Anchor PCA

Benedikt Seiter
ETH Zürich
bseiter@ethz.ch

Anya Fries
Seminar for Statistics
ETH Zürich

Julius von Kügelgen
Seminar for Statistics
ETH Zürich

Jonas Peters
Seminar for Statistics
ETH Zürich

Abstract

Principal component analysis (PCA) is one of the most widely used unsupervised dimension reduction techniques. We study PCA for data from multiple related domains. Since principal components generally differ across domains, one way to obtain a shared low-rank embedding is to perform PCA on the pooled data. However, this approach can focus on spurious directions that exhibit high variation in only a few domains. To find a robust embedding that still explains most variance in unseen but similar domains, we propose instead to focus on shared directions of variation. To this end, we introduce Anchor PCA which trades off overall explained variance with agreement between the shared and domain-specific low-rank embeddings. Anchor PCA amounts to PCA on a modified target matrix and thus can be solved efficiently. Moreover, we show that Anchor PCA recovers a maximal invariant subspace and admits a minimax reconstruction interpretation under bounded domain-specific covariance inflations. On simulated and real-world data with temporal distribution shift, we demonstrate, respectively, that Anchor PCA recovers the maximal invariant subspace and yields embeddings that explain more variance on unseen domains than the pooling baseline and a worst-case alternative. Taken together, these findings establish Anchor PCA as a promising approach to robust unsupervised dimension reduction from multi-domain data.

1 Introduction

Principal component analysis (PCA) [23] is one of the most widely used methods for unsupervised dimensionality reduction. Given a dataset, PCA finds a low-dimensional subspace that captures as much variance as possible. It is used extensively in exploratory data analysis, preprocessing, and representation learning across the natural sciences, engineering, and machine learning [e.g., 28, 31, 45, 52]. Unsupervised learning can serve different goals [22]. Here, we focus on two uses of PCA: providing interpretable directions (that may carry physical meaning) and explaining variance.

We assume that we have data collected across multiple related but heterogeneous settings—different geographic locations, time points or experimental conditions. We refer to these settings as *source domains*. If the principal subspaces differ substantially across the source domains, we may achieve neither of the two goals mentioned above. For example, PCA applied to the average covariance (`poolPCA`) could be dominated by a direction that concentrates variance in only one or two domains. This direction may explain very little variance elsewhere and thus may not yield physically relevant directions; similarly, it may not explain a lot of variance in a target domain.

In this paper, we propose to focus not only on explained variance but also on finding invariant directions. This builds on work in multi-domain dimension reduction and invariant learning. In dimension reduction, recent approaches optimize worst-case performance across domains [17, 51] or enforce invariance of distributions [30]. In supervised learning, invariance of predictive relations or representations across domains is widely studied [3, 49]. In contrast, we introduce a geometric notion of invariance based on agreement of domain-specific principal subspaces, and explicitly trade this against explained variance. Similar in spirit is anchor regression [37], which introduces

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

a soft invariance penalty in the supervised setting. Concretely, we propose $\texttt{AnchorPCA}_\lambda$ (with $\lambda \in \mathbb{R}_{\geq 0} \cup \{\infty\}$), a method that finds a k-dimensional subspace where the parameter λ trades off between explaining variance and finding invariant directions (e.g., for $\lambda = 0$, $\texttt{AnchorPCA}_\lambda$ reduces to $\texttt{poolPCA}$).[1] We further detail related work and how their objectives differ from ours in § A.

We prove that $\texttt{AnchorPCA}$ satisfies two attractive theoretical properties. First, if there exists a nontrivial m-dimensional invariant subspace $\mathcal{S}_\star$, that is, a subspace that lies in the top-k principal subspace of every domain, then $\texttt{AnchorPCA}_\infty$ is guaranteed to contain it, and the leading directions of $\texttt{AnchorPCA}_\lambda$ converge to $\mathcal{S}_\star$ as $\lambda \to \infty$ (Thm. 3.1). Second, $\texttt{AnchorPCA}_\lambda$ admits a minimax reconstruction interpretation: it is the rank-k projector that minimizes worst-case average reconstruction error over test domains that are subject to domain-specific covariance inflations (Thm. 4.1).

We now introduce the setting and build intuition through the motivating example in § 1.2. § 2 introduces Anchor PCA, § 3 and § 4 contain the main theoretical results, § 5 shows learning methods for finite data, and § 6 contains simulated and real-world experiments. § A, § B, § C, and § D contain further related work, further details on Anchor PCA, proofs, and experiment details, respectively.

1.1 Setting and notation

We consider a set of E source domains indexed by $e \in \mathcal{E} := \{1, \dots, E\}$. For each domain e, let $x_e \in \mathbb{R}^p$ be a zero-mean random (column) vector with covariance $\Sigma_e \in \mathbb{S}^p_+$ where $\mathbb{S}^p_+ := \{\Sigma \in \mathbb{R}^{p\times p} : \Sigma = \Sigma^\top,\ \Sigma \succeq 0\}$ is the set of symmetric positive semidefinite matrices and we further assume $\operatorname{Tr}(\Sigma_e) > 0$. Recall that PCA selects the eigenvectors of a covariance matrix Σ corresponding to its k largest eigenvalues, also called a top-k eigenbasis. We denote a top-k eigenbasis of Σ_e by $W_e \in \mathcal{O}_{p\times k}$, where $\mathcal{O}_{p\times k} := \{W \in \mathbb{R}^{p\times k} : W^\top W = I_k\}$ is the Stiefel manifold of $p \times k$ matrices with orthonormal columns. The rank-k principal projection matrices are given by $\Pi_k^{(e)} := W_e W_e^\top$, which can be viewed as a linear autoencoder with encoder $W_e^\top$ and decoder W_e.

Given the domain-specific covariance matrices $(\Sigma_e)_{e\in\mathcal{E}}$ (or i.i.d. samples from the corresponding distributions), we seek to solve the following problem, which we formalize in § 3 and § 4.

> Find a shared k-dimensional subspace that is robust with respect to changing domains.

Denote the average covariance by $\overline{\Sigma} := \frac{1}{E}\sum_{e=1}^E \Sigma_e$ and the average projector by $\overline{\Pi} := \frac{1}{E}\sum_{e=1}^E \Pi_k^{(e)}$. For general $W \in \mathcal{O}_{p\times k}$, we write $\Pi_W := WW^\top$ for the projector onto the space spanned by W. The Frobenius and operator norms of a matrix $X \in \mathbb{R}^{p\times k}$ are defined as $\|X\|_{\mathrm{F}} := \sqrt{\operatorname{Tr}(X^\top X)}$ and $\|X\|_{\mathrm{op}} := \sup_{\|v\|_2=1} \|Xv\|_2$, respectively. For a linear subspace $\mathcal{S} \leq \mathbb{R}^p$, let $\Pi_{\mathcal{S}}$ denote the orthogonal projector onto $\mathcal{S}$; if U has orthonormal columns spanning $\mathcal{S}$, then $\Pi_{\mathcal{S}} = UU^\top$.

1.2 Motivating example

We consider a simple example in $\mathbb{R}^4$ (so $p=4$) with canonical basis c_1, c_2, c_3, c_4, writing $a := c_1$ and $b := c_2$, where we seek a $k=3$-dimensional representation. We observe $E=3$ domains. Table 1 summarizes the eigenvectors and eigenvalues of the covariance matrices (we defer the details to § D.1).

Informally, we consider a direction invariant if it helps explain variance in every domain (the domain acts as an 'anchor' specifying what we would like to be invariant against [see also 37]). In this example, $\operatorname{span}(a, b)$ is an invariant space: in all domains, a and b are among the top-k eigenvectors.

Fig. 1 visualizes the subspaces that are chosen by the different methods. The figure shows the qualitative behavior that will recur throughout the paper. $\texttt{poolPCA}$ follows pooled variance and contains the space spanned by the nuisance directions: it keeps a but misses the invariant direction b. By contrast, both $\texttt{AnchorPCA}_{\lambda=25}$ and $\texttt{AnchorPCA}_\infty$ recover the invariant span exactly. The difference between these two lies in the third direction: $\texttt{AnchorPCA}_{\lambda=25}$ puts a stronger focus on the explained variance and is thus pulled more strongly toward w, which is the direction with largest eigenvalue (see Table 1).

In this paper, we focus on the subspace and not the basis vectors; that is, if in domain 3, the top-k eigenspace contained $\tilde{a}, \tilde{b}$, instead of a, b, such that $\operatorname{span}(a, b) = \operatorname{span}(\tilde{a}, \tilde{b})$, we would still regard $\operatorname{span}(a, b)$ an invariant subspace. The other directions u, v, and w can be considered nuisance directions (in the light of invariance).

[1] For a short visual explanation of Anchor PCA, see this video.

Table 1: **Eigendecomposition and reconstruction errors for the motivating example.** If $k = 3$, the invariant directions are $a = c_1$ and $b = c_2$, whereas u, v, and w are domain-specific directions in the c_3–c_4 plane, and $u^\perp$, $v^\perp$, $w^\perp$ are their orthogonal complements in that plane. The domain-specific columns list eigenvectors and eigenvalues, equivalently the explained variance of each direction. The invariant directions do not need fixed ranks: a is top-ranked in domain 2, while b outranks a in domain 3. The perturbed column evaluates the same reconstruction-error curves shown later in Fig. 2 at the dotted vertical line.

Rank	**Domain 1**		**Domain 2**		**Domain 3**	
	direction	value	direction	value	direction	value
1	u	220	a	120	w	320
2	a	140	v	90	b	120
3	b	90	b	70	a	80
4	$u^\perp$	25	$v^\perp$	10	$w^\perp$	10

Method	**Avg. reconstr. error (↓)**	
	Original covariances	Perturbed covariances
`poolPCA`	93.3	243.3
`AnchorPCA`$_{\lambda=25}$	98.5	163.4
`AnchorPCA`$_{\infty}$	113.8	172.5

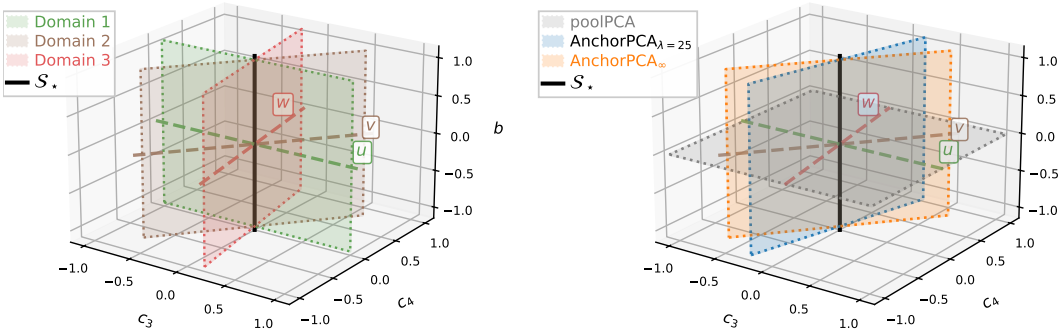

Figure 1: **Geometric view of the motivating example.** Here, we ignore the direction a, which is chosen by all methods, yielding a reduced representation in the coordinates (c_3, c_4, b). Left: the three local top-3 eigenspaces, shown as planes through the invariant line. Right: the recovered rank-3 subspaces in the same coordinates. Unlike `poolPCA`, both `AnchorPCA`$_{\lambda=25}$ and `AnchorPCA`$_\infty$ contain the true invariant direction b and differ only in the third direction ($\lambda = 25$ puts more emphasis on explained variance, so `AnchorPCA`$_{\lambda=25}$ is pulled toward the higher variance direction w).

2 Maximal invariant subspace and Anchor PCA

We now formally define the notion of a maximal invariant subspace and the Anchor PCA problem.

