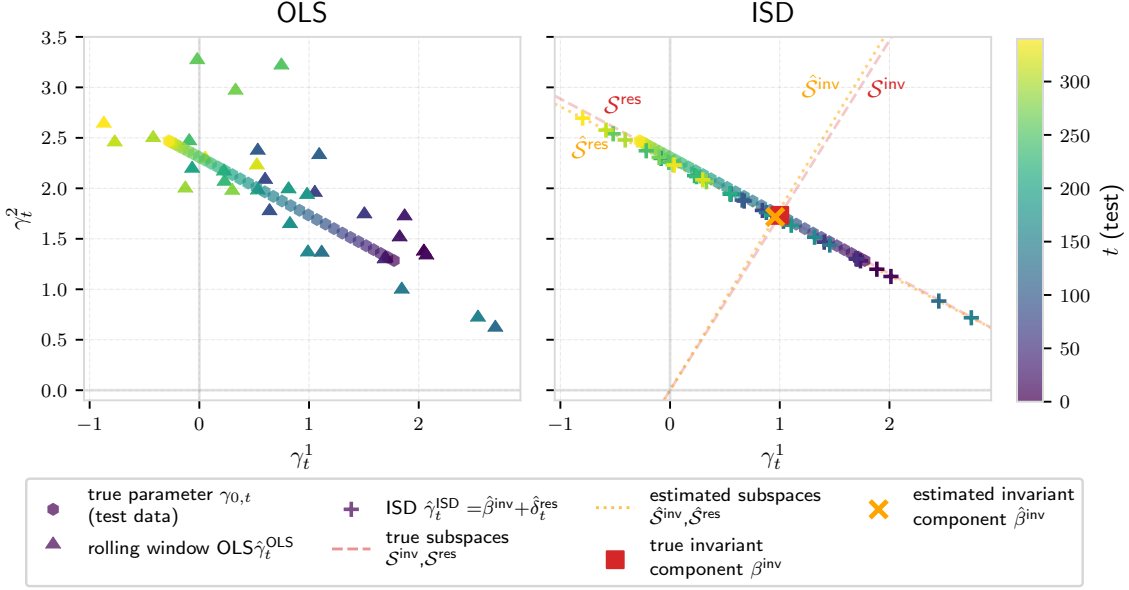
Figure 1: Example of two-dimensional true parameter $\gamma_{0,t}$ varying on a one-dimensional subspace of $\mathbb{R}^2$, and its estimates using ISD (right) compared to rolling window OLS (left). Time is visually encoded using a color map. (Left) True parameter $\gamma_{0,t}$ (hexagons) on 350 test 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 1. 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.

thogonal covariates transformations and showing the explained variance separation (3) in Theorem 1. In Sections 2.2 and 2.3 we define population estimators for the invariant and residual components, solutions to the two sub-problems in (3). We then describe, in Section 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 4 we propose an estimation method for ISD, and provide finite sample generalization guarantees for the proposed estimator. Finally, in Section 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.
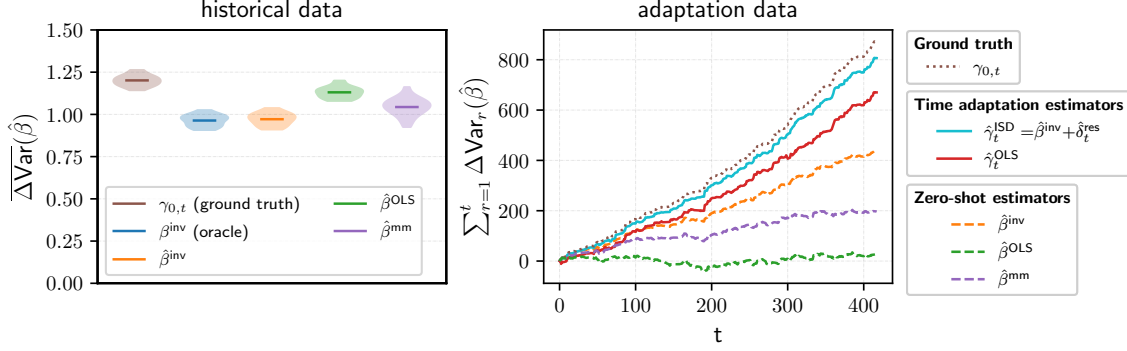
Figure 2: For the data-generating model in Section 5.2, we plot (left) the average explained variance (distribution over 20 runs) obtained at training time (historical data) and (right) the cumulative explained variance obtained testing time (adaptation data) (in one of the 20 runs); in this example, 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 $\hat{\beta}^{\mathrm{inv}}$ of the ISD framework, (iv) the OLS solution $\hat{\beta}^{\mathrm{OLS}}$, (v) the maximin effect $\hat{\beta}^{\mathrm{mm}}$. Starting from $t = 0$ after the observed history, windows of length $3p$ are used to estimate: (vi) the adaptation parameter $\hat{\delta}_t^{\mathrm{res}}$ for $\hat{\beta}^{\mathrm{inv}}$ to obtain the ISD estimate $\hat{\gamma}^{ISD}$ and (vii) the rolling window OLS solution $\hat{\gamma}_t^{\mathrm{OLS}}$. While at training time on historical data the ISD invariant component $\hat{\beta}^{\mathrm{inv}}$ 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 ($\hat{\beta}^{\mathrm{OLS}}, \hat{\beta}^{\mathrm{mm}}$), and can be tuned to new time points to improve on estimators based on adaptation data only ($\hat{\gamma}_t^{\mathrm{OLS}}$).

2. Invariant subspace decomposition

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

Setting 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 (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 $\Sigma_t := \mathrm{Var}(X_t)$ is strictly positive definite. Moreover, for all $n \in \mathbb{N}$, let $[n] := \{1, \ldots, n\}$ and assume we observe n observations $(X_t, Y_t)_{t \in [n]}$ from model (1), which we call historical data. Additionally, let $\mathcal{I}^{\mathrm{ad}} \subseteq \mathbb{N}$ be an interval of consecutive time points with $m := |\mathcal{I}^{\mathrm{ad}}| > p$ and $\min_{t \in \mathcal{I}^{\mathrm{ad}}} t > n$. Assume we observe a second set of observations $(X_t, Y_t)_{t \in \mathcal{I}^{\mathrm{ad}}}$ from the same model but succeeding the historical data, which we call adaptation data. Finally, denote by $t^* > \max_{t \in \mathcal{I}^{\mathrm{ad}}} t$ a time point of interest that occurs after the adaptation data and assume that, for all $t \in \mathcal{I}^{\mathrm{ad}} \cup \{t^*\}$, the quantities $\gamma_{0,t}$, Σ_t and $\mathrm{Var}(\epsilon_t)$ in model (1) are constant (in practice, being approximately constant is sufficient).*

Zero-shot task: observe $(X_t, Y_t)_{t \in [n]}$ and X_{t*} predict Y_{t*}

Adaptation task: observe $(X_t, Y_t)_{t \in [n]}$, $(X_t^{\mathrm{ad}}, Y_t^{\mathrm{ad}})_{t \in \mathcal{I}^{\mathrm{ad}}}$ and X_{t*} predict Y_{t*}
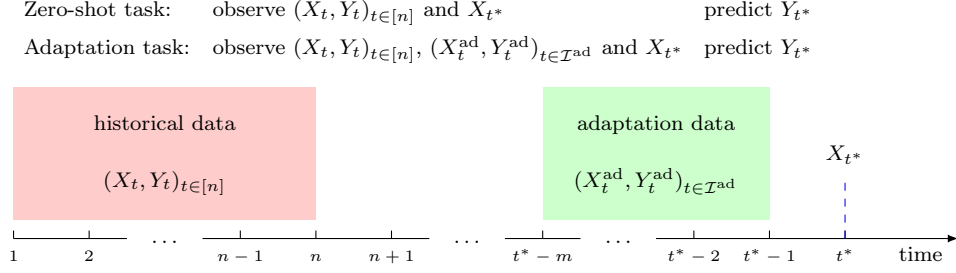

Figure 3: Illustration of the historical and adaptation data and the zero-shot and adaptation tasks.

Our goal is to predict Y_{t*} from X_{t*}. A naive solution is to only consider the adaptation data in $\mathcal{I}^{\mathrm{ad}}$ (see also the analysis in Section 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 $\mathcal{I}^{\mathrm{ad}}$, 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 3. The two tasks correspond to the two sub-problems in (3): the solution to (i) is a time-invariant parameter $\beta^{\mathrm{inv}} \in \mathbb{R}^p$, defined formally below in Section 2.2; the solution to (ii) is an adaptation parameter δ_t^{res} formally defined in Section 2.3 that satisfies $\beta^{\mathrm{inv}} + \delta_t^{\mathrm{res}} = \gamma_{0,t}$.

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

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

Under model (1), the explained variance can be equivalently expressed as

$$
\begin{aligned}
\Delta \mathrm{Var}_t(\beta) &= 2\,\mathrm{Cov}(Y_t, X_t^\top \beta) - \mathrm{Var}(X_t^\top \beta) \\
&= 2\,\mathrm{Cov}(Y_t - X_t^\top \beta, X_t^\top \beta) + \mathrm{Var}(X_t^\top \beta).
\end{aligned}
\tag{4}
$$

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 $t \in \mathbb{N}$, $\Delta \mathrm{Var}_t(\beta) \geq 0$. 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 $\Delta \mathrm{Var}_t(\beta) = \mathrm{Var}(X_t^\top \beta)$ or equivalently, using (4), $\mathrm{Cov}(Y_t - X_t^\top \beta, X_t^\top \beta) = 0$; 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 1 (time-invariance) *We call a parameter $\beta \in \mathbb{R}^p$ time-invariant (over $[n]$) if for all $t \in [n]$,*

$$\mathrm{Cov}(Y_t - X_t^\top \beta, X_t^\top \beta) = 0. \tag{5}$$

---

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 3.

We use this definition, in Section 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 (3). In particular, as we show in Proposition 2, 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 1.

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

Remark 1 (exchanging explained variance with MSPE) *Under model* (1) *maximizing the explained variance at time* $t \in [n]$ *is equivalent to minimizing the MSPE at* t, $\mathrm{MSPE}_t(\beta) :=$ $\mathbb{E}[(Y_t - X_t^\top \beta)^2]$. *More precisely, assuming that all variables have mean zero,*

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

The remaining part of this section is organized as follows. Section 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 we then analyze the time-invariant part of (3), leading to the optimal time-invariant parameter β^{inv}, and show in Proposition 2 (*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 2.3 we analyze the residual part of (3), leading to the optimal parameter δ_t^{res}.

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 (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

$$\mathrm{Cov}(\Pi_{\mathcal{S}_i} X_t, \Pi_{\mathcal{S}_j} X_t) = 0. \tag{6}$$

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 $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ 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 1 *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* $j \in \{1, \ldots, q\}$ *that*

$$\Pi_{\mathcal{S}_j} \gamma_{0,t} = \mathrm{Var}\left(\Pi_{\mathcal{S}_j} X_t\right)^\dagger \mathrm{Cov}\left(\Pi_{\mathcal{S}_j} X_t, Y_t\right) \tag{7}$$

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

8

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.

Lemma 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.[2] Then it holds for all $j \in \{1,\ldots,q\}$ and for all $t \in \mathcal{N}$ that*

$$\arg\max_{\beta \in \mathcal{S}_j} \Delta\mathrm{Var}_t(\beta) = \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^{\dagger} \mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t). \tag{8}$$

Combining Lemmas 1 and 2 implies that

$$\gamma_{0,t} = \sum_{j=1}^q \arg\max_{\beta \in \mathcal{S}_j} \Delta\mathrm{Var}_t(\beta). \tag{9}$$

Therefore, any orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition allows us to split the optimization (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 (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

$$\arg\max_{\beta \in \mathcal{S}} \Delta\mathrm{Var}_t(\beta) = \arg\max_{\beta \in \mathcal{S}} \Delta\mathrm{Var}_s(\beta). \tag{10}$$

For all terms in (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* $\mathcal{S}^{\mathrm{inv}}$ and the *residual subspace* $\mathcal{S}^{\mathrm{res}}$ by

$$\mathcal{S}^{\mathrm{inv}} := \bigoplus_{\substack{j \in \{1,\ldots,q\}: \\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \mathcal{S}_j \quad \text{and} \quad \mathcal{S}^{\mathrm{res}} := \bigoplus_{\substack{j \in \{1,\ldots,q\}: \\ \mathcal{S}_j \text{ not opt-invariant on } [n]}} \mathcal{S}_j, \tag{11}$$

respectively. The invariant subspace $\mathcal{S}^{\mathrm{inv}}$ is opt-invariant (see Lemma 7). Moreover, by definition, it holds that $\mathcal{S}^{\mathrm{res}} = \left(\mathcal{S}^{\mathrm{inv}}\right)^{\perp}$, where $(\cdot)^{\perp}$ denotes the orthogonal complement in $\mathbb{R}^p$, and that $\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ is an orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition (see Lemma 6). 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*

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

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 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

---

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

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 8 and Proposition 1 below). Whenever Assumption 1 holds, the invariant and residual spaces $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$ do not depend on which irreducible orthogonal partition is used in their construction.

Example 1 *This example describes the setting used to generate Figure 1. Consider a 2-dimensional covariate $X_t \in \mathbb{R}^2$, and assume that model (1) is defined as follows. We take, for all $t \in [n]$*

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

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 $\mathrm{Cov}(\Pi_{\mathcal{S}_1}X_t, \Pi_{\mathcal{S}_2}X_t) = \Pi_{\mathcal{S}_1}\Sigma_t\Pi_{\mathcal{S}_2} = \mathbf{0}_{4\times 4}$. 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 8) 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 1 and 2, it holds that

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

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 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 (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 (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 (6) to hold for all $i, j \in \{1, \ldots, q\}$, but only for all $i, j \in \{1, \ldots, q\}$ such that $\mathcal{S}_i \subseteq \mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}_j \subseteq \mathcal{S}^{\mathrm{res}}$.

Theorem 1 *Assume Assumption 2 is satisfied. Let $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$ be defined as in (11) for an arbitrary irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition. Then, it holds for all $t \in \mathbb{N}$ that*

$$\gamma_{0,t} = \underset{\beta\in\mathcal{S}^{\text{inv}}}{\arg\max}\, \overline{\Delta\text{Var}}(\beta) + \underset{\beta\in\mathcal{S}^{\text{res}}}{\arg\max}\, \Delta\text{Var}_t(\beta), \tag{12}$$

where $\overline{\Delta\text{Var}}(\beta) := \frac{1}{n}\sum_{t=1}^{n}\Delta\text{Var}_t(\beta)$. Moreover, if Assumption 1 is satisfied, the separation in (12) is independent of the considered irreducible orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition.

The proof of Theorem 1 can be found in Appendix 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.

2.1.1 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 $\tilde{\Sigma}_t := U^\top \Sigma_t U$, $t \in [n]$, 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 $q_{\max}^U$ be the largest number of such blocks and let $(\tilde{\Sigma}_{t,1})_{t\in[n]}, \ldots, (\tilde{\Sigma}_{t,q_{\max}^U})_{t\in[n]}$ denote the corresponding *common blocks* with dimensions $p_1, \ldots, p_{q_{\max}^U}$ 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 $q_{\max}^U$. 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 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 1   *(i) Let $U \in \mathbb{R}^{p\times p}$ be a joint block diagonalizer of $(\Sigma_t)_{t\in[n]}$. For all $j \in \{1, \ldots, q_{\max}^U\}$, 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 $j \in \{1, \ldots, q_{\max}^U\}$, define*

$$\mathcal{S}_j := \text{span}\{u^k \mid k \in S_j\}.$$

Then, $\{\mathcal{S}_j\}_{j=1}^{q_{\max}^U}$ 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 $\tilde{\Sigma}_t := U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\tilde{\Sigma}_{t,j} =$

$(U^{S_j})^\top \Sigma_t U^{S_j}$, $j \in \{1, \ldots, q\}$ *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 1(i), leads to the same (unique) invariant and residual subspaces defined in (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 $q_{\max}^U$ 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 1 and Lemma 1, by checking for all $j \in \{1, \ldots, q_{\max}^U\}$ whether $\Pi_{\mathcal{S}_j} \gamma_{0,t} = U^{S_j}(U^{S_j})^\top \gamma_{0,t}$ remains constant ($\mathcal{S}_j \subseteq \mathcal{S}^{\mathrm{inv}}$) or not ($\mathcal{S}_j \subseteq \mathcal{S}^{\mathrm{res}}$) 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 1 (continued) *In Example 1 so far, we expressed* $\mathcal{S}_1$ *and* $\mathcal{S}_2$ *in terms of their generating vectors. These can be retrieved using Proposition 1 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 1 and use* $S_1 = \{1\}$ *and* $S_2 = \{2\}$*). In particular, we have that* $\tilde{\Sigma}_t = U^\top \Sigma_t U = \mathrm{diag}(\sigma_{1,t}, \sigma_{2,t})$*: therefore,* $q_{\max}^U = 2$ *and each block has dimension* 1*. Moreover, it holds for all* $t \in [n]$ *that*

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

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

2.2 Invariant component

In Theorem 1 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 $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$. In this section we analyze the result of the first optimization step over the invariant subspace $\mathcal{S}^{\mathrm{inv}}$. To ensure that the space $\mathcal{S}^{\mathrm{inv}}$ is unique, we assume that Assumption 1 is satisfied throughout Section 2.2.

Definition 2 (invariant component) *We denote the parameter that maximizes the explained variance over the invariant subspace by*

$$\beta^{\mathrm{inv}} := \arg\max_{\beta \in \mathcal{S}^{\mathrm{inv}}} \overline{\Delta\mathrm{Var}}(\beta). \tag{13}$$

The parameter β^{inv} corresponds to the pooled OLS solution obtained by regressing Y_t on the projected predictors $\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t$, and can be computed using all observations in $[n]$. The whole procedure to find β^{inv}, by first identifying $\mathcal{S}^{\mathrm{inv}}$ via joint block diagonalization and then using Proposition 2 (i) below, is summarized in Algorithm 1 (see Section 2.4). Proposition 2 summarizes some of the properties of β^{inv}.

Proposition 2 (properties of β^{inv}) *Under Assumption 1, β^{inv} satisfies the following properties:*

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

(ii) β^{inv} is time-invariant over $[n]$, see Definition 1.

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

Definition 1 guarantees, for all $t \in [n]$, that the explained variance for all $\beta \in \mathbb{R}^p$ time-invariant over $[n]$ is $\Delta\mathrm{Var}_t(\beta) = \mathrm{Var}(X^\top \beta) = \beta^\top \Sigma_t \beta$. Point (ii) of Proposition 2 therefore implies that, for all $t \in [n]$, $\Delta\mathrm{Var}_t(\beta^{\mathrm{inv}}) = (\beta^{\mathrm{inv}})^\top \Sigma_t \beta^{\mathrm{inv}}$. 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 $\arg\max_{\beta \in \mathcal{S}^{\mathrm{inv}}} \Delta\mathrm{Var}_t(\beta)$. Moreover, Proposition 2 (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 (12) over $\mathcal{S}^{\mathrm{res}}$. Using U^{inv} as defined in Section 2.1.1, we can express β^{inv} as follows.

$$\beta^{\mathrm{inv}} = U^{\mathrm{inv}}((U^{\mathrm{inv}})^\top \overline{\mathrm{Var}}(X) U^{\mathrm{inv}})^{-1} (U^{\mathrm{inv}})^\top \overline{\mathrm{Cov}}(X, Y) \tag{14}$$

where $\overline{\mathrm{Var}}(X) := \frac{1}{n} \sum_{t=1}^{n} \mathrm{Var}(X_t)$ and $\overline{\mathrm{Cov}}(X, Y) := \frac{1}{n} \sum_{t=1}^{n} \mathrm{Cov}(X_t, Y_t)$. We derive this expression in Lemma 10, and we later use it for estimation in Section 4.2.

Example 1 (continued) *Considering again Example 1, we have that $\mathcal{S}^{\mathrm{inv}} = \mathcal{S}_2$, and therefore the invariant component is given by*

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

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

2.3 Residual component and time adaptation

Under the generalization assumption (Assumption 2), Theorem 1 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 $\mathcal{S}^{\mathrm{res}}$, i.e.,

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

In this section we focus on the second optimization step over $\mathcal{S}^{\mathrm{res}}$. We assume that Assumptions 1 and 2 hold, and for all $t \in \mathbb{N}$ we define the residual component $\delta_t^{\mathrm{res}} := \arg\max_{\beta \in \mathcal{S}^{\mathrm{res}}} \Delta \mathrm{Var}_t(\beta)$. Using (7), (8) and (11), we can express δ_t^{res} as

$$\delta_t^{\mathrm{res}} = \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_t)^{\dagger} \mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_t, Y_t) \tag{15}$$
$$= \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_t)^{\dagger} \mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_t, Y_t - X_t^{\top} \beta^{\mathrm{inv}}).$$

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

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

Moreover, the orthogonal projection matrix onto $\mathcal{S}^{\mathrm{res}}$ is given by $\Pi_{\mathcal{S}^{\mathrm{res}}} = U^{\mathrm{res}} (U^{\mathrm{res}})^{\top}$. Hence, using Lemma 5 we get that

$$\delta_t^{\mathrm{res}} = U^{\mathrm{res}} \mathrm{Var}((U^{\mathrm{res}})^{\top} X_t)^{-1} \mathrm{Cov}((U^{\mathrm{res}})^{\top} X_t, Y_t - X_t^{\top} \beta^{\mathrm{inv}}), \tag{16}$$

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

Example 1 (continued) *In Example 1, we have that for all $t \in [n]$ the residual component δ_t^{res} is given by*

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

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 $U^{\top} \delta_t^{\mathrm{res}} = \begin{bmatrix} 3 - 2t/n, & 0 \end{bmatrix}^{\top}$, 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 $3 - 2t/n \in [1, 3]$.