2.1 Maximal invariant subspace

Given source domains $\mathcal{E}$ with covariances $(\Sigma_e)_{e\in\mathcal{E}}$, a natural choice is to ignore the domain structure and solve PCA on the pooled covariance $\overline{\Sigma}$. We refer to this procedure as `poolPCA`. Formally, a matrix $W^\star$ solves rank-k `poolPCA` if $W^\star \in \arg\max_{W\in\mathcal{O}_{p\times k}} \mathrm{Tr}(W^\top \overline{\Sigma} W)$. Indeed, for any $W = [w_1, \ldots, w_k] \in \mathcal{O}_{p\times k}$, we have $\mathrm{Tr}(W^\top \overline{\Sigma} W) = \sum_{j=1}^k w_j^\top \overline{\Sigma} w_j = \mathrm{Tr}(\overline{\Sigma}\Pi_W)$, which is the total variance explained by projecting onto the candidate subspace $\mathrm{span}(W)$; hence, maximizing the trace recovers the k leading eigenvectors of $\overline{\Sigma}$. By definition, `poolPCA` focuses on the average explained variance, which may result in large differences in explained variance across source domains.

In this paper, we propose to focus on directions of variation that are shared among all domains. In particular, we can consider the directions appearing within the top-k principal directions in all domains.

Definition 2.1 (Maximal invariant subspace)**.** Let $k < p$ be fixed. The *maximal invariant subspace* and its dimension are defined as

$$\mathcal{S}_\star := \bigcap_{e=1}^{E} \mathrm{Im}\left(\Pi_k^{(e)}\right), \qquad m := \dim(\mathcal{S}_\star).$$

(If there are eigenvalue ties, $\mathcal{S}_\star$ and $\overline{\Pi}$ and the AnchorPCA solutions—defined below—depend on a fixed choice of a rank-k principal projector in each domain.) The definition of $\mathcal{S}_\star$ depends on k: in general, the dimension m of $\mathcal{S}_\star$ is at most k, but varies according to the problem at hand. In Example 1.2, $m = 2$ and $\mathcal{S}_\star = \operatorname{span}(a, b)$. Equivalently, we can define $\mathcal{S}_\star$ as the eigenspace of $\overline{\Pi}$ associated with eigenvalue 1; indeed, for every unit vector v, we have $v^\top \overline{\Pi} v = \frac{1}{E}\sum_{e=1}^{E} v^\top \Pi_k^{(e)} v \le 1$, and equality holds if and only if $\Pi_k^{(e)} v = v$ for every e, that is, if and only if $v \in \mathcal{S}_\star$. We will use this characterization when estimating $\mathcal{S}_\star$ from data (§ 5.2).

2.2 Anchor PCA

In practice, we may not expect to find any (nonzero) direction that is exactly invariant, in which case $\mathcal{S}_\star = \{0\}$ is trivial. Furthermore, we may be interested in finding a subspace of fixed dimension k, rather than one whose dimension can vary depending on the concrete set of covariance matrices. We therefore propose selecting a rank-k subspace that minimizes disagreement with the domain-specific top-k eigenspaces. Such a space may be represented using different bases; we propose to choose one that maximizes pooled explained variance. We refer to this approach as $\texttt{AnchorPCA}_\infty$.

Definition 2.2 ($\texttt{AnchorPCA}_\infty$)**.** A matrix $W_k^{\perp,\infty} \in \mathcal{O}_{p\times k}$ solves *rank-k* $\texttt{AnchorPCA}_\infty$ if

$$W_k^{\perp,\infty} \in \operatorname*{arg\,max}_{W\in\mathcal{O}_{p\times k}} \operatorname{Tr}\left(W^\top \overline{\Sigma} W\right) \qquad \text{subject to:} \qquad W \in \operatorname*{arg\,min}_{\widetilde{W}\in\mathcal{O}_{p\times k}} \sum_{e=1}^{E} \left\| \Pi_{\widetilde{W}} - \Pi_k^{(e)} \right\|_{\mathrm{F}}^2. \tag{1}$$

We prove in § 4 that $\texttt{AnchorPCA}_\infty$ satisfies an asymptotic minimax-optimality guarantee for reconstruction error under additional variance inflations of arbitrary strengths in certain directions. If we only expect mild inflations, however, $\texttt{AnchorPCA}_\infty$ is overly conservative. We therefore also consider the following version in which the hard invariance constraint is relaxed to a soft penalty.

Definition 2.3 ($\texttt{AnchorPCA}_\lambda$)**.** For $\lambda \ge 0$, a matrix $W_k^{\perp,\lambda} \in \mathcal{O}_{p\times k}$ solves *rank-k* $\texttt{AnchorPCA}_\lambda$ if

$$W_k^{\perp,\lambda} \in \operatorname*{arg\,max}_{W\in\mathcal{O}_{p\times k}} \left\{ \operatorname{Tr}\left(W^\top \overline{\Sigma} W\right) - \lambda \sum_{e=1}^{E} \left\| \Pi_W - \Pi_k^{(e)} \right\|_{\mathrm{F}}^2 \right\}. \tag{2}$$

For $\lambda \in [0, \infty]$, the corresponding *rank-k* $\texttt{AnchorPCA}_\lambda$ *subspace* and *projection matrix* are given by $\mathcal{S}_k^{\perp,\lambda} := \operatorname{Im}(W_k^{\perp,\lambda})$ and $\Pi_k^{\perp,\lambda} := W_k^{\perp,\lambda}(W_k^{\perp,\lambda})^\top$, respectively.

While $\texttt{AnchorPCA}_\infty$ only considers maximally invariant solutions and explained variance serves as a secondary objective, $\texttt{AnchorPCA}_\lambda$ balances the two, with the trade-off controlled by the invariance level λ. We prove in § B.2.1 that, as λ grows, the two solutions converge to each other: under suitable assumptions, we have

$$\left\| \Pi_k^{\perp,\lambda} - \Pi_k^{\perp,\infty} \right\|_{\mathrm{op}} \to 0 \qquad \text{as } \lambda \to \infty.$$

Thm. C.8 (ii) studies the convergence at the level of directions rather than subspaces and projections.

Both Anchor PCA methods can be extended to a soft agreement score: even when $\mathcal{S}_\star$ is too small or trivial, directions with larger $\overline{\Pi}$-eigenvalues are still preferred, rather than requiring exact equality of eigenvalues to one.

2.3 Projection-based formulation

The optimization problems (1) and (2) allow for an equivalent, projection-based view: expanding the penalty in (1) gives $\sum_{e=1}^{E} \|\Pi_W - \Pi_k^{(e)}\|_{\mathrm{F}}^2 = \sum_{e=1}^{E} 2k - 2\operatorname{Tr}(\Pi_W \Pi_k^{(e)}) = 2Ek - 2E\operatorname{Tr}(W^\top \overline{\Pi} W)$. Hence, the constraint is equivalent to maximizing $\operatorname{Tr}(W^\top \overline{\Pi} W)$ over $\mathcal{O}_{p\times k}$. Similarly, for (2), solutions are obtained by solving a standard rank-k PCA on a modified target matrix M_λ

$$\max_{W\in\mathcal{O}_{p\times k}} \operatorname{Tr}(W^\top M_\lambda W), \qquad \text{where} \qquad M_\lambda := \overline{\Sigma} + 2E\lambda\,\overline{\Pi}. \tag{3}$$

This view has two advantages: (i) It suggests computationally feasible ways to solve the optimization problems of $\texttt{AnchorPCA}_\infty$ and $\texttt{AnchorPCA}_\lambda$ (see § 5). (ii) It allows for a geometric interpretation: $\operatorname{Tr}(\Pi_W \Pi_k^{(e)})$ is the sum of squared cosines of the principal angles between $\operatorname{span}(W)$ and the principal subspace of domain e. Thus, $\texttt{AnchorPCA}_\infty$ first maximizes average overlap with the domain-specific principal subspaces and then maximizes pooled explained variance among those solutions.

2.4 Choosing an ordered basis

Both objectives (1) and (2) depend on W only through $\Pi_W = WW^\top$, so any orthonormal basis of a solution space defined by W forms another solution; statements about individual columns require choosing an ordered basis. We propose to use the following convention for choosing an ordered basis (see § B.1 for details). For $\texttt{AnchorPCA}_\infty$, we first order directions by decreasing agreement, that is, by decreasing $\overline{\Pi}$-eigenvalue; this creates blocks of equal eigenvalues (which we refer to in the algorithm and proof sections); we then choose a basis that diagonalizes $\overline{\Sigma}$ restricted to the $\overline{\Pi}$-eigenspace corresponding to that block and order the vectors by non-increasing $\overline{\Sigma}$-eigenvalue, equivalently, explained variance. For $\texttt{AnchorPCA}_\lambda$, we use an ordered top-$k$ eigenbasis of M_λ, see Eq. (3), giving the usual sequential PCA interpretation for the modified covariance M_λ. Both of these conventions are implemented in our code, see § 5.

We will see in § 3 and § 4 that both versions of `AnchorPCA` can find invariant directions and come with improved recovery and robustness properties compared to `poolPCA`.

3 Recovery guarantees

We now study when `AnchorPCA` recovers invariant structure. As motivated in Example 1.2, the subspace target is $\mathcal{S}_\star = \cap_{e=1}^{E} \mathrm{Im}(\Pi_k^{(e)})$, the part of the local top-k principal subspaces shared by all domains. Since $\texttt{AnchorPCA}_\lambda$ and $\texttt{AnchorPCA}_\infty$ return rank-k subspaces, there are two natural recovery questions: whether these subspaces contain $\mathcal{S}_\star$, and whether their leading $m = \dim(\mathcal{S}_\star)$ directions span $\mathcal{S}_\star$. We first prove that the $\texttt{AnchorPCA}_\infty$ subspace $\mathcal{S}_k^{\perp,\infty}$ contains $\mathcal{S}_\star$, and that $\texttt{AnchorPCA}_\lambda$ identifies $\mathcal{S}_\star$ asymptotically as $\lambda \to \infty$.

Theorem 3.1 (Invariant subspace identification and asymptotic convergence)**.** *Let $m = \dim(\mathcal{S}_\star) \le k$. Let $\Pi_k^{\perp,\infty}$ be any rank-k $\texttt{AnchorPCA}_\infty$ projection matrix. Then*

$$\Pi_k^{\perp,\infty}\Pi_{\mathcal{S}_\star} = \Pi_{\mathcal{S}_\star}, \qquad \textit{equivalently} \qquad \mathcal{S}_\star \subseteq \mathrm{Im}(\Pi_k^{\perp,\infty}).$$

Moreover, for $\lambda > 0$, choose the $\texttt{AnchorPCA}_\lambda$ solution $W_k^{\perp,\lambda} = (w_1^{\perp,\lambda}, \dots, w_k^{\perp,\lambda})$ to be an ordered orthonormal eigenbasis of M_λ associated with its k largest eigenvalues (see § 2.4). Let $\Pi_m^{\perp,\lambda}$ denote the projection matrix onto $\mathrm{span}(w_1^{\perp,\lambda}, \dots, w_m^{\perp,\lambda})$. Then, there exists λ_0 such that for all $\lambda > \lambda_0$, the span of these first m eigenvectors is unique and

$$\mathrm{span}(w_1^{\perp,\lambda}, \dots, w_m^{\perp,\lambda}) \to \mathcal{S}_\star, \quad \textit{i.e.,} \quad \left\|\Pi_m^{\perp,\lambda} - \Pi_{\mathcal{S}_\star}\right\|_{\mathrm{op}} \to 0 \quad \textit{as } \lambda \to \infty.$$

Under the ordered basis convention in § B.1.2, also the first m $\texttt{AnchorPCA}_\infty$ directions span $\mathcal{S}_\star$ (see Thm. C.8). Moreover, § B.2.2 shows, under an additional block-diagonality assumption on $\overline{\Sigma}$, that $\texttt{AnchorPCA}_\lambda$ recovers $\mathcal{S}_\star$ already for finite values of λ (this assumption holds in the example of § 1.2).