2.4 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.,

$$\gamma_{0,t} = \beta^{\mathrm{inv}} + \delta_t^{\mathrm{res}}. \tag{17}$$

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

---

Algorithm 1 Population ISD

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

Output: $\beta^{\mathrm{inv}}, \delta_n^{\mathrm{res}}$

1: $(\Sigma_t)_{t \in [n]} \leftarrow \{\mathrm{Var}(X_t)\}_{t \in [n]}$

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

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

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

5: $S^{\mathrm{inv}}, S^{\mathrm{res}} \leftarrow \emptyset$

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

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

8:      **if** $\Pi_{\mathcal{S}_j} \gamma_{0,t}$ constant in $[n]$ **then** $S^{\mathrm{inv}} \leftarrow S^{\mathrm{inv}} \cup S_j$ ▷ see Prop. 2

9:      **else** $S^{\mathrm{res}} \leftarrow S^{\mathrm{res}} \cup S_j$

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

11: $\Pi_{\mathcal{S}^{\mathrm{inv}}} \leftarrow U^{\mathrm{inv}}(U^{\mathrm{inv}})^\top$

12: $\beta^{\mathrm{inv}} \leftarrow \frac{1}{n} \sum_{t=1}^n \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t)^\dagger \mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t, Y_t)$ ▷ see Prop. 2, Lemmas 1,2

13: $\delta_n^{\mathrm{res}} \leftarrow U^{\mathrm{res}} \mathrm{Var}((U^{\mathrm{res}})^\top X_n)^{-1} \mathrm{Cov}((U^{\mathrm{res}})^\top X_n, Y_n - X_n^\top \beta^{\mathrm{inv}})$ ▷ see Eq. (16)

---

the end of Section 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 (17) and therefore the output of Algorithm 1 depends on the irreducible orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition used.

In Appendix C we show that a decomposition of the true time-varying parameter similar to (17) is also obtained when considering $(X_t)_{t \in [n]}$-decorrelating partitions that are not orthogonal, i.e., such that (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 (14) and (16), respectively, where the matrix U is now a non-orthogonal irreducible joint block diagonalizer.

Remark 2 (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 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 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 $\Pi_{\mathcal{S}^{\mathrm{inv}}} \gamma_{0,t}$ on the invariant subspace remains constant across time points. We provide an example in which $\gamma_{0,t}$ is quickly varying in Appendix A.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 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.

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), $\mathcal{S}^{\mathrm{inv}}$ is opt-invariant on $\mathbb{N}$ and we can characterize all possible models defined as in (1) by the possible variations of the true time-varying parameter in the residual space $\mathcal{S}^{\mathrm{res}}$, or, equivalently, by all possible values of $\gamma_{0,t} - \beta^{\mathrm{inv}} \in \mathcal{S}^{\mathrm{res}}$. We then obtain that β^{inv} is worst case optimal in the following sense.

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

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

We further obtain from Proposition 2 (*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 3. In absence of further information on $\delta_{t^*}^{\mathrm{res}}$, Theorem 2 suggests to use an estimate of $\hat{\beta}^{\mathrm{inv}}$ of β^{inv} to predict Y_{t^*}, i.e., $\hat{Y}_{t^*} = X_{t^*}^{\top}\hat{\beta}^{\mathrm{inv}}$.

Remark 3 (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 $\Sigma := \mathrm{Var}(X_t)$ 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

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

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 (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 5.1 and shown in Figure 2). Furthermore, the decomposition allows for time adaptation (see Section 2.3), which would not be possible starting from the maximin effect.

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

$$\inf_{\hat{\gamma}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(X^\top(\hat{\gamma} - \gamma))^2] \geq \sigma^2 \frac{p}{n}, \tag{18}$$

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 (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 $\hat{\delta}_{t^*}$ 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 $\frac{\dim(\mathcal{S}^{\mathrm{inv}})}{n} + \frac{\dim(\mathcal{S}^{\mathrm{res}})}{m}$. In Section 4, we prove that $\hat{\gamma}_{t^*}^{\mathrm{ISD}}$ indeed achieves this bound in Theorem 3. If the invariant subspace is non-degenerate and therefore $\dim(\mathcal{S}^{\mathrm{res}}) < p$, and n is sufficiently large, this implies that $\hat{\gamma}_{t^*}^{\mathrm{ISD}}$ has better finite sample performance than estimators based on the adaptation data alone.

---

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 invariant 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.

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 and on Algorithm 1 described in Section 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 $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$, employing a joint block diagonalization algorithm (Section 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 4.2. In Section 4.3 we then show the advantage of separating the optimization as in (3) to estimate γ_{0,t^*} at the previously unobserved time point t^*, by providing finite sample guarantees for the ISD estimator.

4.1 Estimating the subspace decomposition

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, $k \in \{1, \ldots, K\}$, 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 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), which solves AJD for the set $\{\hat{\Sigma}_k\}_{k=1}^K$, i.e., it is such that for all $k \in \{1, \ldots, K\}$ 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 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 $k \in \{1, \ldots, K\}$ 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 $\{\hat{\mathcal{S}}_j\}_{j=1}^{q_{\max}^{\hat{U}}}$ the estimated irreducible orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition.

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 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 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 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 $j \in \{1, \ldots, q_{\max}^{\hat{U}}\}$, we take the sets of indices S_j as defined in Proposition 1, 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 $\hat{\mathcal{S}}_j$ of the estimated irreducible orthogonal partition. It follows from Lemma 1 that we can find the opt-invariant subspaces by checking for all $j \in \{1, \ldots, q_{\max}^{\hat{U}}\}$ whether $\Pi_{\hat{\mathcal{S}}_j} \hat{\gamma}_k$ remains approximately constant for $k \in \{1, \ldots, K\}$. To do so, let $\hat{\gamma} := (\sum_{k=1}^K \text{Var}(\hat{\gamma}_k)^{-1})^{-1} \sum_{k=1}^K \text{Var}(\hat{\gamma}_k)^{-1}\hat{\gamma}_k$ be the average of the estimated regression coefficients inversely weighted by their variance. We further use that, by Lemma 9, 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 (5). This approach is motivated by Proposition 2 *(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 (5), we can equivalently consider the correlation in place of the covariance, that is, $\text{corr}(Y_t - X_t^\top \beta, X_t^\top \beta) = 0$: an estimate of this correlation allows us to obtain a normalized measure of (5) that is comparable across different experiments. Formally, we consider for all $k \in \{1, \ldots, K\}$ and for all $j \in \{1, \ldots, q_{\max}^{\hat{U}}\}$

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

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

$$\frac{1}{K} \sum_{k=1}^K \left|\hat{c}_k^j\right| \leq \lambda \tag{19}$$

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 A.4 in the appendix). An estimator of the invariant and residual subspaces is then given by

$$\hat{\mathcal{S}}^{\text{inv}} = \bigoplus_{\substack{j \in \{1, \ldots, q_{\max}^{\hat{U}}\}: \\ \text{(19) is satisfied}}} \hat{\mathcal{S}}_j, \quad \hat{\mathcal{S}}^{\text{res}} = \bigoplus_{\substack{j \in \{1, \ldots, q_{\max}^{\hat{U}}\}: \\ \text{(19) is not satisfied}}} \hat{\mathcal{S}}_j, \tag{20}$$

where we approximate opt-invariance with the inequality (19) being satisfied.

4.2 Estimating the invariant and residual components

Let $\hat{U}^{\text{inv}}$ and $\hat{U}^{\text{res}}$ be the submatrices of $\hat{U}$ whose columns span $\hat{\mathcal{S}}^{\text{inv}}$ and $\hat{\mathcal{S}}^{\text{res}}$, respectively. We propose to estimate β^{inv} using the following plug-in estimator for (14)

$$\hat{\beta}^{\text{inv}} := \hat{U}^{\text{inv}}((\hat{U}^{\text{inv}})^\top \mathbf{X}^\top \mathbf{X}\hat{U}^{\text{inv}})^{-1}(\hat{U}^{\text{inv}})^\top \mathbf{X}^\top \mathbf{Y}. \tag{21}$$

---

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 B in the appendix.

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 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 (16) we obtain the following plug-in estimator for $\delta_{t^*}^{\mathrm{res}}$

$$\hat{\delta}_{t^*}^{\mathrm{res}} := \hat{U}^{\mathrm{res}}((\hat{U}^{\mathrm{res}})^\top (\mathbf{X}^{\mathrm{ad}})^\top \mathbf{X}^{\mathrm{ad}} \hat{U}^{\mathrm{res}})^{-1} (\hat{U}^{\mathrm{res}})^\top (\mathbf{X}^{\mathrm{ad}})^\top (\mathbf{Y}^{\mathrm{ad}} - \mathbf{X}^{\mathrm{ad}} \hat{\beta}^{\mathrm{inv}}). \tag{22}$$

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

$$\hat{\gamma}_{t^*}^{\mathrm{ISD}} := \hat{\beta}^{\mathrm{inv}} + \hat{\delta}_{t^*}^{\mathrm{res}}. \tag{23}$$

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 B.

Example 1 (continued) *Assume that in Example 1 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 $\Pi_{\mathcal{S}_2}\gamma_{0,t}^{\mathrm{ad}} = \Pi_{\mathcal{S}_2}\gamma_{0,t} = \beta^{\mathrm{inv}}$, 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

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

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 $\mathcal{I}^{\mathrm{ad}} = \{t^* - m, \dots, t^* - 1\}$ of length $m = 16$. After estimating β^{inv} on historical data, we use the adaptation data to estimate $\delta_{t^*}^{\mathrm{res}}$ and the OLS solution $\gamma_{t^*}^{\mathrm{OLS}}$. We repeat this online step 350 times (each time increasing t^* by one). Figure 1 shows the results of this experiment.

4.3 Finite sample generalization guarantee

Considering the setting described in Section 2.3, we now compare the ISD estimator in (23) with the OLS estimator computed on $\mathcal{I}^{\mathrm{ad}}$, i.e., $\hat{\gamma}_{t^*}^{\mathrm{OLS}} := ((\mathbf{X}^{\mathrm{ad}})^\top \mathbf{X}^{\mathrm{ad}})^{-1} (\mathbf{X}^{\mathrm{ad}})^\top \mathbf{Y}^{\mathrm{ad}}$. We assume that we are given the (oracle) subspaces $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$, and consider the expected explained variance at t^* of $\hat{\gamma}_{t^*}^{\mathrm{ISD}}$ and $\hat{\gamma}_{t^*}^{\mathrm{OLS}}$. More in detail, the expected explained variance at t^* of an arbitrary estimator $\hat{\gamma}$ of γ_{0,t^*} is given by

$$\mathbb{E}[\Delta \mathrm{Var}_{t^*}(\hat{\gamma})] := \mathbb{E}[\mathrm{Var}(Y_{t^*}) - \mathrm{Var}(Y_{t^*} - X_{t^*}^\top \hat{\gamma} \mid \hat{\gamma})]. \tag{24}$$

Evaluating (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 (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 1). 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 3 *Assume Assumption 2 and that, in model* (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*

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

From the proof of Theorem 3 it follows that C_{res} can be chosen close to 1 if we only consider sufficiently large m. Moreover, because $\text{MSPE}(\hat{\gamma}_{t^*}^{\text{OLS}}) - \text{MSPE}(\hat{\gamma}_{t^*}^{\text{ISD}}) = \mathbb{E}[\Delta\text{Var}_{t^*}(\hat{\gamma}_{t^*}^{\text{ISD}})] - \mathbb{E}[\Delta\text{Var}_{t^*}(\hat{\gamma}_{t^*}^{\text{OLS}})]$, Theorem 3 implies that if $\dim(\mathcal{S}^{\text{inv}}) \geq 1$ and n sufficiently large then

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

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 $\mathcal{I}^{\text{ad}}$: there is a higher gain in using $\hat{\gamma}_{t^*}^{\text{ISD}}$ instead of $\hat{\gamma}_{t^*}^{\text{OLS}}$ if the dimension of $\mathcal{S}^{\text{inv}}$ is large and only a small amount of time-points are available are available in the adaptation data.

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 $\hat{\beta}^{\text{inv}}$ for increasing sample size n of the historical data. The second simulation experiment compares the predictive accuracy of $\hat{\gamma}_t^{\text{ISD}}$ and $\hat{\gamma}_t^{\text{OLS}}$ 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 3.

In both simulation experiments we let the dimension of the covariates be $p = 10$, and $\dim(\mathcal{S}^{\text{inv}}) = 7$, $\dim(\mathcal{S}^{\text{res}}) = 3$, 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 $\sigma^2_{\epsilon_t} = 0.64$. The dimensions of the subspaces, $\mathcal{S}^{\text{inv}}$ and $\mathcal{S}^{\text{res}}$, and the true time-varying parameter $\gamma_{0,t}$ are chosen to ensure that, in the historical data, $X_t^\top \beta^{\text{inv}}$ and $X_t^\top \delta_t^{\text{res}}$ explain approximately half of the variance of Y_t each. This choice allows for better visualization in the described experiments.

In Section 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 is taken from the `smooth_polarizers` experiment in the `lt_walks_v1` dataset, available at `https://github.com/juangamella/causal-chamber`. The 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).

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 4). Figure 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 1 (we refer to this window as test data). We compare the predictive performance of the parameter $\hat{\beta}^{\text{inv}}$ on
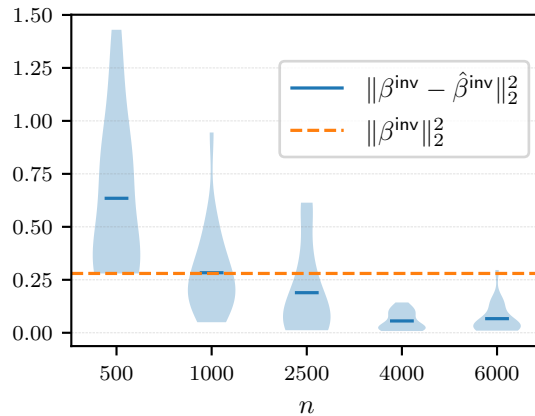

Figure 4: MSE of $\hat{\beta}^{\text{inv}}$ for increasing size of the historical data n (see Section 5.1). For larger values of n, the estimation of the invariant subspace decomposition becomes more precise and leads to smaller errors in the estimated invariant component $\hat{\beta}^{\text{inv}}$.
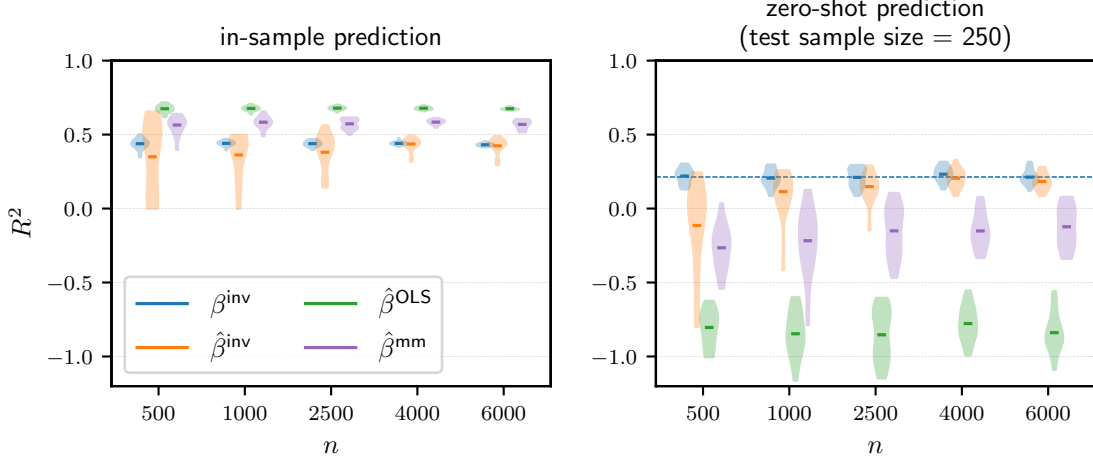
Figure 5: Normalized explained variance (R^2) by $\hat{\beta}^{\mathrm{inv}}$ and comparison with β^{inv}, $\hat{\beta}^{\mathrm{mm}}$ and $\hat{\beta}^{\mathrm{OLS}}$: training (historical data, left) and zero-shot generalization (test data, right), for different sizes n of the historical data (see Section 5.1). The dashed line indicates the population value of the (normalized) explained variance by β^{inv}.

the historical and test data with that of the oracle invariant parameter β^{inv}, the maximin effect $\hat{\beta}^{\mathrm{mm}}$ (computed using the magging estimator proposed by Bühlmann and Meinshausen, 2015), and the OLS solution $\hat{\beta}^{\mathrm{OLS}}$, both computed using the historical data. We show in Figure 5 the results in terms of the R^2 coefficient, given by $R^2 = \frac{\sum_{t=1}^{n}(\widehat{\mathrm{Var}}(Y_t) - \widehat{\mathrm{Var}}(Y_t - X_t^\top \hat{\beta}))}{\sum_{t=1}^{n} \widehat{\mathrm{Var}}(Y_t)}$.

Figure 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 $\hat{\beta}^{\mathrm{inv}}$. Using $\hat{\beta}^{\mathrm{OLS}}$ or $\hat{\beta}^{\mathrm{mm}}$ 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 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 $\mathcal{S}^{\mathrm{inv}}$, to avoid including part of the residual subspace into $\hat{\mathcal{S}}^{\mathrm{inv}}$. 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 4 and 5.

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 $\hat{\beta}^{\mathrm{inv}}$, 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
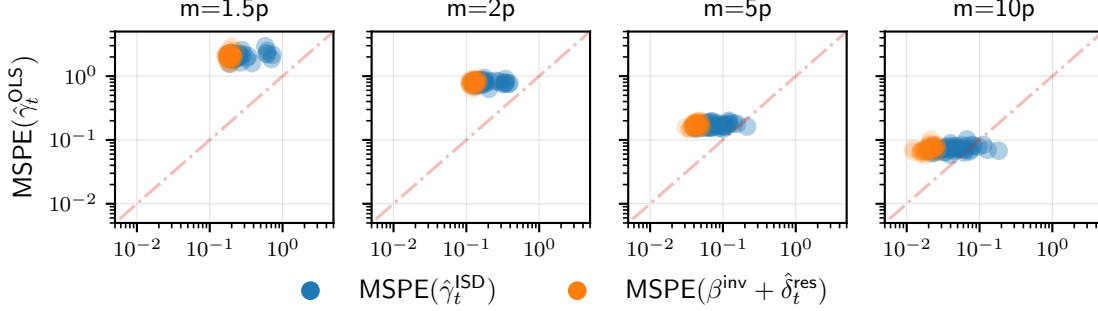
Figure 6: MSPE comparison: $\hat{\gamma}_t^{\mathrm{ISD}}$ (blue dots) vs. $\hat{\gamma}_t^{\mathrm{OLS}}$ for $p = 10$ and various adaptation window lengths m (see Section 5.2). The ISD estimator achieves lower MSPE than the OLS for smaller sizes m of the adaptation window, while the two estimators become comparable for increasing m. The orange dots show the MSPE of β^{inv}(oracle) $+ \hat{\delta}_t^{\mathrm{res}}$ vs. $\hat{\gamma}_t^{\mathrm{OLS}}$: if the subspace decomposition is known, then the ISD always achieves lower MSPE than the OLS.

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 $\hat{\delta}_t^{\mathrm{res}}$ and the OLS solution $\hat{\gamma}_t^{\mathrm{OLS}}$. We then compute the squared prediction error of $\hat{\gamma}_t^{\mathrm{ISD}} = \hat{\beta}^{\mathrm{inv}} + \hat{\delta}_t^{\mathrm{res}}$ and $\hat{\gamma}_t^{\mathrm{OLS}}$ 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 $\hat{\gamma}_t^{\mathrm{OLS}}$ against $\hat{\gamma}_t^{\mathrm{ISD}}$. The result is shown in Figure 6, and empirically supports Theorem 3, in particular that the difference in the MSPEs of the OLS and ISD estimators is proportional to the ratio $\frac{\dim(\mathcal{S}^{\mathrm{inv}})}{m}$ 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 A.5 in the Appendix). The ISD framework is particularly helpful in scenarios in which the size of the available adaptation window is small (two first plots from the left in Figure 6). Indeed, from Theorem 3 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 $\hat{\delta}_t^{\mathrm{res}}$ for small lengths m of the adaptation window where $\dim(\mathcal{S}^{\mathrm{res}}) < m < p$ and OLS is not feasible.

We run a similar experiment to show (Figure 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 $\hat{\delta}_t^{\mathrm{res}}$ and $\hat{\gamma}_t^{\mathrm{OLS}}$ on a rolling adaptation window of size $m = 3p$. The plot in Figure 7 shows that, on average, the ISD framework, by exploiting invariance properties in the observed data, allows us to accurately explain the
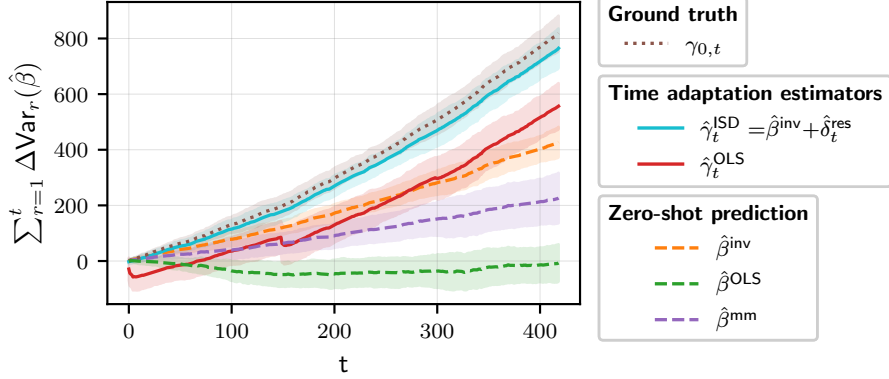
Figure 7: Average cumulative explained variance on the adaptation data and one standard deviation intervals (over 20 runs) by the true time-varying parameter $\gamma_{0,t}$ and various estimators (see Section 5.2). When few data points are available for adaptation at time t, the explained variance of 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.

variance of the response by using small windows for time adaptation, significantly improving on the OLS solution in the same time windows.

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 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 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.
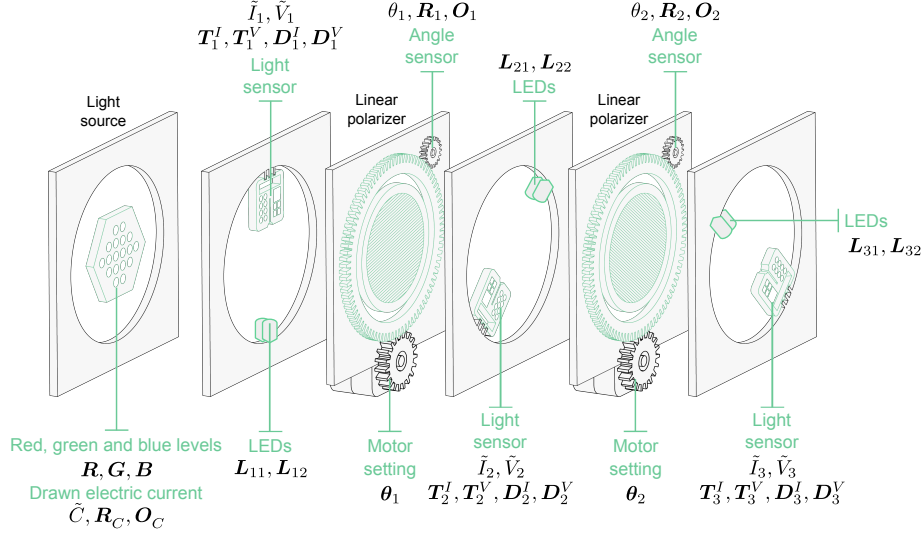
25

Figure 8: Illustration of the light tunnel, see Section 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 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 5.3.