4 Optimal reconstruction under perturbations

We now study robustness of the solutions under structured covariance inflations. It is straightforward to see that $\mathcal{S}_\star$ satisfies the following robustness guarantee under eigengap conditions on the covariances Σ_e (see Prop. B.5 for details): $\mathcal{S}_\star$ is invariant to arbitrarily large admissible domain-specific top-k variance inflations; more precisely, it remains the maximally invariant subspace under perturbations of the form $\Sigma_e' = \Sigma_e + \Delta_e$ with $0 \preceq \Delta_e$ and $\mathrm{Im}(\Delta_e) \subseteq \mathrm{Im}(\Pi_k^{(e)})$ for all $e \in \mathcal{E}$.

We now prove that both versions of `AnchorPCA` satisfy a similar guarantee under bounded versions of the same perturbations. More precisely, for a covariance matrix Σ, the expected reconstruction error of a rank-k projector Π_W is given by $\mathbb{E}[\|x - \Pi_W x\|_2^2] = \mathrm{Tr}(\Sigma(I_p - \Pi_W))$. We further define, for any $\lambda \ge 0$, the following set of perturbed covariances

$$\mathcal{C}_\lambda := \left\{ (\Sigma_1', \dots, \Sigma_E') \;\middle|\; \Sigma_e' = \Sigma_e + \Delta_e \in \mathbb{S}_+^p, \;\; 0 \preceq \Delta_e \preceq 2E\lambda\, \Pi_k^{(e)} \text{ for all } e \in \mathcal{E} \right\} \tag{4}$$

and obtain the following result.

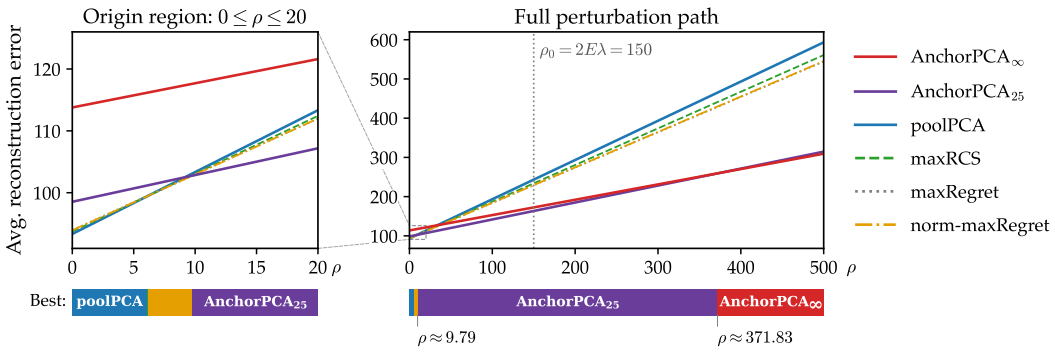

Figure 2: Average reconstruction error along the perturbation path $\Sigma_e^{(\rho)} = \Sigma_e + \rho\,\Pi_k^{(e)}$ for the example in § 1.2. As suggested by Thm. 4.1, `poolPCA` is best only for small perturbations ($0 \le \rho \lesssim 6.08$). Among the wcPCA [17] baselines, `maxRCS` is best only on the tiny interval $6.08 \lesssim \rho \lesssim 6.15$, after which `norm-maxRegret` is best until $\rho \lesssim 9.79$. `AnchorPCA`$_{25}$ dominates on the broad middle regime $9.79 \lesssim \rho \lesssim 371.83$, and `AnchorPCA`$_\infty$ performs best for strong perturbations $\rho \gtrsim 371.83$.

Theorem 4.1 (Robust reconstruction). *For $W \in \mathcal{O}_{p\times k}$, we define the average reconstruction error over $\{\Sigma_e\}_{e\in\mathcal{E}}$ as $\mathcal{R}(W;\Sigma_1,\dots,\Sigma_E) := \frac{1}{E}\sum_{e=1}^{E} \mathrm{Tr}(\Sigma_e(I_p - \Pi_W))$. Then, (i),*

$$W_k^{\perp,\lambda} \in \operatorname*{arg\,min}_{W\in\mathcal{O}_{p\times k}} \max_{(\Sigma_1',\dots,\Sigma_E')\in\mathcal{C}_\lambda} \mathcal{R}(W;\Sigma_1',\dots,\Sigma_E') \iff W_k^{\perp,\lambda} \text{ solves rank-}k \texttt{ AnchorPCA}_\lambda.$$

and, (ii), with $\mathcal{B} := \arg\min_{\widetilde{W}\in\mathcal{O}_{p\times k}} \sum_{e=1}^{E} \|\Pi_{\widetilde{W}} - \Pi_k^{(e)}\|_{\mathrm{F}}^2$

$$W_k^{\perp,\infty} \in \operatorname*{arg\,min}_{W\in\mathcal{B}} \mathcal{R}(W;\Sigma_1,\dots,\Sigma_E) \iff W_k^{\perp,\infty} \text{ solves rank-}k \texttt{ AnchorPCA}_\infty.$$

Thus, if we consider target domains that are subject to covariance inflations described by the set $\mathcal{C}_\lambda$, then `AnchorPCA`$_\lambda$ minimizes the worst-case reconstruction error over such domains. We now aim to provide intuition for part (ii). The proof of Thm. 4.1 shows that

$$\max_{(\Sigma_1',\dots,\Sigma_E')\in\mathcal{C}_\lambda} \mathcal{R}(W;\Sigma_1',\dots,\Sigma_E') = \frac{1}{E}\sum_{e=1}^{E} \mathrm{Tr}\big(\Sigma_e(I_p - \Pi_W)\big) + \lambda \sum_{e=1}^{E} \|\Pi_W - \Pi_k^{(e)}\|_{\mathrm{F}}^2. \quad (5)$$

Equation (5) shows that the worst-case reconstruction error is affine in λ, with slope equal to the projector-disagreement penalty. `AnchorPCA`$_\infty$ first minimizes this slope over rank-k projectors and then, among projectors with minimal slope, minimizes the nominal reconstruction error.[2]

In finite samples, covariances from new environments will not exactly lie in the set $\mathcal{C}_\lambda$ (4). For example, covariance estimation errors generally do not only lie in the local top-k subspaces, are not necessarily positive semidefinite, and they may be full rank. We provide a relaxation of Thm. 4.1 in Corollary C.9, in which we allow the deviations Δ_e to be full-rank with positive and negative eigenvalues, provided the resulting covariances $\Sigma_e + \Delta_e$ remain positive semidefinite and satisfy $\Sigma_e + \Delta_e \preceq \Sigma_e + 2E\lambda\Pi_k^{(e)} + \varepsilon_e I_p$. Here, $\varepsilon_e \ge 0$ specifies how much extra variance is allowed in any direction in domain e. `AnchorPCA`$_\lambda$ still minimizes the worst-case reconstruction error over this broader class, though allowing more extra variance (larger ε_e) increases the worst-case reconstruction error.

Fig. 2 visualizes parts of Thm. 4.1. It shows that `poolPCA` is optimal only for small perturbations, whereas `AnchorPCA`$_\lambda$ and `AnchorPCA`$_\infty$ guard against medium-sized and large perturbations. Further details are provided in § D.2. The experiment in § 6.2 shows that robustness guarantees like the ones shown in Thm. 4.1 may be of practical relevance.

[2] A single minimax problem over the unbounded union $\bigcup_{\lambda>0} \mathcal{C}_\lambda$ is typically degenerate, as can be seen from Equation (5). So the constrained formulation is the natural non-asymptotic statement.

Algorithm 1 Block-stabilized empirical $\texttt{AnchorPCA}_\infty$

Input: $\widehat{\Sigma}$ (average of domain-specific covariances), $\widehat{\Pi}$ (average rank-k projectors), k, tol_n
Output: $\widehat{W}_k^{\perp,\infty}$, $\widehat{\Pi}_k^{\perp,\infty}$, $\widehat{\mathcal{S}_\star}$, $\widehat{m}_\infty$
▷ *Step 1: Measure cross-domain agreement.*
Eigendecompose $\widehat{\Pi}$ as $\widehat{\Pi} = \sum_{j=1}^{p} \widehat{\rho}_j \widehat{v}_j \widehat{v}_j^\top$ with $\widehat{\rho}_1 \geq \cdots \geq \widehat{\rho}_p$.
▷ *Step 2: Group eigenvalues that lie within* tol_n *of each other.*
Set $a \leftarrow 1$ and $\ell \leftarrow 1$.
while $a \leq p$ **do**
  Set $b \leftarrow a$.
  while $b < p$ and $\widehat{\rho}_a - \widehat{\rho}_{b+1} \leq \mathrm{tol}_n$ **do**
    Set $b \leftarrow b + 1$.
  end while
  Set $I_\ell \leftarrow \{a, \ldots, b\}$ and $\widehat{\mathrm{Eig}}_\ell \leftarrow \mathrm{span}\{\widehat{v}_j : j \in I_\ell\}$.
  Set $a \leftarrow b + 1$ and $\ell \leftarrow \ell + 1$.
end while
▷ *Step 3: Fill k directions block by block, ranking by variance within a block.*
Let r be the first block index such that $\sum_{\ell \leq r} \dim(\widehat{\mathrm{Eig}}_\ell) \geq k$, and set $q \leftarrow k - \sum_{\ell < r} \dim(\widehat{\mathrm{Eig}}_\ell)$.
For all $\ell < r$, take an ordered eigenbasis of $\widehat{\Sigma}$ within $\widehat{\mathrm{Eig}}_\ell$.
Take the top-q eigenvectors of $\widehat{\Sigma}$ within $\widehat{\mathrm{Eig}}_r$.
Concatenate these vectors to obtain $\widehat{W}_k^{\perp,\infty}$.
Set $\widehat{\Pi}_k^{\perp,\infty} \leftarrow \widehat{W}_k^{\perp,\infty} (\widehat{W}_k^{\perp,\infty})^\top$.
▷ *Step 4: Estimate the invariant subspace.*
Set $\widehat{\mathcal{S}_\star} \leftarrow \widehat{\mathrm{Eig}}_1$ and $\widehat{m}_\infty \leftarrow \dim(\widehat{\mathrm{Eig}}_1)$.

5 Learning methods for Anchor PCA

We now describe the finite-sample methods for both `AnchorPCA` versions. Suppose domain $e \in \mathcal{E}$ provides observations $x_{e,1}, \ldots, x_{e,n_e} \in \mathbb{R}^p$. Define $\bar{x}_e := \frac{1}{n_e} \sum_{i=1}^{n_e} x_{e,i}$ and the empirical covariance $\widehat{\Sigma}_e := \frac{1}{n_e - 1} \sum_{i=1}^{n_e} (x_{e,i} - \bar{x}_e)(x_{e,i} - \bar{x}_e)^\top$. Let $\widehat{W}_e \in \mathbb{R}^{p \times k}$ be a top-k eigenbasis of $\widehat{\Sigma}_e$, and write the empirical versions of the domain-specific projection matrix as $\widehat{\Pi}_k^{(e)} := \widehat{W}_e \widehat{W}_e^\top$, average covariance as $\widehat{\Sigma} := \sum_{e=1}^{E} \widehat{\Sigma}_e / E$, and average projection matrix as $\widehat{\Pi} := \sum_{e=1}^{E} \widehat{\Pi}_k^{(e)} / E$, where we omit the bar compared to their population-level counterparts $\overline{\Sigma}, \overline{\Pi}$ for readability. We have weighted each domain here equally regardless of sample size. Alternatively, one may weight domains in proportion to their sample sizes, $w_e = n_e / \sum_j n_j$, which corresponds to weighting individual observations equally across domains.