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, 2015) and the OLS solution $\hat{\beta}^{\text{OLS}}$ on historical data. We then use these estimated parameters for zero-shot prediction on test data. Table 1 shows the R^2 coefficient, defined as in Section 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 expected: since the RGB light source is stronger than the two LEDs at the end of the tunnel, most of the variance
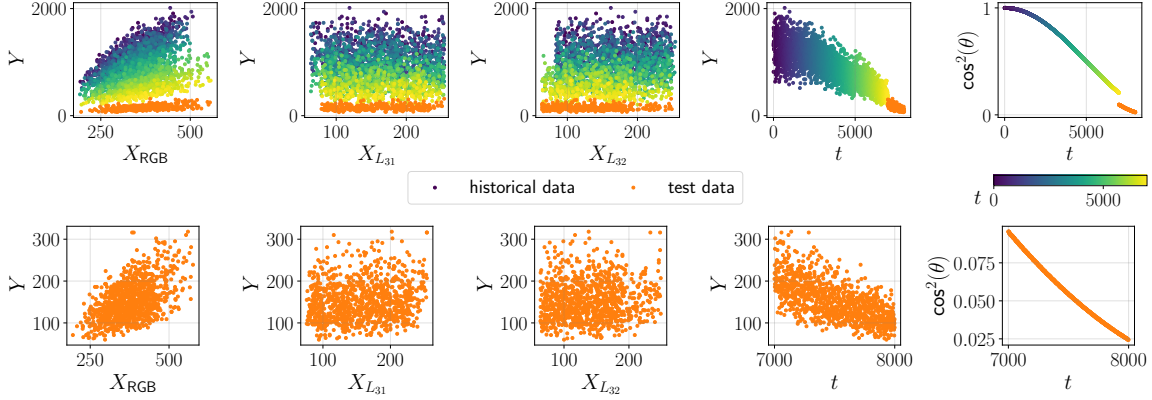
26

Figure 9: Dataset of the experiment discussed in Section 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).
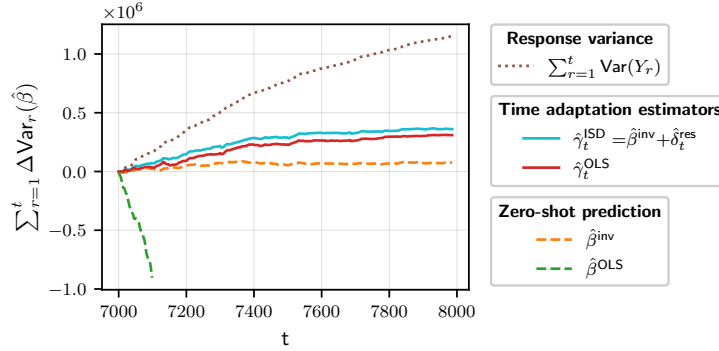

Figure 10: Cumulative explained variance on the adaptation data by the time adaptation estimators $\hat{\gamma}^{\mathrm{ISD}}$ and $\hat{\gamma}^{\mathrm{OLS}}$ and by the zero-shot predictors $\hat{\beta}^{\mathrm{inv}}$ and $\hat{\beta}^{\mathrm{OLS}}$; see Section 5.3.

in Y is explained by the non-invariant component X_{RGB} (see Figure 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 $\mathrm{corr}(X_{\mathrm{RGB}}, X_{L_{31}}) = 0.004$ with p-value 0.723 and $\mathrm{corr}(X_{\mathrm{RGB}}, X_{L_{32}}) = -0.107$ 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 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 ($\dim(\mathcal{S}^{\mathrm{inv}}) = 1$) and by Theorem 3, we expect only a small improvement.

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 $\mathcal{S}^{\mathrm{inv}}$ and a residual subspace $\mathcal{S}^{\mathrm{res}}$, such that the maximizer of the explained variance over $\mathcal{S}^{\mathrm{inv}}$ is time-invariant. The estimation of the invariant component β^{inv} on a large historical dataset and the reduced dimensionality of $\mathcal{S}^{\mathrm{res}}$ 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. For instance, Eastwood et al. (2023) show how to adapt invariant nonlinear features to heterogeneous environments in classification tasks: extending similar ideas to the regression setting considered here could be of interest. Future work may further study how to incorporate the ISD framework in specific applied settings such as contextual bandits.

Acknowledgments

References

K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

P. Bühlmann and N. Meinshausen. Magging: maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, 104(1):126–135, 2015.

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. Durbin and S. J. Koopman. *Time series analysis by state space methods*, volume 38. OUP Oxford, 2012.

C. Eastwood, S. Singh, A. L. Nicolicioiu, M. Vlastelica Pogančić, J. von Kügelgen, and B. Schölkopf. Spuriosity didn't kill the classifier: Using invariant predictions to harness spurious features. *Advances in Neural Information Processing Systems*, 36:18291–18324, 2023.

J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179, 2008.

F. Feng, B. Huang, K. Zhang, and S. Magliacane. Factored adaptation for non-stationary reinforcement learning. *Advances in Neural Information Processing Systems*, 35:31957–31971, 2022.

C. Févotte and F. J. Theis. Pivot selection strategies in Jacobi joint block-diagonalization. In *International Conference on Independent Component Analysis and Signal Separation*, pages 177–184. Springer, 2007.

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, Jan 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00964-x. URL `https://doi.org/10.1038/s42256-024-00964-x`.

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.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

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.

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

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.

D. Nion. A tensor framework for nonunitary joint block diagonalization. *IEEE Transactions on Signal Processing*, 59(10):4585–4594, 2011.

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

N. Pfister, 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.

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.

J. R. Schott. *Matrix analysis for statistics*. John Wiley & Sons, 2016.

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.

B. Sun, J. Feng, and K. Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*, pages 153–171. Springer, 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.

W. Yao, G. Chen, and K. Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 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.

Appendix A. Supporting examples and remarks

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 1) 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.

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

We repeat the same simulation described in Section 5.1 but now consider non-smooth variations of $\gamma_{0,t}$. More specifically, the only difference from the simulation described in Section 5.1 is that we let the 3 time-varying entries of $\gamma_{0,t}$ (prior to its rotation by U) 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
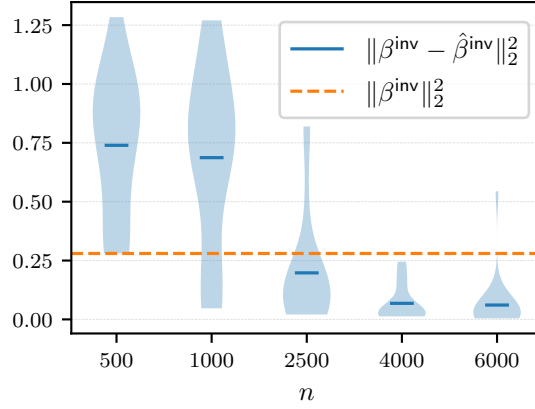
Figure 11: MSE of $\hat{\beta}^{\text{inv}}$ for increasing size of the historical data n. For larger values of n, the estimation of the invariant subspace decomposition becomes more precise and leads to smaller errors in the estimated invariant component $\hat{\beta}^{\text{inv}}$.
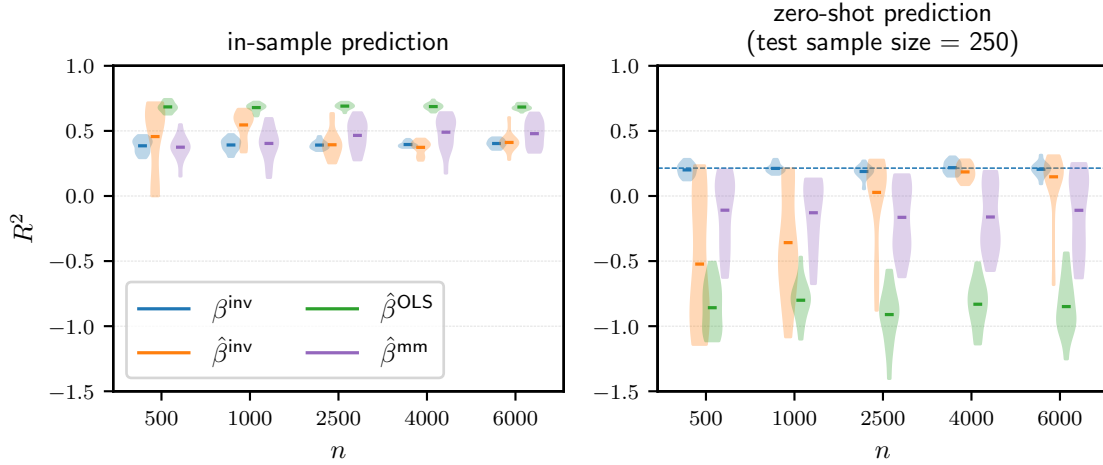
Figure 12: Normalized explained variance (R^2) by $\hat{\beta}^{\text{inv}}$ and comparison with β^{inv}, $\hat{\beta}^{\text{mm}}$ and $\hat{\beta}^{\text{OLS}}$: training (historical data, left) and zero-shot generalization (test data, right), for different sizes n of the historical data. The dashed line indicates the population value of the (normalized) explained variance by β^{inv}. In the historical data, $\gamma_{0,t}$ is quickly varying.

data (which is generated as in Section 5.1). This experiment supports Remark 2, showing that we do not need assumptions on the type of changes in $\gamma_{0,t}$ in the historical data.

In particular, Figure 11 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 12 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 $\hat{\beta}^{\mathrm{inv}}$ remains positive even for values of the test data that lie outside of the observed support.

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; Tichavský 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 Tichavský 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 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

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

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

34

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 $\hat{U} = VP^*$.