5.1 Finite-sample implementation

We define empirical $\texttt{AnchorPCA}_\lambda$ as follows: we compute rank-k PCA of $\widehat{M}_\lambda := \widehat{\Sigma} + 2E\lambda\,\widehat{\Pi}$, see Equation (3), and return an ordered top-k eigenbasis $\widehat{W}_k^{\perp,\lambda}$ of $\widehat{M}_\lambda$. $\texttt{AnchorPCA}_\infty$ ranks directions first by cross-domain agreement, measured by the eigenvalues of $\overline{\Pi}$ (with population eigenvalue 1 corresponding to $\mathcal{S}_\star$), and then, among directions with equal agreement, by explained variance. A plug-in implementation replacing population quantities by their empirical counterparts would not yield a consistent method: for example, due to estimation error, all eigenvalues of $\widehat{\Pi}$ would generally be distinct, creating blocks of equal eigenvalues of size one (see § 2.4). Instead, we group consecutive $\widehat{\Pi}$-eigenspaces whenever eigenvalues remain within tol_n of the first eigenvalue of that group, where $n := \min_{e \in \mathcal{E}} n_e$; for each of these groups of eigenspaces, we compute the direct sum and denote the resulting spaces by $\widehat{\mathrm{Eig}}_1, \widehat{\mathrm{Eig}}_2, \ldots$, and rank directions within each group by their variance under $\widehat{\Sigma}$; see Algorithm 1 for details. If $\mathcal{S}_\star \neq \{0\}$ and tol_n separates the empirical eigenvalue-1 block from the rest of the spectrum (in our experiments we use $\mathrm{tol}_n = \min\{0.05, 0.5n^{-0.4}\}$ as a default choice), $\widehat{\mathrm{Eig}}_1$ consistently estimates $\mathcal{S}_\star$ (see § 5.3). Thus, the algorithm also returns estimates of the invariant subspace and its dimension.

Choosing k. The dimension k determines the output dimension and the local top-k spaces whose intersection defines $\mathcal{S}_\star$. Similar to PCA, we suggest to use domain knowledge when available, inspect pooled and source-domain scree plots [8], or choose the smallest k meeting a prespecified source explained variance requirement. For a downstream task, source-domain validation can select the smallest k within one standard error of the best score.

Selecting the penalty. The penalty λ controls the trade-off between explained variance and agreement across domains. For fixed k, we propose leave-one-source-domain-out cross-validation of $2E\lambda$, rather than λ, since $2E\lambda$ is the coefficient that weights $\overline{\Pi}$ in $M_\lambda = \overline{\Sigma} + 2E\lambda\overline{\Pi}$. With at least three source domains, we select the candidate with the highest mean held-out proportion of variance explained, then refit on all sources. We call this $\texttt{AnchorPCA}_{\widehat{\lambda}}$. If source labels and a downstream task are available, the held-out score can instead be a task metric. The candidate grid and implementation details are given in § D.4.

5.2 $\texttt{FindS}_\star$: Identifying the invariant subspace $\mathcal{S}_\star$

In some settings, the object of interest is not a full k-dimensional subspace (as estimated by AnchorPCA), but only the maximally invariant subspace $\mathcal{S}_\star$. We now describe how to estimate an unordered basis of $\mathcal{S}_\star$ directly.

An eigenvalue problem. Recall that $\mathcal{S}_\star$ is the eigenspace of $\overline{\Pi}$ associated with eigenvalue 1 (§ 2). A natural approach would therefore be to identify eigenvalues equal to 1. While this characterization is exact at the population level, it cannot be applied directly to finite samples. Although bootstrapping projection matrices is feasible [e.g., 5], the null hypothesis $\lambda_i(\overline{\Pi}) = 1$ lies on the boundary of the parameter space, and empirical eigenvalues will have value strictly less than 1, making the use of bootstrap-based confidence intervals difficult. Moreover, bootstrapping the null or using a parametric bootstrap is challenging, as the null is composite and depends on an unknown invariant subspace.

Testing the dimension of $\mathcal{S}_\star$. We instead directly test the dimension of $\mathcal{S}_\star$. If $\dim(\mathcal{S}_\star) = m$ were known, $\mathcal{S}_\star$ is estimated by the eigenspace of $\widehat{\Pi}$ associated with its m largest eigenvalues. Let $m_{\min} := \max(0, p - E(p-k))$. This motivates the sequential testing procedure

$$H_{0,s} : \dim(\mathcal{S}_\star) = s \quad \text{vs} \quad H_{A,s} : \dim(\mathcal{S}_\star) < s, \tag{6}$$

for $s \in \{k, k-1, \ldots, m_{\min}+1\}$, stopping at the first s for which the null is not rejected; if all nulls are rejected, estimate $\dim(\mathcal{S}_\star)$ as $m_{\min}$. To construct such tests, we adapt ideas from partial common principal component analysis (PCPCA) subspaces [42]. In the PCPCA setting, it is assumed that for all $e \in \mathcal{E}$, we observe a sample of i.i.d. Gaussian observations with population positive-definite covariance Σ_e. Additionally, Σ_e is assumed to satisfy a strict eigengap between its kth and $(k+1)^{\text{st}}$ eigenvalues. Under these assumptions, Wald-type tests are available for the dimension of the union of the top-k eigenspaces of $(\Sigma_e)_{e\in\mathcal{E}}$. To translate this to our setting, we consider the orthogonal complement projector matrices $I_p - \Pi_k^{(e)}$ and let r be the dimension of the union of their eigenspaces. Since a vector belongs to $\mathcal{S}_\star$ if and only if it lies in the null space of $I_p - \Pi_k^{(e)}$ for every $e \in \mathcal{E}$, $r = p - \dim(\mathcal{S}_\star)$. We apply the PCPCA subspace test to the bottom-$(p-k)$ eigenspaces of the covariance matrices, equivalently to $\mathrm{Im}(I_p - \Pi_k^{(e)})$, and thus obtain a sequential testing procedure for (6). Combining this sequential test with the eigenspace estimator above, we define $\texttt{FindS}_\star$ as the procedure that first estimates $\dim(\mathcal{S}_\star)$ and then returns the eigenspace of $\widehat{\Pi}$ associated with its estimated number of largest eigenvalues. The test statistic and further details are given in § B.3.

5.3 Consistency of Anchor PCA

Both AnchorPCA methods are consistent under mild conditions (the full statement and proof are given in § C.4). Assume that the minimum domain sample size $n := \min_{e\in\mathcal{E}} n_e \to \infty$, that observations are centered with finite second moments, and that each domain covariance Σ_e has a strict eigengap at rank k. Then $\widehat{\Sigma} \to \overline{\Sigma}$ and $\widehat{\Pi} \to \overline{\Pi}$ almost surely, and $\texttt{AnchorPCA}_\lambda$ is consistent whenever M_λ itself has a strict eigengap at rank k. For $\texttt{AnchorPCA}_\infty$, consistency requires a little more care: the block tolerance used in the algorithm must shrink to zero, but slowly enough relative to the estimation error of $\widehat{\Pi}$. Any tolerance of the form $\mathrm{tol}_n = cn^{-\alpha}$ with $0 < \alpha < 1/2$ works under finite fourth moments. Under these conditions, the full rank-k projector is consistent when the population $\texttt{AnchorPCA}_\infty$ solution projector is unique; moreover, if $\mathcal{S}_\star \neq \{0\}$, the first block returned by $\texttt{AnchorPCA}_\infty$ is a

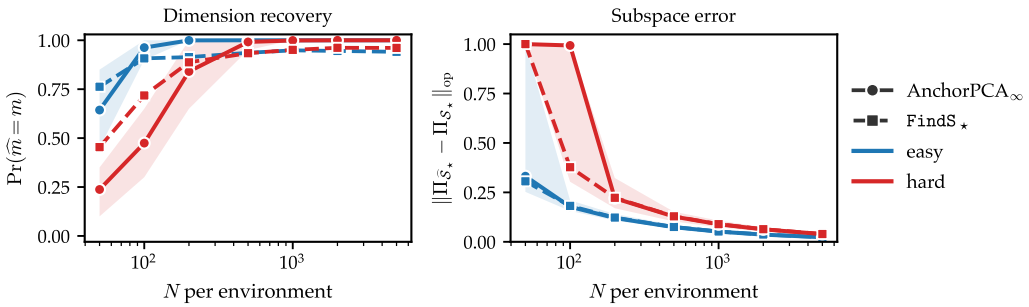

Figure 3: **Recovery of $\mathcal{S}_\star$ by AnchorPCA$_\infty$ and FindS$_\star$ (§ 5.2).** Gaussian data with $E = 5$, $p = 10$, $k = 5$, and $m = \dim(\mathcal{S}_\star) = 2$. We show mean correct-dimension probability (left) and median operator-norm projector error (right) against the true $\mathcal{S}_\star$. As predicted by § 5.3, the first grouped eigenspace of AnchorPCA$_\infty$ (solid) recovers $\mathcal{S}_\star$ with growing sample size and FindS$_\star$ (dashed) has dimension-recovery error close to the nominal test size of 0.05. For each sample size N, results are aggregated over 100 independent distribution draws, for each of which we draw 20 independent datasets. Shaded bands show 10–90% quantiles across distribution draws (shown only for AnchorPCA$_\infty$).

consistent estimator of $\mathcal{S}_\star$. For FindS$_\star$, Schott's asymptotic theory [42] implies that the probability of correctly estimating $\dim(\mathcal{S}_\star)$ converges to $1 - \alpha$, where α is the nominal level of the tests.

6 Experiments

Code to reproduce all experiments and figures is publicly available at https://github.com/benesei/anchor-pca.

6.1 Recovering the invariant subspace $\mathcal{S}_\star$

In this experiment, all domains have equal sample size, which we call N (which thus also equals $n = \min_{e \in \mathcal{E}} n_e$). § 5.3 proves that the first grouped eigenspace, $\widehat{\mathrm{Eig}}_1$, of the block-stabilized AnchorPCA$_\infty$ construction consistently estimates $\mathcal{S}_\star$ when the tolerance shrinks while dominating the empirical perturbation. We test this statement in a random-subspace model where $\mathcal{S}_\star$ is known, using the default block tolerance ($\mathrm{tol}_N = \min\{0.05, 0.5N^{-0.4}\}$). Below we choose the value $m = \dim(\mathcal{S}_\star)$ and let $\widehat{m}_\infty$ denote the dimension of the estimate $\widehat{\mathcal{S}_\star}$. We measure subspace error by $\|\Pi_{\widehat{\mathcal{S}}_\star} - \Pi_{\mathcal{S}_\star}\|_{\mathrm{op}}$.

The distribution draws are as follows. We draw a random m-dimensional subspace $\mathcal{S}_\star \subset \mathbb{R}^p$ and, for each domain, a different rank-k principal subspace containing $\mathcal{S}_\star$. The covariance in that domain is then constructed so that this subspace is exactly its top-k eigenspace. We consider two DGP regimes. In the easy setting, variances along shared and domain-specific top directions are comparable; in the difficult setting, shared directions have lower variance. § D.3 provides details and two more challenging configurations: a smaller invariant subspace ($m = 1$) and a configuration with only two domains ($E = 2 < k = 5$).

Fig. 3 is in agreement with the consistent recovery of $\mathcal{S}_\star$. In the easy setting, the correct block dimension is already recovered in about 64% of runs at $N = 50$ and reaches probability 1 by $N = 5000$; in the hard setting, the corresponding probability is about 24% at $N = 50$ but also reaches 1 by $N = 5000$. The median projector error decreases to about 0.02 in the easy setting and 0.04 in the difficult setting at $N = 5000$. The dashed curves in Fig. 3 show that FindS$_\star$ achieves similarly low subspace error, while its dimension-recovery error remains close to 0.05 for large sample sizes, the nominal level of the individual tests, consistent with § 5.3. Our theory (§ 5.3) for Anchor PCA does not assume Gaussianity. Indeed, § D.3.4 shows the same qualitative behavior for Gaussian-mixture samples and § D.3.5 additionally compares explained variance on an unseen domain. AnchorPCA$_\infty$ and poolPCA retain nearly identical target explained variance (EV), but AnchorPCA$_\infty$ recovers the true invariant subspace $\mathcal{S}_\star$ more accurately than poolPCA when using each method's first m directions.