A.4 Threshold selection for opt-invariant subspaces

In the simulations, we select the threshold λ in (19) by cross-validation. More in detail, we define the grid of possible thresholds λ by

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

where, for all $j \in \{1, \ldots, q_{\max}^{\hat{U}}\}$ and $k \in \{1, \ldots, K\}$, $\hat{c}_k^j := \widehat{\mathrm{Corr}}(\mathbf{Y}_k - \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma}), \mathbf{X}_k(\Pi_{\hat{\mathcal{S}}_j}\hat{\gamma}))$. We then split the historical data into $L = 10$ disjoint blocks of observations, and for all $j \in \{1, \ldots, J\}$ 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 $\ell \in \{1, \ldots, L\}$, we compute an estimate for the invariant component as in (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 (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 $\overline{\Delta\mathrm{Var}}_\ell^\lambda := \frac{1}{|\mathcal{I}_\ell|} \sum_{t^* \in \mathcal{I}_\ell} \widehat{\Delta\mathrm{Var}}_{t^*}(\lambda)$. For all $\lambda \in \Lambda$, we denote the average explained variance over the L folds by $\overline{\Delta\mathrm{Var}}^\lambda := \frac{1}{L} \sum_{\ell=1}^L \overline{\Delta\mathrm{Var}}_\ell^\lambda$ and the standard error (across the L folds) of such explained variance as $\mathrm{se}(\overline{\Delta\mathrm{Var}}^\lambda) := \frac{1}{L} \sqrt{\sum_{\ell=1}^L (\overline{\Delta\mathrm{Var}}_\ell^\lambda - \overline{\Delta\mathrm{Var}}^\lambda)^2}$. Moreover, let $\lambda^{\max} := \arg\max_{\lambda \in \Lambda} \overline{\Delta\mathrm{Var}}^\lambda$. Then, we choose the optimal threshold as

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

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

A.5 Further simulation details: MSPE comparison

In Section 5.2 we present a simulated experiment in which we compare the ISD estimator and the OLS estimator on the time adaptation task. Figure 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 3, we show in Figure 13 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 3, 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 14 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 3.
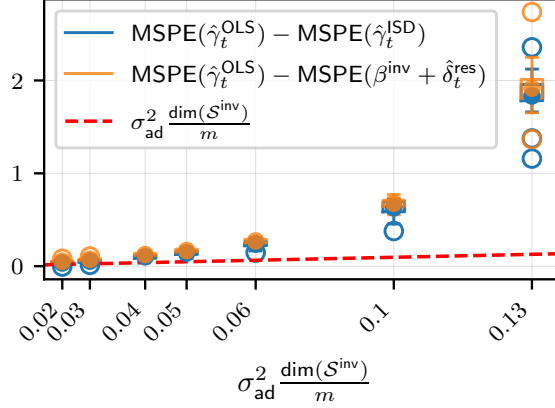
Figure 13: Difference in the MSPE of $\hat{\gamma}_t^{\text{OLS}}$ and $\hat{\gamma}_t^{\text{ISD}}$ (and an oracle version of $\hat{\gamma}_t^{\text{ISD}}$ based on the true β^{inv}) with respect to $\sigma_{\text{ad}}^2 \frac{\dim(\mathcal{S}^{\text{inv}})}{m}$ for different values of m. The computed 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 3. The filled dots in the boxplots show the mean over 20 runs (while the empty dots represent the outliers).

Figure 14: 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 3. The filled dots in the boxplots show the mean over 20 runs (while the empty dots represent the outliers).
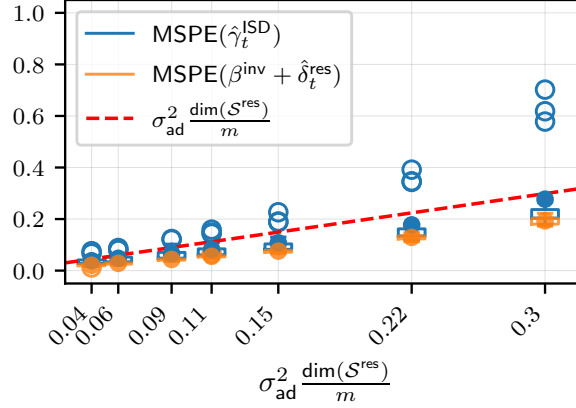
Appendix B. ISD estimation algorithm

We provide here the pseudocode summarizing the ISD procedure described in Section 4. The algorithm includes the estimation of the intercept, that is, it considers for all $t \in \mathbb{N}$ the model

$$Y_t = \gamma_{0,t}^0 + X_t^\top \gamma_{0,t} + \epsilon_t \qquad (25)$$

satisfying the assumptions of Setting 1 but with the addition of $\gamma_{0,t}^0 \in \mathbb{R}$. In the 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 $\hat{\gamma}_{t^*}^0$ 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 4). In case (b), the intercept can instead be estimated simultaneously to the residual component $\hat{\delta}_t^{\text{res}}$: 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 4 *Recall that the population OLS solution for the linear model* (25) *at time t can be found by adding a 1 to the vector X_t and solving*

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

37

The problem has closed form solution

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

Algorithm 2 ISD: estimation
___

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

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

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

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

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

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

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

6: $\hat{U}, \{\hat{S}_j\}_{j=1}^q \leftarrow \mathrm{approxIrreducibleJointBlockDiagonalizer}((\hat{\Sigma}_t)_{t \in [n]})$ $\triangleright$ see Sec. 4.1.1

7: $\triangleright$ Find the opt-invariant subspaces to estimate $\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}$:

8: $\hat{S}^{\mathrm{inv}}, \hat{S}^{\mathrm{res}} \leftarrow \emptyset$

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

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

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

12:      **if** $\frac{1}{K} \sum_{k=1}^K \left| \hat{c}^j_k \right| \le \lambda$ **then** $\hat{S}^{\mathrm{inv}} \leftarrow \hat{S}^{\mathrm{inv}} \cup \hat{S}_j$ $\triangleright$ see Eq. (19)

13:      **else** $\hat{S}^{\mathrm{res}} \leftarrow \hat{S}^{\mathrm{res}} \cup \hat{S}_j$

14: $\triangleright$ Invariant component estimation:

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

16: $\hat{\beta}^{\mathrm{inv}} \leftarrow \hat{U}^{\mathrm{inv}} \widehat{\operatorname{Var}}(\mathbf{X}\hat{U}^{\mathrm{inv}})^{-1} \widehat{\operatorname{Cov}}(\mathbf{X}\hat{U}^{\mathrm{inv}}, \mathbf{Y})$ $\triangleright$ see Eq. (21)

17: $\triangleright$ Adaptation step:

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

19: $\hat{\delta}^{\mathrm{res}}_{t^*} \leftarrow \hat{U}^{\mathrm{res}} \widehat{\operatorname{Var}}(\mathbf{X}^{\mathrm{ad}}\hat{U}^{\mathrm{res}})^{-1} \widehat{\operatorname{Cov}}(\mathbf{X}^{\mathrm{ad}}\hat{U}^{\mathrm{res}}, \mathbf{Y}^{\mathrm{ad}} - \mathbf{X}^{\mathrm{ad}}\hat{\beta}^{\mathrm{inv}})$ $\triangleright$ see Eq. (22)

20: $\triangleright$ Intercept estimation:

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

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

23: $\hat{\gamma}_{t^*} \leftarrow \hat{\beta}^{\mathrm{inv}} + \hat{\delta}^{\mathrm{res}}_{t^*}$ $\triangleright$ see Eq. (23)
___

Appendix C. Extension to non-orthogonal subspaces

In Section 2.1 we have defined an orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition as a collection $\{\mathcal{S}_j\}_{j \in \{1,\ldots,q\}}$ of pairwise orthogonal linear subspaces of $\mathbb{R}^p$ satisfying (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 (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 but where the subspaces in the partition are not necessarily orthogonal. To do so, we define a collection of (not necessarily orthogonal) linear subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_q \subseteq \mathbb{R}^p$ with $\bigoplus_{j=1}^q \mathcal{S}_j = \mathbb{R}^p$ and satisfying (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 2.1.1 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 $j \in \{1, \ldots, q\}$, the columns of U^{S_j} are orthonormal vectors. A version of Proposition 1 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 2.1.1, and we additionally define the matrix $W := U^{-\top}$.

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 $\bigoplus_{i \in \{1,\ldots,q\} \setminus \{j\}} \mathcal{S}_i$: 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 $j \in \{1, \ldots, q\}$ 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 2.1, 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 1 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 $\mathcal{S}^{\mathrm{inv}}$ and residual subspace $\mathcal{S}^{\mathrm{res}}$ as

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

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

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

We show in the following proposition that the expressions (14) and (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 3 *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 $j \in \{1, \ldots, q\}$

$$P_{\mathcal{S}_j | \mathcal{S}_{-j}} \gamma_{0,t} = U^{S_j} \operatorname{Var}((U^{S_j})^\top X_t)^{-1} \operatorname{Cov}((U^{S_j})^\top X_t, Y_t). \tag{26}$$

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

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

(iii)

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

Proposition 3 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 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, $\dim(\mathcal{S}^{\mathrm{res}})$.

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 $\mathcal{S}^{\mathrm{inv}}$ and $\mathcal{S}^{\mathrm{res}}$.

Example 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*

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

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

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

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.

Appendix D. Auxiliary results

Lemma 3 *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 $\Pi_{\mathcal{S}_j} := U^{S_j} (U^{S_j})^\top$.

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 $A := \Pi_{\mathcal{S}_j} B \Pi_{\mathcal{S}_j}$ and $A^\dagger := \Pi_{\mathcal{S}_j} B^{-1} \Pi_{\mathcal{S}_j}$. Moreover, define $\widetilde{B} := U^\top B U$ and for all $k \in \{1, \ldots, q\}$, $\widetilde{B}_k := (U^{S_k})^\top B U^{S_k}$. 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 $\Pi_{\mathcal{S}}$. 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 $\widetilde{B}$,

$$\widetilde{B}^{-1} = \begin{bmatrix} \widetilde{B}_1^{-1} & & \\ & \ddots & \\ & & \widetilde{B}_q^{-1} \end{bmatrix} = \begin{bmatrix} (U^{S_1})^\top \\ \vdots \\ (U^{S_q})^\top \end{bmatrix} B^{-1} \begin{bmatrix} U^{S_1} & \ldots & U^{S_q} \end{bmatrix}. \tag{27}$$

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 3. ■

Lemma 4 *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 $\tilde{\Sigma}_t := U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\tilde{\Sigma}_{t,j} = (U^{S_j})^\top \Sigma_t U^{S_j}$, $j \in \{1, \ldots, q\}$ and of dimension $|S_j| = \dim(\mathcal{S}_j)$, where $S_j \subseteq \{1, \ldots, p\}$ indexes a subset of the columns of U. Moreover, $\Pi_{\mathcal{S}_j} = U^{S_j}(U^{S_j})^\top$.*

Proof Let $U = (u_1, \ldots, u_p) \in \mathbb{R}^{p \times p}$ be an orthonormal matrix with columns $u_1, \ldots, u_p$ such that for all $j \in \{1, \ldots, q\}$ there exists $S_j \subseteq \{1, \ldots, p\}$ such that $\mathcal{S}_j = \text{span}(\{u_k \mid k \in S_j\})$. 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 $i, j \in \{1, \ldots, q\}$ 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 $j \in \{1, \ldots, q\}$ 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

(6)) it holds that, for all $t \in \mathcal{N}$,

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

where for all $j \in \{1, \ldots, q\}$ we defined $\widetilde{\Sigma}_{t,j} := (U^{S_j})^\top \Sigma_t U^{S_j}$. ∎

Lemma 5 *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 $j \in \{1, \ldots, q\}$, that*

$$\text{Var} \left(\Pi_{\mathcal{S}_j} X_t \right)^\dagger = \Pi_{\mathcal{S}_j} \text{Var} \left(X_t \right)^{-1} \Pi_{\mathcal{S}_j}. \tag{28}$$

Proof Let $U \in \mathbb{R}^{p \times p}$ be an orthonormal matrix such that, for all $t \in [n]$, $\tilde{\Sigma}_t := U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\widetilde{\Sigma}_{t,1}, \ldots, \widetilde{\Sigma}_{t,q}$ of dimensions $\dim(\mathcal{S}_1), \ldots, \dim(\mathcal{S}_q)$. Such a matrix exists by Lemma 4 and each diagonal block is given by $\widetilde{\Sigma}_{t,j} := (U^{S_j})^\top \Sigma_t U^{S_j}$ 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 3. ∎