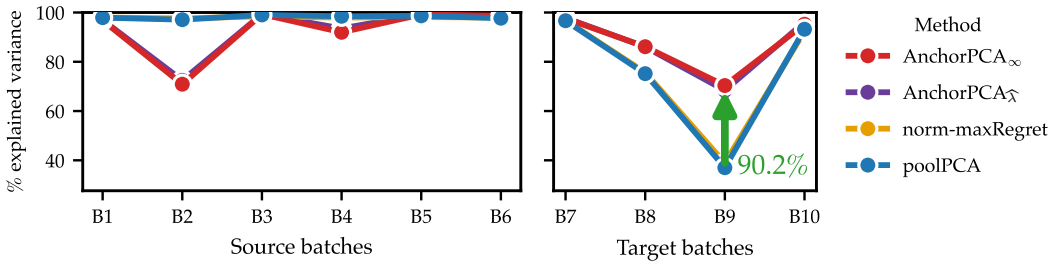

Figure 4: **Gas-sensor drift, source B1–B6 and target B7–B10.** All methods use $k = 20$ and source-only preprocessing/fitting; `norm-maxRegret` is the normalized regret wcPCA baseline [17]. `AnchorPCA`$_{\widehat{\lambda}}$ (which chooses $2E\lambda = 100$) exhibits the source–target compromise: it retains more source explained variance than `AnchorPCA`$_\infty$ but yields smaller average target gains. `AnchorPCA`$_\infty$ gives up more source explained variance, achieves the highest mean target EV, and improves over `poolPCA` on every held-out target batch; the arrow marks its largest relative target gain over `poolPCA` (B9, 90.2%).

6.2 Gas sensor array drift

We evaluate `AnchorPCA` on gas-sensor drift data [15, 46], consisting of 13,910 recordings from 16 chemical sensors, summarized by $p = 128$ features and grouped into ten temporal batches. We treat batches as domains: source batches are used for fitting and later batches are held out. The data are known to exhibit sensor drift, i.e., gradual changes in sensor response caused by aging, poisoning, or environmental effects, so that the same gas can induce different response distributions over time [15, 47]. The hope is that the projector agreement of `AnchorPCA` across early batches downweights drift-specific high-variance directions and retains directions that remain informative under future temporal shifts. We evaluate by normalized batch explained variance $100 \cdot \mathrm{Tr}(W^\top \widehat{\Sigma}_b W)/\,\mathrm{Tr}(\widehat{\Sigma}_b)$, where W is fit only on source batches.

Fig. 4 shows the results for the split B1–B6 $\to$ B7–B10 for $k = 20$. `AnchorPCA`$_{\widehat{\lambda}}$ (which chooses $2E\lambda = 100$) illustrates the following trade-off: its mean source explained variance of 93.4% lies between those of `poolPCA` (98.1%) and `AnchorPCA`$_\infty$ (92.8%), but its mean target explained variance (87.1%) outperforms `poolPCA` (75.5%). `AnchorPCA`$_\infty$ has the weakest source fit (mainly due to B2) but yields a slightly higher target explained variance (87.3%) and improves over `poolPCA` on every held-out target batch. In § D.4, we consider alternative splits and values of k: `AnchorPCA`$_{\widehat{\lambda}}$ has higher mean target EV than `poolPCA` in 14/18 configurations, while `AnchorPCA`$_\infty$ improves in 16/18. In § D.4.1 we examine the 2 configurations ($s = 8$ and $k = 20, 30$), where `poolPCA` outperforms Anchor PCA. We also examine weighting domains in proportion to sample size in § D.4.2; under both schemes, `AnchorPCA`$_\infty$ has higher target EV than `poolPCA` when averaged over the rolling splits, at all three ranks.

To assess whether the representations also help a downstream task, we predict the six gas classes. Linear classifiers on the 20-dimensional representations raise mean target accuracy from 68.6% for `poolPCA` to 74.6% for `AnchorPCA`$_{\widehat{\lambda}}$ and 75.6% for `AnchorPCA`$_\infty$, competitive with unreduced features (§ D.4.3).

7 Discussion and future work

We introduced Anchor PCA, a novel framework for PCA on multiple domains. We prove recovery guarantees and robustness results and show its effectiveness on simulated and real data. Our invariance notion (membership in the local top-k principal subspace of every domain) depends on the choice of k and can be sensitive to near eigenvalue ties. The minimax guarantee in Thm. 4.1 covers bounded variance inflations inside local top-k eigenspaces, but not mean shifts, variance deflations, or rotations of the principal subspaces. While we have focused on robust linear dimension reduction based on PCA, we discuss possible nonlinear extensions of `AnchorPCA`$_\lambda$ based on autoencoders in § E.

Acknowledgments

The authors used GPT-5.4–6.1 and Claude Opus 4.6–5.5 for assistance with manuscript drafting, proofreading of human-written text and mathematical proofs, and development of simulations, figures, and diagrams, but take full responsibility for the content. JvK is supported by The Branco Weiss Fellowship—Society in Science.

References

[1] Abubakar Abid, Martin J. Zhang, Vivek K. Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134, 2018. doi: 10.1038/s41467-018-04608-8. 16

[2] Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2704–2729. PMLR, 2022. 46

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 17

[4] Krishna B. Athreya and Soumendra N. Lahiri. *Measure Theory and Probability Theory*. Springer Texts in Statistics. Springer, New York, 2006. doi: 10.1007/978-0-387-35434-7. 24

[5] Rudolf Beran and Muni S. Srivastava. Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 13(1):95–115, 1985. doi: 10.1214/aos/1176346579. 8

[6] Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, 1997. doi: 10.1007/978-1-4612-0653-8. 17

[7] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441. 29, 30

[8] Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966. doi: 10.1207/s15327906mbr0102_10. 8

[9] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2022. doi: 10.1109/TPAMI.2021.3094760. 17

[10] Petros Drineas and Ilse C. F. Ipsen. Low-rank matrix approximations do not need a singular value gap. *SIAM Journal on Matrix Analysis and Applications*, 40(1):299–319, 2019. doi: 10.1137/18M1163658. 23

[11] Eugene F. Fama and Kenneth R. French. 49 Industry Portfolios [Daily] [Data set]. Kenneth R. French Data Library, Tuck School of Business, Dartmouth College, 2026. Accessed July 26, 2026. 43

[12] Qing Feng, Meilei Jiang, Jan Hannig, and J. S. Marron. Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265, 2018. doi: 10.1016/j.jmva.2018.03.008. 16

[13] Bernhard N. Flury. Common principal components in K groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984. doi: 10.1080/01621459.1984.10477108. 16

[14] Bernhard N. Flury and Walter Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986. doi: 10.1137/0907013. 16

[15] Jordi Fonollosa, Irene Rodríguez-Luján, and Ramón Huerta. Chemical gas sensor array dataset. *Data in Brief*, 3:85–89, 2015. doi: 10.1016/j.dib.2015.01.003. 10, 38

[16] Francesco Freni, Anya Fries, Linus Kühne, Markus Reichstein, and Jonas Peters. Maximum risk minimization with random forests. *arXiv preprint arXiv:2512.10445*, 2025. 46

[17] Anya Fries, Markus Reichstein, David Blei, and Jonas Peters. Worst-case low-rank approximations. *arXiv preprint arXiv:2603.11304*, 2026. 1, 6, 10, 16, 33, 39, 43

[18] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. 46

[19] Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943. doi: 10.2307/1905714. 16

[20] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. 45

[21] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. doi: 10.1017/CBO9780511810817. 20, 23, 25, 26, 27, 32

[22] Aapo Hyvärinen. What is the goal of unsupervised machine learning? *arXiv preprint arXiv:2610.11697*, 2026. 1

[23] Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2 edition, 2002. doi: 10.1007/b98835. 1

[24] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 2021. 42

[25] Margherita Lazzaretto, Jonas Peters, and Niklas Pfister. Invariant subspace decomposition. *Journal of Machine Learning Research*, 26(95):1–56, 2025. 17

[26] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, 2013. doi: 10.1214/12-AOAS597. 16

[27] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022. 17

[28] Mirco Migliavacca, Talie Musavi, Miguel D. Mahecha, Jacob A. Nelson, Jürgen Knauer, Dennis D. Baldocchi, Oscar Perez-Priego, Rune Christiansen, Jonas Peters, et al. The three major axes of terrestrial ecosystem function. *Nature*, 598(7881):468–472, 2021. doi: 10.1038/s41586-021-03939-9. 1

[29] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18. PMLR, 2013. 16

[30] Yotam Norman and Ron Meir. Unsupervised representation learning – an invariant risk minimization perspective. In *International Conference on Learning Representations*, 2026. 1, 16

[31] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008. doi: 10.1038/nature07331. 1

[32] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. doi: 10.1109/TNN.2010.2091281. 16

[33] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161. 16

[34] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167. 16

[35] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586. 45

[36] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. 17

[37] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B*, 83(2):215–246, 2021. doi: 10.1111/rssb.12398. 1, 2, 17, 34

[38] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0. 45

[39] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. 42

[40] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 15

[41] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1255–1262. Omnipress, 2012. 16

[42] James R. Schott. Partial common principal component subspaces. *Biometrika*, 86(4):899–908, 1999. doi: 10.1093/biomet/86.4.899. 8, 9, 16, 21

[43] Xinwei Shen and Nicolai Meinshausen. Distributional principal autoencoders. *arXiv preprint arXiv:2404.13649*, 2024. 16, 46

[44] Uthaipon Tantipongpipat, Samira Samadi, Mohit Singh, Jamie Morgenstern, and Santosh Vempala. Multi-criteria dimensionality reduction with applications to fairness. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 16

[45] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. doi: 10.1162/jocn.1991.3.1.71. 1

[46] Alexander Vergara. Gas sensor array drift at different concentrations. UCI Machine Learning Repository, 2012. Dataset. 10, 38, 39

[47] Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margie A. Ryan, Margie L. Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166–167:320–329, 2012. doi: 10.1016/j.snb.2012.01.074. 10

[48] Bingkai Wang, Xi Luo, Yi Zhao, and Brian Caffo. Semiparametric partial common principal component analysis for covariance matrices. *Biometrics*, 77(4):1175–1186, 2021. doi: 10.1111/biom.13369. 16

[49] Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23018–23033. PMLR, 17–23 Jul 2022. 1, 17

[50] Kexin Wang, Salil Bhate, João M. Pereira, Joe Kileel, Matylda Figlerowicz, and Anna Seigal. Multi-context principal component analysis. *arXiv preprint arXiv:2601.15239*, 2026. 16

[51] Zhenyu Wang, Molei Liu, Jing Lei, Francis Bach, and Zijian Guo. StablePCA: Distributionally robust learning of shared representations from multi-source data. *arXiv preprint arXiv:2505.00940*, 2025. 1, 16

[52] Ian J. Wright, Peter B. Reich, Mark Westoby, David D. Ackerly, Zdravko Baruch, et al. The worldwide leaf economics spectrum. *Nature*, 428(6985):821–827, 2004. doi: 10.1038/nature02403. 1

[53] Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008. 23

Contents of the Appendix

A Further related work **15**

A.1 Multi-domain dimension reduction 15

A.2 Invariance for prediction 16

B Further details on Anchor PCA **17**

B.1 Ordered bases for direction-level statements 17

B.2 Additional theoretical results 18

B.3 Test statistic of $\texttt{FindS}_\star$ 21

C Proofs **23**

C.1 Additional intermediate results 23

C.2 Proof of Thm. 3.1 26

C.3 Proof of Thm. 4.1 29

C.4 Consistency of Anchor PCA 30

D Details on experiments **33**

D.1 Details of the motivating example 33

D.2 Details on the perturbation-path illustration 33

D.3 Details on the simulations 34

D.4 Gas sensor array drift details and additional analyses 38

D.5 Fama–French industry portfolios 43

E Nonlinear extensions **45**

E.1 Vanilla autoencoder (AE) 45

E.2 Anchor autoencoder (`AnchorAE`) 45

E.3 Distributional principal autoencoder (DPA) 46

E.4 Multi-domain DPA 46

A Further related work

`AnchorPCA` is a dimensionality reduction over several source domains that builds on ideas of invariance. We summarize work in multi-domain dimension reduction and related ideas in prediction that use invariance.