D.1 Proof of Lemma 1

Proof Let $U \in \mathbb{R}^{p \times p}$ be an orthonormal matrix such that, for all $t \in [n]$, $\tilde{\Sigma}_t := U^\top \Sigma_t U$ is block diagonal with q diagonal blocks $\widetilde{\Sigma}_{t,1}, \ldots, \widetilde{\Sigma}_{t,q}$ of dimensions $\dim(\mathcal{S}_1), \ldots, \dim(\mathcal{S}_q)$. Such a matrix exists by Lemma 4 and each diagonal block is given by $\widetilde{\Sigma}_{t,j} := (U^{S_j})^\top \Sigma_t U^{S_j}$ 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 (1) that $\gamma_{0,t} = \Sigma_t^{-1} \text{Cov}(X_t, Y_t)$ and hence we use U to get the following expansion

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

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 $\tilde{\Sigma}_t$,

$$\tilde{\Sigma}_t^{-1} = \begin{bmatrix} \tilde{\Sigma}_{t,1}^{-1} & & \\ & \ddots & \\ & & \tilde{\Sigma}_{t,q}^{-1} \end{bmatrix} = \begin{bmatrix} (U^{S_1})^\top \\ \vdots \\ (U^{S_q})^\top \end{bmatrix} \Sigma_t^{-1} \begin{bmatrix} U^{S_1} & \cdots & U^{S_q} \end{bmatrix}. \tag{29}$$

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 5. 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 and the third equality follows from Lemma 5. ∎

D.2 Proof of Lemma 2

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

$$
\begin{aligned}
\Delta \mathrm{Var}_t(\beta) &= 2 \, \mathrm{Cov}(Y_t, X_t)\beta - \beta^\top \, \mathrm{Var}(X_t)\beta \\
&= 2 \, \mathrm{Cov}(Y_t, (\textstyle\sum_{k=1}^q \Pi_{\mathcal{S}_k} X_t))\Pi_{\mathcal{S}_j}\beta - \beta^\top \Pi_{\mathcal{S}_j} \, \mathrm{Var}(\textstyle\sum_{k=1}^q \Pi_{\mathcal{S}_k} X_t)\Pi_{\mathcal{S}_j}\beta \\
&= 2 \, \mathrm{Cov}(Y_t, (\Pi_{\mathcal{S}_j} X_t))\Pi_{\mathcal{S}_j}\beta - \beta^\top \Pi_{\mathcal{S}_j} \, \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\Pi_{\mathcal{S}_j}\beta \\
&= 2 \, \mathrm{Cov}(Y_t, (\Pi_{\mathcal{S}_j} X_t))\beta - \beta^\top \, \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)\beta, \tag{30}
\end{aligned}
$$

where the first equality follows by (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

$$
\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \, \mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t) + (I_p - \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \, \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t))w \tag{31}
$$

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 4. We now observe that

$$
\begin{aligned}
\mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^\dagger \, \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t) &= \Pi_{\mathcal{S}_j} \Sigma_t^{-1} \Pi_{\mathcal{S}_j} \Pi_{\mathcal{S}_j} \Sigma_t \Pi_{\mathcal{S}_j} \tag{32} \\
&= U^{S_j}(U^{S_j})^\top \Sigma_t^{-1} U^{S_j}(U^{S_j})^\top \Sigma_t U^{S_j}(U^{S_j})^\top \\
&= U^{S_j} \tilde{\Sigma}_{t,j}^{-1} \tilde{\Sigma}_{t,j}(U^{S_j})^\top \\
&= \Pi_{\mathcal{S}_j}.
\end{aligned}
$$

where the first equality follows from Lemma 3, since U jointly block diagonalizes the matrices $(\Sigma_t)_{t \in \mathcal{N}}$ by Lemma 4. We can therefore rewrite (31) 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. ∎

Lemma 6 $\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ *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

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

∎

Lemma 7 $\mathcal{S}^{\mathrm{inv}}$ *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

$$\underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \, \Delta \mathrm{Var}_t(\beta) = \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t)^{\dagger} \mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t, Y_t)$$

$$= \sum_{\substack{j \in \{1,\ldots,q\}: \\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \mathrm{Var}(\Pi_{\mathcal{S}_j} X_t)^{\dagger} \mathrm{Cov}(\Pi_{\mathcal{S}_j} X_t, Y_t)$$

$$= \sum_{\substack{j \in \{1,\ldots,q\}: \\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \underset{\beta \in \mathcal{S}_j}{\arg\max} \, \Delta \mathrm{Var}_t(\beta)$$

$$= \sum_{\substack{j \in \{1,\ldots,q\}: \\ \mathcal{S}_j \text{ opt-invariant on } [n]}} \underset{\beta \in \mathcal{S}_j}{\arg\max} \, \Delta \mathrm{Var}_s(\beta)$$

$$= \underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \, \Delta \mathrm{Var}_s(\beta).$$

The first equality holds by Lemma 2, since $\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ is indeed an orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition, see Lemma 6. The second equality follows from the definition of $\mathcal{S}^{\mathrm{inv}}$ and can be proved by Lemma 1 and observing that the set of subspaces $\{\mathcal{S}_j \mid j \in \{1,\ldots,q\} : \mathcal{S}_j \text{ opt-invariant on } [n]\}$ is an orthogonal and $(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t)_{t \in [n]}$-decorrelating partition. The third equality holds by Lemma 2 and the fourth by definition of an opt-invariant subspace on $[n]$. ∎

Lemma 8 *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 2.1.1. 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 $q_{\max}^U = \sum_{j=1}^q \bar{m}_j$, 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 $U^{\mathcal{S}_j}$, with $j \in \{1,\ldots,q\}$ indexing the diagonal blocks and

$S_j \subseteq \{1, \ldots, p\}$ 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.,

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

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 $j \in \{1, \ldots, q\}$, $\bar{m}_j = 1$. Such condition implies that the diagonal blocks indexed by S_j, $j \in \{1, \ldots, q\}$, 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 $\{\mathcal{Q}_1, \ldots, \mathcal{Q}_p\}$ be the set of eigenspaces of A and for all $i \in \{1, \ldots, p\}$ let $m_i := \dim(\mathcal{Q}_i) = 1$. The proposition then implies that for all $i \in \{1, \ldots, p\}$ there exists $j \in \{1, \ldots, q\}$ 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 $j \in \{1, \ldots, q\}$, $\bar{m}_j = 1$. This concludes the proof for Lemma 8. ∎

Lemma 9 *Let $\{\mathcal{S}_j\}_{j=1}^q$ be an irreducible orthogonal and $(X_t)_{t \in [n]}$-decorrelating partition. Then, for all $j \in \{1, \ldots, q\}$ it holds that*

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

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

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

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 4. For all $j \in \{1, \ldots, q\}$ and for all $\beta \in \mathcal{S}_j$ it follows from (30) that

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

which has gradient

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

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

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

47

By Lemma 3, 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 3, 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

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

$$=\beta^* + \left(I_p - \Pi_{\mathcal{S}_j}\right)w.$$

The equality follows from the following computation

$$\left(\frac{1}{n}\sum_{t=1}^{n}\mathrm{Var}(\Pi_{\mathcal{S}_j}X_t)\right)^{\dagger}\left(\frac{1}{n}\sum_{t=1}^{n}\mathrm{Var}(\Pi_{\mathcal{S}_j}X_t)\right)$$

$$= \Pi_{\mathcal{S}_j}\left(\sum_{t=1}^{n}\mathrm{Var}(X_t)\right)^{-1}\Pi_{\mathcal{S}_j}\Pi_{\mathcal{S}_j}\left(\sum_{t=1}^{n}\mathrm{Var}(X_t)\right)\Pi_{\mathcal{S}_j}$$

$$= U^{S_j}(U^{S_j})^{\top}\left(\sum_{t=1}^{n}\Sigma_t\right)^{-1}U^{S_j}(U^{S_j})^{\top}\left(\sum_{t=1}^{n}\Sigma_t\right)U^{S_j}(U^{S_j})^{\top}$$

$$= U^{S_j}\left(\sum_{t=1}^{n}\tilde{\Sigma}_t\right)^{-1}\left(\sum_{t=1}^{n}\tilde{\Sigma}_t\right)(U^{S_j})^{\top}$$

$$= \Pi_{\mathcal{S}_j}.$$

The first equality follows again from Lemma 3. 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

$$\arg\max_{\beta \in \mathcal{S}_j} \overline{\Delta\mathrm{Var}}(\beta)$$

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

$$= \arg\max_{\beta \in \mathcal{S}_j} \Delta\mathrm{Var}_t(\beta)$$

$$= \mathrm{Var}(\Pi_{\mathcal{S}_j}X_t)^{\dagger}\mathrm{Cov}(\Pi_{\mathcal{S}_j}X_t, Y_t)$$

$$= \Pi_{\mathcal{S}_j}\gamma_{0,t},$$

where we used the definition of opt-invariance on $[n]$ for the second equality, Lemma 2 for the third equality and Lemma 1 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

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

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

Lemma 10 *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

$$
\begin{aligned}
\beta^{\mathrm{inv}} &= \underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \, \overline{\Delta \mathrm{Var}}(\beta) \\
&= \underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \, \frac{1}{n} \sum_{t=1}^n \Delta \mathrm{Var}_t(\beta) \\
&= \underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \, \frac{1}{n} \sum_{s=1}^n (2 \, \mathrm{Cov}(Y_t, \Pi_{\mathcal{S}^{\mathrm{inv}}} X_t) \beta - \beta^\top \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t) \beta) \\
&= \left(\frac{1}{n} \sum_{t=1}^n \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t) \right)^\dagger \frac{1}{n} \sum_{t=1}^n \mathrm{Cov}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_t, Y_t) \\
&= \Pi_{\mathcal{S}^{\mathrm{inv}}} \overline{\mathrm{Var}}(X)^{-1} \Pi_{\mathcal{S}^{\mathrm{inv}}} \overline{\mathrm{Cov}}(X,Y) \\
&= U^{\mathrm{inv}}((U^{\mathrm{inv}})^\top \overline{\mathrm{Var}}(X) U^{\mathrm{inv}})^{-1} (U^{\mathrm{inv}})^\top \overline{\mathrm{Cov}}(X,Y).
\end{aligned}
$$

The fourth equality follows from Lemma 9. The fifth equality follows by Lemma 3 (see the proof of Lemma 9). 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

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

$\blacksquare$

Appendix E. Proofs

E.1 Proof of Theorem 1

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 (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 1 and Lemma 2, 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 1. $\blacksquare$

E.2 Proof of Proposition 1

Proof (i) For all $j \in \{1, \ldots, q_{\mathrm{max}}^U\}$, 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

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

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 4, taking $\mathcal{N} = [n]$. Irreducibility of the joint block diagonalizer follows from the irreducibility of the orthogonal partition. ∎

E.3 Proof of Proposition 2

Proof (i) First, observe that, by Lemma 6, $\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ is an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition and that, by Lemma 7, $\mathcal{S}^{\mathrm{inv}}$ is opt-invariant on $[n]$. Then, by definition of β^{inv} and by Lemma 9 we get for all $t \in [n]$ that

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

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

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

where the third equality uses $\beta^{\mathrm{inv}} = \Pi_{\mathcal{S}^{\mathrm{inv}}} \gamma_{0,t}$ by Proposition 2 (i) and the last equality follows from the fact that $\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}$ are an orthogonal and $(X_t)_{t\in[n]}$-decorrelating partition by Lemma 6.