A.1 Multi-domain dimension reduction

Multi-domain dimension reduction can be separated into three groups: worst-case methods, invariance-based methods, and other methods (e.g., those which do not relate to generalization, but consider multiple groups and PCA).

Worst-case methods. Worst-case methods replace pooled explained variance or reconstruction by a max-min (or min-max) objective, which seeks to optimize performance on the worst domain, group, or criterion. FairPCA [40] minimizes the maximum increase in reconstruction error incurred by using

a shared projection instead of the domain-specific optimal projection. Multi-criteria dimensionality reduction [44] generalizes this idea to several criteria over multiple groups. However, neither provides a robustness perspective. In contrast, wcPCA [17] includes out-of-sample guarantees for target covariances in the convex hull of the observed source covariances. Along similar lines, StablePCA [51] casts multi-source PCA as a group-DRO problem. $\texttt{AnchorPCA}_\lambda$ is also worst-case optimal, but over a different uncertainty set of perturbed covariances (Thm. 4.1). Importantly, it does not only optimize a worst-case objective: it also recovers the maximal invariant subspace $\mathcal{S}_\star$ (Thm. 3.1), a target none of the methods above consider.

Invariance-based methods. Invariance-based dimension reduction methods seek representations whose distributions, moments, or predictive relations are stable across domains, often with the goal of improving downstream prediction for a target Y. Transfer Component Analysis (TCA) [32] aligns source and target distributions through maximum mean discrepancy (MMD) criteria while preserving variance; it requires target domain information while Anchor PCA does not. Domain-Invariant Component Analysis (DICA) [29] learns transformations that reduce dissimilarity across training domains while preserving predictive structure; it requires a downstream task and response Y, while Anchor PCA is fully unsupervised. In contrast, Unsupervised DICA (UDICA) [29] and Principal Invariant Component Analysis (PICA) [30] do not require information about a response Y or the target domain. UDICA retains variance in kernel features while penalizing the distances between the mean embeddings of the source domains. PICA finds linear components whose projections have the same distribution across domains. For a single, unit norm component v, with centered Gaussian data, PICA's objective is to maximize pooled explained variance subject to $v^\top \Sigma_e v = v^\top \Sigma_f v$ for all source domains e, f. Both PICA and UDICA aim for distributional invariance; we instead consider a geometric notion of invariance, namely agreement between the domain-specific top-k principal projectors. A direction can therefore be invariant even if its variance, or its rank within the top k, differs across domains. Additionally, unlike PICA, which imposes exact invariance constraints, we always return a non-empty subspace of the chosen dimension k, even when $\mathcal{S}_\star = \{0\}$. While DICA provides generalization guarantees for downstream prediction, TCA, DICA, UDICA, and PICA do not provide unsupervised PCA-style robustness guarantees for explained variance or reconstruction of the kind studied here.

Other methods. Other multi-group dimensionality reduction methods model shared structure for statistical or exploratory purposes rather than for OOD robustness. Common principal component analysis (CPCA) assumes simultaneous diagonalizability of the covariance matrices [13, 14] (that is, there exists an orthogonal U such that for all domains e, $\Sigma_e = U D_e U^\top$ with D_e diagonal and domain-specific); partial common principal component methods relax this assumption and estimate or test for shared eigendirections or shared principal subspaces [42, 48]. Our $\texttt{FindS}_\star$ procedure builds on this testing procedure but the population target is different: $\mathcal{S}_\star$ is the intersection of local top-k eigenspaces, not a set of eigenvectors that are shared as individual axes or ordered in the same way across domains. Multi-context PCA (MCPCA) [50] also seeks shared directions of variation and extends the CPCA model $\Sigma_e = U D_e U^\top$ by allowing nonorthogonal columns of U that may appear in only some domains. Wang et al. [50] establish identifiability up to sign and ordering when the columns are linearly independent and no two have proportional weights across domains (that is, $((D_e)_{jj})_{e\in\mathcal{E}} \not\propto ((D_e)_{kk})_{e\in\mathcal{E}}$ for all $j \neq k$). Anchor PCA instead learns a single subspace by balancing explained variance and agreement with the local top-k subspaces, without assuming this covariance factorization. Multi-block methods such as JIVE [26] and AJIVE [12] decompose data sources into joint and individual low-rank variation, whereas contrastive PCA [1] searches for directions enriched in one dataset relative to a background dataset. These methods provide useful decompositions, but they do not couple shared-structure estimation to the projector-agreement and structured-perturbation guarantees studied here. Nonlinear PCA and autoencoder variants provide another axis of generalization [43]; § E discusses how the same agreement principle could be ported to autoencoder-based representations.

A.2 Invariance for prediction

In supervised domain generalization, invariance is usually formulated as stability of a conditional law, a prediction rule, or a risk across domains. This principle is rooted in causal invariance [19, 33, 34, 41] and appears in invariant models for causal transfer learning, invariant risk minimization, invariant-feature subspace recovery, invariant causal representation learning, and invariant

subspace decomposition for time-varying prediction problems [3, 25, 27, 36, 49]. Several prediction methods relax strict invariance rather than enforcing it exactly: for example, anchor regression interpolates between ordinary least squares and an invariant solution through a penalty parameter [9, 37]. $\texttt{AnchorPCA}_\lambda$ imports this soft-invariance logic into unsupervised dimension reduction: finite λ penalizes projector disagreement, while $\texttt{AnchorPCA}_\infty$ recovers the hard agreement target and, when it exists, the maximal invariant subspace $\mathcal{S}_\star$.

B Further details on Anchor PCA

B.1 Ordered bases for direction-level statements

The optimization problems defining $\texttt{AnchorPCA}_\lambda$ (Defn. 2.3) and $\texttt{AnchorPCA}_\infty$ (Defn. 2.2) depend on W only through the projection matrix $\Pi_W = WW^\top$. Hence, the canonical population objects are solution subspaces or solution projectors. However, the projection-based formulation allows us to fix an ordered representation, see § 2.3. We now state their precise construction.

B.1.1 $\texttt{AnchorPCA}_\lambda$

Solutions of $\texttt{AnchorPCA}_\lambda$ can be obtained by solving rank-k PCA on $M_\lambda = \overline{\Sigma} + 2E\lambda\,\overline{\Pi}$ (§ 2.3). Hence, we use the standard PCA convention and choose $W_k^{\perp,\lambda} = (w_1^{\perp,\lambda}, \ldots, w_k^{\perp,\lambda})$ as an ordered orthonormal eigenbasis of M_λ associated with its k largest eigenvalues. Thus, $w_1^{\perp,\lambda}$ maximizes $w^\top M_\lambda w$ over unit vectors, $w_2^{\perp,\lambda}$ maximizes the same quantity subject to being orthogonal to $w_1^{\perp,\lambda}$, and so on. If M_λ has eigenvalue ties, the ordered basis is not unique and we choose any orthonormal basis inside tied eigenspaces. If all ties occur within the selected top-k spectrum, this affects only the solution basis, not the rank-k solution subspace and projector. If there is no eigengap at rank-k, i.e. $\lambda_k(M_\lambda) = \lambda_{k+1}(M_\lambda)$, then the rank-$k$ solution projector is non-unique, and any maximizing rank-k solution projector may be used.

B.1.2 $\texttt{AnchorPCA}_\infty$

We now define an analogous ordered basis for $\texttt{AnchorPCA}_\infty$ in a two-stage construction. First, we determine the solution subspace: directions are ordered by agreement with the domain-specific top-k principal subspaces (equivalently, by decreasing eigenvalue of $\overline{\Pi}$). Second, within each eigenspace, we choose a basis by iteratively maximizing the pooled explained variance over directions in that block.

Step 1a: Ordering by agreement. The agreement constraint of $\texttt{AnchorPCA}_\infty$ in (1) is equivalent to maximizing

$$\mathrm{Tr}(W^\top \overline{\Pi} W) = \sum_{i=1}^{k} w_i^\top \overline{\Pi} w_i \tag{7}$$

over $W = [w_1, \ldots, w_k] \in \mathcal{O}_{p\times k}$ (§ 2.3). We now characterize this solution. Let $\rho_1 > \rho_2 > \cdots > \rho_L$ be the distinct eigenvalues of $\overline{\Pi}$, and let

$$\mathrm{Eig}_\ell := \ker(\overline{\Pi} - \rho_\ell I_p), \qquad d_\ell := \dim(\mathrm{Eig}_\ell).$$

Eig_ℓ is the eigenspace of $\overline{\Pi}$ corresponding to the eigenvalue ρ_ℓ and d_ℓ is its dimension. Larger ρ_ℓ implies larger average agreement with the domain-specific top-k principal subspaces. Then, let r be the unique index satisfying

$$\sum_{\ell<r} d_\ell < k \le \sum_{\ell\le r} d_\ell.$$

By Ky Fan's maximum principle [6, Problem I.6.15], the trace in (7) is maximized precisely when the columns of W are drawn from the leading eigenspaces of $\overline{\Pi}$: every maximizer spans a k-dimensional subspace of the form

$$\mathrm{Eig}_1 \oplus \cdots \oplus \mathrm{Eig}_{r-1} \oplus \mathcal{U}_r,$$

where $\mathcal{U}_r \le \mathrm{Eig}_r$ is an arbitrary subspace of dimension $q := k - \sum_{\ell<r} d_\ell$.

Step 1b: Determining $\mathcal{U}_r$. The eigenspaces $\mathrm{Eig}_1, \ldots, \mathrm{Eig}_{r-1}$ are selected completely, and only the final eigenspace $\mathcal{U}_r$ can be determined via the secondary $\overline{\Sigma}$-criterion of (2.2), which chooses $\mathcal{U}_r$ to maximize pooled explained variance among all q-dimensional subspaces of Eig_r. Equivalently, inside Eig_r, the subspace $\mathcal{U}_r$ is chosen by the top-q eigenspace of $\Pi_{\mathrm{Eig}_r} \overline{\Sigma} \Pi_{\mathrm{Eig}_r}$.

Step 2: Choosing an ordered basis. Within each fully selected eigenspace $\mathrm{Eig}_1, \ldots, \mathrm{Eig}_{r-1}$, we choose an orthonormal eigenbasis ordered by non-increasing eigenvalue of $\overline{\Sigma}$, and concatenate these bases in decreasing order of ρ_ℓ. We then append an ordered top-q eigenbasis of $\overline{\Sigma}$ within Eig_r. With this convention, $\texttt{AnchorPCA}_\infty$ directions are ordered first by invariance, that is, by decreasing $\overline{\Pi}$-eigenvalue. Inside a fixed eigenspace, they are ordered by pooled explained variance.

Remark B.1 (Handling non-uniqueness.)**.** If $\overline{\Sigma}$ has eigenvalue ties in one of the fully selected eigenspaces $\mathrm{Eig}_1, \ldots, \mathrm{Eig}_{r-1}$, the basis within the tied eigenspace is not unique; we make an arbitrary choice, and this affects only direction-level quantities, not the solution subspace or projector. If instead $\Pi_{\mathrm{Eig}_r} \overline{\Sigma} \Pi_{\mathrm{Eig}_r}$ has ties among its top q eigenvalues, the subspace $\mathcal{U}_r$ itself is not unique, and the rank-k $\texttt{AnchorPCA}_\infty$ solution subspace and projector are likewise non-unique; the uniqueness statements elsewhere in the paper (e.g., Thm. C.11(ii)) exclude this case by assumption.

B.2 Additional theoretical results

Notation. Throughout the appendix, all matrices are represented with respect to the standard basis of $\mathbb{R}^p$, unless otherwise stated. For pairwise orthogonal subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_j \subseteq \mathbb{R}^p$, we write $\mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_j$ for their orthogonal direct sum. Suppose that a symmetric matrix $M \in \mathbb{R}^{p \times p}$ preserves each $\mathcal{S}_i$, that is, $Mv \in \mathcal{S}_i \ \forall i \in \{1, \ldots, j\}$ and $\forall v \in \mathcal{S}_i$. Then we can block-diagonalize M with respect to $(\mathcal{S}_i)_{i=1}^j$ as follows. Let $B_i \in \mathbb{R}^{p \times d_i}$ be an orthonormal basis matrix for $\mathcal{S}_i$, where $d_i := \dim(\mathcal{S}_i)$, and let $Q := [B_1 \ \cdots \ B_j]$. In this aligned basis, M is block diagonal, that is

$$Q^\top M Q = \mathrm{diag}(A_1, \ldots, A_j) = \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_j \end{bmatrix}, \qquad A_i := B_i^\top M B_i.$$

We abbreviate this block-diagonal representation by writing $M = A_1 \oplus \cdots \oplus A_j$ with respect to the decomposition $\mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_j$.

B.2.1 $\texttt{AnchorPCA}_\infty$ as a limit of $\texttt{AnchorPCA}_\lambda$

Intuitively, as the penalty λ grows, solutions of $\texttt{AnchorPCA}_\lambda$ approach the hard-agreement solutions of $\texttt{AnchorPCA}_\infty$. The following proposition states this at the level of projection matrices. Thm. C.8 gives a complementary result at the level of individual directions, showing that the leading $m = \dim(\mathcal{S}_\star)$ $\texttt{AnchorPCA}_\lambda$ directions converge to $\mathcal{S}_\star$.

Proposition B.2 (Large-penalty limit)**.** *For all $\lambda \geq 0$, let $\Pi_k^{\perp,\lambda}$ be any rank-k $\texttt{AnchorPCA}_\lambda$ solution projector. If $\lambda_n \to \infty$, every accumulation point of $(\Pi_k^{\perp,\lambda_n})$ is a rank-k $\texttt{AnchorPCA}_\infty$ solution projector.*

If, in addition, the population rank-k $\texttt{AnchorPCA}_\infty$ solution projector is unique, denote it by $\Pi_k^{\perp,\infty}$, then

$$\left\| \Pi_k^{\perp,\lambda} - \Pi_k^{\perp,\infty} \right\|_{\mathrm{op}} \to 0 \qquad \text{as } \lambda \to \infty.$$

Proof. Let

$$\mathcal{P}_k := \{X \in \mathbb{R}^{p \times p} : X = WW^\top, \ W \in \mathcal{O}_{p \times k}\}$$

denote the set of rank-k orthogonal projectors. Since $\mathcal{P}_k$ is closed and bounded in $\mathbb{R}^{p \times p}$, it is compact.

For $X \in \mathcal{P}_k$, define

$$H(X) := \mathrm{Tr}(\overline{\Sigma} X), \qquad G(X) := \mathrm{Tr}(\overline{\Pi} X), \qquad F_\lambda(X) := H(X) + 2E\lambda\, G(X).$$

Since both $\texttt{AnchorPCA}$ objectives depend on W only through $WW^\top$, rank-k $\texttt{AnchorPCA}_\lambda$ is the problem of maximizing F_λ over $\mathcal{P}_k$, while rank-k $\texttt{AnchorPCA}_\infty$ is the problem of first maximizing G over $\mathcal{P}_k$ and then maximizing H among the maximizers of G.

Fix a sequence $\lambda_n \to \infty$, and for each n, let $\Pi_k^{\perp,\lambda_n} \in \mathcal{P}_k$ be a rank-k $\texttt{AnchorPCA}_\lambda$ solution projector at penalty λ_n. By compactness of $\mathcal{P}_k$, the sequence $(\Pi_k^{\perp,\lambda_n})$ has accumulation points in $\mathcal{P}_k$. Let $\Pi_\star$ be any such accumulation point; passing to a subsequence (and relabeling), $\left\|\Pi_k^{\perp,\lambda_n} - \Pi_\star\right\|_{\mathrm{op}} \to 0$.

Step 1: $\Pi_\star$ maximizes G over $\mathcal{P}_k$. By contradiction, assume that $\Pi_\star$ does not maximize G and that there exist $Y \in \mathcal{P}_k$ and $\delta > 0$ such that $G(Y) \geq G(\Pi_\star) + 2\delta$. By continuity of G,

$$G(\Pi_k^{\perp,\lambda_n}) \leq G(\Pi_\star) + \delta$$

for all sufficiently large n. Further, for all $X \in \mathcal{P}_k$, $|H(X)| = |\operatorname{Tr}(\overline{\Sigma}X)| \leq k \left\|\overline{\Sigma}\right\|_{\mathrm{op}}$ (because X is a rank-k orthogonal projector), and hence $H(Y) - H(\Pi_k^{\perp,\lambda_n}) \geq -2k \left\|\overline{\Sigma}\right\|_{\mathrm{op}}$. Therefore, for all sufficiently large n,

$$\begin{aligned} F_{\lambda_n}(Y) - F_{\lambda_n}(\Pi_k^{\perp,\lambda_n}) &= \big(H(Y) - H(\Pi_k^{\perp,\lambda_n})\big) + 2E\lambda_n\big(G(Y) - G(\Pi_k^{\perp,\lambda_n})\big) \\ &\geq -2k \left\|\overline{\Sigma}\right\|_{\mathrm{op}} + 2E\lambda_n\delta. \end{aligned}$$

where $G(Y) - G(\Pi_k^{\perp,\lambda_n}) \geq G(\Pi_\star) + 2\delta - (G(\Pi_\star) + \delta) = 2\delta - \delta = \delta$. Since $\lambda_n \to \infty$, the right-hand side eventually is positive, contradicting the optimality of $\Pi_k^{\perp,\lambda_n}$. Thus, $\Pi_\star$ maximizes G over $\mathcal{P}_k$.

Step 2: $\Pi_\star$ maximizes H among the maximizers of G. By contradiction, assume that there exists $Y \in \mathcal{P}_k$ and $\delta > 0$ such that Y maximizes G and $H(Y) \geq H(\Pi_\star) + 2\delta$. Since Y maximizes G over $\mathcal{P}_k$, $G(Y) \geq G(\Pi_k^{\perp,\lambda_n})$ for all n. Further, by continuity of H,

$$H(\Pi_k^{\perp,\lambda_n}) \leq H(\Pi_\star) + \delta$$

for all sufficiently large n. Hence, for all sufficiently large n,

$$F_{\lambda_n}(Y) - F_{\lambda_n}(\Pi_k^{\perp,\lambda_n}) = \big(H(Y) - H(\Pi_k^{\perp,\lambda_n})\big) + 2E\lambda_n\big(G(Y) - G(\Pi_k^{\perp,\lambda_n})\big) \geq \delta,$$

which contradicts the optimality of $\Pi_k^{\perp,\lambda_n}$. Therefore, $\Pi_\star$ is a rank-k $\texttt{AnchorPCA}_\infty$ solution projector.

Step 3: full convergence under uniqueness. Assume now that the rank-k $\texttt{AnchorPCA}_\infty$ solution projector is unique and equals $\Pi_k^{\perp,\infty}$. By contradiction, assume that there exist $\varepsilon > 0$ and a sequence $\lambda_n \to \infty$ such that

$$\left\|\Pi_k^{\perp,\lambda_n} - \Pi_k^{\perp,\infty}\right\|_{\mathrm{op}} \geq \varepsilon \qquad \text{for all } n.$$

By compactness of $\mathcal{P}_k$, the sequence $(\Pi_k^{\perp,\lambda_n})$ has an accumulation point $\Pi_\star$. By the first part of the proposition, $\Pi_\star$ is a rank-k $\texttt{AnchorPCA}_\infty$ solution projector, hence $\Pi_\star = \Pi_k^{\perp,\infty}$, a contradiction. Therefore

$$\left\|\Pi_k^{\perp,\lambda} - \Pi_k^{\perp,\infty}\right\|_{\mathrm{op}} \to 0.$$

□

B.2.2 Recovery of $\mathcal{S}_\star$ by $\texttt{AnchorPCA}_\lambda$ with finite penalty λ

The following proposition shows that $\texttt{AnchorPCA}_\lambda$ recovers the invariant subspace at two levels: if $\lambda > \lambda^\star$, every rank-k solution subspace contains $\mathcal{S}_\star$, and if $\lambda > \lambda^{\star\star}$, the leading rank-$m$ $\texttt{AnchorPCA}_\lambda$ directions (in the convention of § B.1) span $\mathcal{S}_\star$.

Proposition B.3 (Finite-penalty recovery). *Assume $1 \leq m \leq k < p$ and that $\overline{\Sigma}$ is block-diagonal with respect to $\mathcal{S}_\star \oplus \mathcal{S}_\star^\perp$. Let $\overline{\Sigma}_{\mathcal{S}_\star} := \Pi_{\mathcal{S}_\star}\overline{\Sigma}\Pi_{\mathcal{S}_\star}$ and $\overline{\Sigma}_{\mathcal{S}_\star^\perp} := \Pi_{\mathcal{S}_\star^\perp}\overline{\Sigma}\Pi_{\mathcal{S}_\star^\perp}$ be the diagonal blocks of $\overline{\Sigma}$ with respect to $\mathcal{S}_\star \oplus \mathcal{S}_\star^\perp$ and let $\overline{R} := \Pi_{\mathcal{S}_\star^\perp}\overline{\Pi}\Pi_{\mathcal{S}_\star^\perp}$. If $\Pi_k^{\perp,\lambda}$ is a rank-k $\texttt{AnchorPCA}_\lambda$ solution projector, then the following statements hold.*

(i) If $\lambda > \lambda^\star := \big(\lambda_1(\overline{\Sigma}_{\mathcal{S}_\star^\perp}) - \lambda_m(\overline{\Sigma}_{\mathcal{S}_\star})\big)/2E\big(1 - \lambda_{k-m+1}(\overline{R})\big)$, then $\Pi_k^{\perp,\lambda}\Pi_{\mathcal{S}_\star} = \Pi_{\mathcal{S}_\star}$, equivalently $\mathcal{S}_\star \subseteq \mathcal{S}_k^{\perp,\lambda}$.

(ii) If $\lambda > \lambda^{\star\star} := \big(\lambda_1(\overline{\Sigma}_{\mathcal{S}_\star^\perp}) - \lambda_m(\overline{\Sigma}_{\mathcal{S}_\star})\big)/2E\big(1-\lambda_1(\overline{R})\big)$*, then the top* m*-dimensional eigenspace of* $M_\lambda = \overline{\Sigma} + 2E\lambda\overline{\Pi}$ *is* $\mathcal{S}_\star$ *.*

Proof. Since rank-k `AnchorPCA`$_\lambda$ is rank-k PCA on $M_\lambda = \overline{\Sigma} + 2E\lambda\,\overline{\Pi}$, the proof proceeds by block-diagonalizing M_λ along $\mathcal{S}_\star \oplus \mathcal{S}_\star^\perp$ and then showing that, once λ is large enough, the $\mathcal{S}_\star$-block eigenvalues dominate the relevant $\mathcal{S}_\star^\perp$-block eigenvalues.

Step 1: block-diagonalize M_λ along $\mathcal{S}_\star \oplus \mathcal{S}_\star^\perp$. Let $S \in \mathcal{O}_{p\times m}$ and $Z \in \mathcal{O}_{p\times(p-m)}$ have orthonormal columns respectively spanning $\mathcal{S}_\star$ and $\mathcal{S}_\star^\perp$. Let

$$A := S^\top\overline{\Sigma}S, \qquad B := Z^\top\overline{\Sigma}Z, \qquad R := Z^\top\overline{\Pi}Z$$

Then, in the orthonormal coordinates (S, Z),

$$(S,Z)^\top M_\lambda (S,Z) = \begin{pmatrix} A + 2E\lambda I_m & 0 \\ 0 & B + 2E\lambda R \end{pmatrix}$$

because $S^\top\overline{\Sigma}Z = 0$ (because, by assumption, $\overline{\Sigma}$ is block-diagonal with respect to $\mathcal{S}_\star \oplus \mathcal{S}_\star^\perp$) and $S^\top\overline{\Pi}S = I_m$ and $S^\top\overline{\Pi}Z = 0$ (as $\mathcal{S}_\star$ is the eigenspace of $\overline{\Pi}$ associated with eigenvalue 1).