(iii) Let $\mathcal{B}^n$ denote the set of all time-invariant parameters over $[n]$. By assumption, we have that $\mathcal{B}^n \subseteq \mathcal{S}^{\mathrm{inv}}$. Moreover, by point (ii), we have that $\beta^{\mathrm{inv}} \in \mathcal{B}^n$. Therefore, by definition of β^{inv} we obtain that $\beta^{\mathrm{inv}} = \arg\max_{\beta \in \mathcal{B}^n} \overline{\Delta\mathrm{Var}}(\beta)$. This completes the proof of Proposition 2. ∎

E.4 Proof of Theorem 2

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

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

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

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

Therefore,

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

$$= \underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \Delta \mathrm{Var}_t(\beta).$$

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

$$\underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \Delta \mathrm{Var}_t(\beta) = \underset{\beta \in \mathcal{S}^{\mathrm{inv}}}{\arg\max} \overline{\Delta \mathrm{Var}(\beta)}.$$

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

E.5 Proof of Theorem 3

Proof We assume without loss of generality that the observed predictors X_t have zero mean in $t \in \mathcal{I}^{\mathrm{ad}} \cup t^*$. Alternatively, as mentioned in Section 4.1.2, a constant term could be added to X_t to account for the mean. We also observe that $\hat{\gamma}_{t^*}^{\mathrm{OLS}}$ and $\hat{\delta}_{t^*}^{\mathrm{res}}$ are both unbiased estimators for γ_{0,t^*} and $\delta_{t^*}^{\mathrm{res}}$, respectively (this can be checked using standard OLS analysis). We now start by computing the out-of-sample MSPE for $\hat{\gamma}_{t^*}^{\mathrm{ISD}}$.

$$\begin{aligned}
\mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{ISD}}) &= \mathbb{E}[(X_{t^*}^\top(\hat{\gamma}_{t^*}^{\mathrm{ISD}} - \gamma_{0,t^*}))^2] \\
&= \mathbb{E}[(X_{t^*}^\top(\Pi_{\mathcal{S}^{\mathrm{inv}}} + \Pi_{\mathcal{S}^{\mathrm{res}}})(\hat{\beta}^{\mathrm{inv}} + \hat{\delta}_{t^*}^{\mathrm{res}} - \beta^{\mathrm{inv}} - \delta_{t^*}^{\mathrm{res}}))^2] \\
&= \mathrm{trace}(\mathbb{E}[(\hat{\beta}^{\mathrm{inv}} - \beta^{\mathrm{inv}})(\hat{\beta}^{\mathrm{inv}} - \beta^{\mathrm{inv}})^\top] \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*})) \\
&\quad + \mathrm{trace}(\mathbb{E}[(\hat{\delta}_{t^*}^{\mathrm{res}} - \delta_{t^*}^{\mathrm{res}})(\hat{\delta}_{t^*}^{\mathrm{res}} - \delta_{t^*}^{\mathrm{res}})^\top] \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_{t^*})) \\
&= \mathrm{trace}(\mathbb{E}[\mathrm{Var}(\hat{\beta}^{\mathrm{inv}} \mid \mathbf{X})] \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*})) \\
&\quad + \mathrm{trace}(\mathbb{E}[\mathrm{Var}(\hat{\delta}_{t^*}^{\mathrm{res}} \mid \mathbf{X}^{\mathrm{ad}})] \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_{t^*})) \\
&= \mathrm{trace}(\mathbb{E}[\mathrm{Var}(\hat{\beta}^{\mathrm{inv}} \mid \mathbf{X})] \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*})) \\
&\quad + \frac{\sigma_{\mathrm{ad}}^2}{m} \mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{res}}} \mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top \mathbf{X}^{\mathrm{ad}})^{-1}] \Pi_{\mathcal{S}^{\mathrm{res}}} \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_{t^*})),
\end{aligned}$$

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

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

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

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

where $\preceq$ denotes the Loewner order, and it follows from $\mathrm{diag}_n(\sigma_{\epsilon,\min}^2) \preceq \mathrm{Var}(\epsilon_t) \preceq \mathrm{diag}_n(\sigma_{\epsilon,\max}^2)$, with $\mathrm{diag}_n(\cdot)$ 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 $\mathrm{MSPE}(\hat{\gamma}_{t^*}^{\mathrm{ISD}})$ is lower bounded by

$$\mathrm{trace}(\mathbb{E}[\mathrm{Var}(\hat{\beta}^{\mathrm{inv}} \mid \mathbf{X})] \, \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*}))$$
$$\geq \sigma_{\epsilon,\min}^2 \, \mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}} \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] \Pi_{\mathcal{S}^{\mathrm{inv}}} \, \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*}))$$
$$\geq \frac{\sigma_{\epsilon,\min}^2}{n} \, \mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}} \boldsymbol{\Sigma}^{-1} \Pi_{\mathcal{S}^{\mathrm{inv}}} \Sigma_{t^*} \Pi_{\mathcal{S}^{\mathrm{inv}}}),$$

where $\boldsymbol{\Sigma} := \mathbb{E}[\frac{1}{n} \mathbf{X}^\top \mathbf{X}]$. Using now that, by assumption, $\{\mathcal{S}^{\mathrm{inv}}, \mathcal{S}^{\mathrm{res}}\}$ 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 4 and let $U^{\mathrm{inv}} \in \mathbb{R}^{p \times \dim(\mathcal{S}^{\mathrm{inv}})}$ be the submatrix of U whose columns form an orthonormal basis for $\mathcal{S}^{\mathrm{inv}}$, such that $\Pi_{\mathcal{S}^{\mathrm{inv}}} = U^{\mathrm{inv}}(U^{\mathrm{inv}})^\top$. Moreover, let $\tilde{\boldsymbol{\Sigma}}^{\mathrm{inv}} := (U^{\mathrm{inv}})^\top \boldsymbol{\Sigma} U^{\mathrm{inv}}$ and $\tilde{\Sigma}_{t^*}^{\mathrm{inv}} := (U^{\mathrm{inv}})^\top \Sigma_{t^*} U^{\mathrm{inv}}$ denote, as in Lemma 4, the diagonal block corresponding to $\mathcal{S}^{\mathrm{inv}}$ of the block diagonal matrices $U^\top \boldsymbol{\Sigma} U$ and $U^\top \Sigma_{t^*} U$, respectively. By the properties of orthogonal matrices, $(\tilde{\boldsymbol{\Sigma}}^{\mathrm{inv}})^{-1} = (U^{\mathrm{inv}})^\top \boldsymbol{\Sigma}^{-1} U^{\mathrm{inv}}$. For all $i \in \{1, \ldots, \dim(\mathcal{S}^{\mathrm{inv}})\}$, let λ_i denote the i-th eigenvalue in decreasing order. Then, we can further express the above lower bound as

$$\frac{\sigma_{\epsilon,\min}^2}{n} \, \mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}} \boldsymbol{\Sigma}^{-1} \Pi_{\mathcal{S}^{\mathrm{inv}}} \Sigma_{t^*} \Pi_{\mathcal{S}^{\mathrm{inv}}})$$
$$= \frac{\sigma_{\epsilon,\min}^2}{n} \, \mathrm{trace}(U^{\mathrm{inv}}(U^{\mathrm{inv}})^\top \boldsymbol{\Sigma}^{-1} U^{\mathrm{inv}}(U^{\mathrm{inv}})^\top \Sigma_{t^*} U^{\mathrm{inv}}(U^{\mathrm{inv}})^\top)$$
$$= \frac{\sigma_{\epsilon,\min}^2}{n} \, \mathrm{trace}(U^{\mathrm{inv}}(\tilde{\boldsymbol{\Sigma}}^{\mathrm{inv}})^{-1} \tilde{\Sigma}_{t^*}^{\mathrm{inv}} (U^{\mathrm{inv}})^\top)$$
$$\geq \sigma_{\epsilon,\min}^2 \frac{\dim(\mathcal{S}^{\mathrm{inv}})}{n} \frac{\lambda_{\min}(\tilde{\Sigma}_{t^*}^{\mathrm{inv}})}{\lambda_{\max}(\tilde{\boldsymbol{\Sigma}}^{\mathrm{inv}})}$$
$$= \sigma_{\epsilon,\min}^2 \frac{\dim(\mathcal{S}^{\mathrm{inv}})}{n} c_{\mathrm{inv}},$$

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

$$\mathrm{trace}(\mathbb{E}[\mathrm{Var}(\hat{\beta}^{\mathrm{inv}} \mid \mathbf{X})] \, \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*}))$$
$$\leq \sigma_{\epsilon,\max}^2 \, \mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{inv}}} \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] \Pi_{\mathcal{S}^{\mathrm{inv}}} \, \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{inv}}} X_{t^*}))$$
$$\leq \sigma_{\epsilon,\max}^2 \frac{\dim(\mathcal{S}^{\mathrm{inv}})}{n} C_{\mathrm{inv}},$$

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

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

where $h_{\mathrm{res}}(m) := \mathrm{trace}(\Pi_{\mathcal{S}^{\mathrm{res}}} \mathbb{E}[(\tfrac{1}{m}(\mathbf{X}^{\mathrm{ad}})^\top \mathbf{X}^{\mathrm{ad}})^{-1}]\Pi_{\mathcal{S}^{\mathrm{res}}} \, \mathrm{Var}(\Pi_{\mathcal{S}^{\mathrm{res}}} X_{t^*}))$. Since we have assumed that for all $t \in \mathcal{I}^{\mathrm{ad}} \cup \{t^*\}$ the distribution of X_t does not change, by Jensen's inequality and by the fact that $\Pi_{\mathcal{S}^{\mathrm{res}}} \Sigma_{t^*}^{-1} \Pi_{\mathcal{S}^{\mathrm{res}}} \Sigma_{t^*} \Pi_{\mathcal{S}^{\mathrm{res}}} = \Pi_{\mathcal{S}^{\mathrm{res}}}$ (see Lemma 6 and (32)) we

have that $h_{\mathrm{res}}(m) \geq \dim(\mathcal{S}^{\mathrm{res}})$, and $\lim_{m \to \infty} h_{\mathrm{res}}(m) = \dim(\mathcal{S}^{\mathrm{res}})$. Moreover, we can find an upper bound for $h_{\mathrm{res}}(m)$ in the following way.

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

Summarizing, we obtain that

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where $C_{\mathrm{ad,inv}} := c^{-1}\lambda_{\max}(\Sigma_{t^*})$. In particular, this difference is always positive if n is sufficiently large. Finally, we can observe that the expected explained variance (24) for $\hat{\gamma}$, where $\hat{\gamma} = \hat{\gamma}_{t^*}^{\mathrm{ISD}}$ or $\hat{\gamma} = \hat{\gamma}_{t^*}^{\mathrm{OLS}}$, is

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

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

E.6 Proof of Proposition 3

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

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

The last relation is obtained by observing that

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

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 $i, j \in \{1, \ldots, q\}$ 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 $\bigoplus_{i\in\{1,\ldots,q\}\setminus\{j\}}\mathcal{S}_i$. Fix $j\in\{1,\ldots,q\}$. Then, using these relations we obtain that

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

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 $t\in[n]$,

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

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

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

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

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