Step 2: compare the top eigenvalues across the two blocks. First, observe that

$$\overline{\Sigma}_{\mathcal{S}_\star} = SAS^\top, \qquad \overline{\Sigma}_{\mathcal{S}_\star^\perp} = ZBZ^\top, \qquad \overline{R} = ZRZ^\top.$$

Hence the eigenvalues of A, B and R that we will require are as follows

$$\lambda_m(A) = \lambda_m(\overline{\Sigma}_{\mathcal{S}_\star}), \qquad \lambda_1(B) = \lambda_1(\overline{\Sigma}_{\mathcal{S}_\star^\perp}), \qquad \lambda_i(R) = \lambda_i(\overline{R}),\ i \in \{1,\dots,p-m\}.$$

Furthermore, $\lambda_1(R) < 1$: to prove this, assume, by contradiction, that $Rv = v$ for some nonzero $v \in \mathbb{R}^{p-m}$, and set $z := Zv \in \mathcal{S}_\star^\perp$, so that $\|z\|_2^2 = \|v\|_2^2$. Since $R = Z^\top\overline{\Pi}Z$, $z^\top\overline{\Pi}z = v^\top Rv = \|v\|_2^2 = \|z\|_2^2 = z^\top z$. As $0 \preceq \overline{\Pi} \preceq I_p$, $I_p - \overline{\Pi}$ is positive semi-definite and $z^\top(I_p - \overline{\Pi})z = 0$ gives $\overline{\Pi}z = z$, i.e. $\Pi_k^{(e)}z = z$ for every e, so $z \in \mathcal{S}_\star$ — contradicting $0 \neq z \in \mathcal{S}_\star^\perp$.

The smallest eigenvalue of the invariant block and the eigenvalues of the non-invariant block are thus

$$a_m(\lambda) := \lambda_m(A) + 2E\lambda, \qquad b_j(\lambda) := \lambda_j(B + 2E\lambda R) \text{ for } j \in \{1,\dots,p-m\}.$$

Step 3: show inclusion of $\mathcal{S}_\star$ for $\lambda > \lambda^\star$ (part (i)). By Weyl's inequality [21, Theorem 4.3.7] used in decreasing-order notation, applied to $B,\ 2E\lambda R$ with indices $(1, k-m+1)$ (recall that we are using decreasing order notation of eigenvalues, whereas Thm. 4.3.7 in [21] uses increasing order),

$$b_{k-m+1}(\lambda) \le \lambda_1(B) + 2E\lambda\,\lambda_{k-m+1}(R).$$

Since $\lambda^\star = \frac{\lambda_1(B)-\lambda_m(A)}{2E\{1-\lambda_{k-m+1}(R)\}}$, the condition $\lambda > \lambda^\star$ is equivalent to

$$2E\lambda\{1-\lambda_{k-m+1}(R)\} > \lambda_1(B) - \lambda_m(A).$$

and therefore

$$a_m(\lambda) = \lambda_m(A) + 2E\lambda > \lambda_1(B) + 2E\lambda\,\lambda_{k-m+1}(R) \ge b_{k-m+1}(\lambda).$$

Hence at most $k-m$ eigenvalues from the non-invariant coordinate block are at least $a_m(\lambda)$, whereas all m eigenvalues from the invariant coordinate block are at least $a_m(\lambda)$. Therefore every top-k eigenspace of M_λ contains $\mathcal{S}_\star$, that is, $\Pi_k^{\perp,\lambda}\Pi_{\mathcal{S}_\star} = \Pi_{\mathcal{S}_\star}$.

Step 4: show that the leading m directions span $S^\star$ for $\lambda > \lambda^{\star\star}$ (part (ii)). Again by Weyl's inequality [21, Theorem 4.3.7] used in decreasing-order notation, with indices $(1, 1)$,

$$b_1(\lambda) \le \lambda_1(B) + 2E\lambda\,\lambda_1(R).$$

Since $\lambda^{\star\star} = \frac{\lambda_1(B)-\lambda_m(A)}{2E\{1-\lambda_1(R)\}}$, the condition $\lambda > \lambda^{\star\star}$ is equivalent to

$$2E\lambda\{1-\lambda_1(R)\} > \lambda_1(B) - \lambda_m(A),$$

and therefore

$$a_m(\lambda) = \lambda_m(A) + 2E\lambda > \lambda_1(B) + 2E\lambda\,\lambda_1(R) \ge b_1(\lambda).$$

Thus every eigenvalue of the invariant block strictly exceeds every eigenvalue of the non-invariant block. Consequently, the leading m-dimensional eigenspace of M_λ is $\mathcal{S}_\star$.

□

Remark B.4 (Basis-level interpretation)**.** Part (ii) has a basis-level interpretation: the first m directions of any ordered $\texttt{AnchorPCA}_\lambda$ representative basis (in the sense of § B.1) span $\mathcal{S}_\star$ and diagonalize $\overline{\Sigma}$ within $\mathcal{S}_\star$. Under the weaker threshold in part (i), only the subspace-level statement $\mathcal{S}_\star \subseteq \mathcal{S}_k^{\perp,\lambda}$ is guaranteed. If the eigenvalues of an invariant block and a non-invariant block of M_λ tie, the ordered eigenbasis is determined only up to a rotation inside the tied eigenspace. Hence, the k ordered directions need not contain m vectors spanning $\mathcal{S}_\star$. Specifically, one can always choose a basis whose first m columns span $\mathcal{S}_\star$, but this need not hold for every ordered basis.

B.2.3 Invariance of $S_\star$ to variance inflation

Proposition B.5. *Assume $\lambda_k(\Sigma_e) > \lambda_{k+1}(\Sigma_e)$ for every $e \in \mathcal{E}$, and let $\mathcal{S}_\star = \bigcap_{e\in\mathcal{E}} \mathrm{Im}(\Pi_k^{(e)})$ be the maximal invariant subspace of $(\Sigma_e)_{e\in\mathcal{E}}$. Then $\mathcal{S}_\star$ is also the maximal invariant subspace of $(\Sigma_e + \Delta_e)_{e\in\mathcal{E}}$ for every choice of symmetric matrices Δ_e with $0 \preceq \Delta_e$ and $\mathrm{Im}(\Delta_e) \subseteq \mathrm{Im}(\Pi_k^{(e)})$, for all $e \in \mathcal{E}$.*

Proof. We first show that each admissible perturbation leaves the rank-k principal subspaces unchanged. Intersecting these unchanged subspaces then shows that $\mathcal{S}_\star$ is unchanged. Fix $e \in \mathcal{E}$ and write $\mathcal{S}_e := \mathrm{Im}(\Pi_k^{(e)})$. Since Δ_e is symmetric with $\mathrm{Im}(\Delta_e) \subseteq \mathcal{S}_e$, its kernel contains $\mathcal{S}_e^\perp$, so both Σ_e and Δ_e are block-diagonal with respect to $\mathbb{R}^p = \mathcal{S}_e \oplus \mathcal{S}_e^\perp$. On $\mathcal{S}_e^\perp$, the perturbed covariance $\Sigma_e + \Delta_e$ agrees with Σ_e and so has largest eigenvalue $\lambda_{k+1}(\Sigma_e)$. On $\mathcal{S}_e$, since $\Delta_e \succeq 0$, the smallest eigenvalue of $\Sigma_e + \Delta_e$ is at least $\lambda_k(\Sigma_e)$. The gap $\lambda_k(\Sigma_e) > \lambda_{k+1}(\Sigma_e)$ then forces the rank-k principal subspace of $\Sigma_e + \Delta_e$ to be exactly $\mathcal{S}_e$, so its rank-k principal projector, which we denote by $\Pi_{k,\Delta}^{(e)}$ equals $\Pi_k^{(e)}$. Intersecting over e gives that the maximal invariant subspace of the perturbed covariances is $\mathcal{S}_\star$, and averaging gives $E^{-1}\sum_{e=1}^E \Pi_{k,\Delta}^{(e)} = \overline{\Pi}$.

□

B.3 Test statistic of $\texttt{FindS}_\star$

We now state the Wald-type test statistic used in the sequential procedure for estimating $\dim(\mathcal{S}_\star)$. Since our construction is adapted from Schott's partial common principal component subspace test [42], we first recall the standard version in our notation and then explain the modification for $\mathcal{S}_\star$. We present only the reduced representation, since this is the form used in computation. To simplify notation, throughout this appendix section we hide the dependence on k. For example, we let $\Pi_e := \Pi_k^{(e)}$ be the projector onto the top-k eigenspace in domain e.

B.3.1 Standard construction.

Suppose that, for each $e \in \{1, \dots, E\}$, we observe an i.i.d. Gaussian sample with population covariance matrix Σ_e, and assume that the kth and $(k+1)$st eigenvalues of Σ_e are separated by a strict eigengap. Let Π_e denote the projector onto the top-k eigenspace of Σ_e, and define r as the dimension of the union of the top-k eigenspaces of $(\Sigma_e)_{e\in\mathcal{E}}$, that is, $r := \dim\big(\mathrm{span}\big(\mathrm{Im}(\Pi_1) \cup \cdots \cup \mathrm{Im}(\Pi_E)\big)\big)$. Schott considers the hypotheses

$$H_{0,s} : r = s \qquad \text{versus} \qquad H_{A,s} : r > s.$$

We now show how the test statistic is constructed. Let $\widehat{\Pi}_e$ be the sample projector onto the top-k eigenspace of the sample covariance matrix $\widehat{\Sigma}_e$, and let $\widehat{\Pi}_0$ be the projector onto the span of the top s eigenvectors of $\widehat{\Pi}_1 + \cdots + \widehat{\Pi}_E$. Let $\widehat{C}_e \in \mathbb{R}^{p\times k}$ be an orthonormal basis of $\mathrm{Im}(\widehat{\Pi}_e)$, and let $\widehat{C}_0 \in \mathbb{R}^{p\times(p-s)}$ be an orthonormal basis of $\mathrm{Im}(I_p - \widehat{\Pi}_0)$. The statistic depends on three objects: the deviation vector $v^\star$, its estimated asymptotic covariance matrix $\widehat{W}^\star$, and the projector $\widehat{F}$ onto the eigenspace of $\widehat{W}^\star$ corresponding to its $n_s = (Ek - s)(p - s)$ largest eigenvalues.

Deviation vector $v^\star$. Define

$$v^\star := \begin{pmatrix} \mathrm{vec}(\widehat{C}_0^\top \widehat{C}_1) \\ \vdots \\ \mathrm{vec}(\widehat{C}_0^\top \widehat{C}_E) \end{pmatrix}.$$

Covariance matrix $\widehat{W}^\star$. Here, we need several intermediate components. First, let

$$\widehat{\Pi}_\star := \sum_{j=1}^{s} \widehat{\lambda}_j \widehat{q}_j \widehat{q}_j^\top, \qquad \widehat{\Pi}_\star^+ := \sum_{j=1}^{s} \widehat{\lambda}_j^{-1} \widehat{q}_j \widehat{q}_j^\top,$$

where $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_s$ and $\widehat{q}_1, \ldots, \widehat{q}_s$ are the top-s eigenvalues and corresponding eigenvectors of $\widehat{\Pi}_1 + \cdots + \widehat{\Pi}_E$. Further, for all $e \in \{1, \ldots, E\}$, let $\widehat{\lambda}_{e,1} \geq \cdots \geq \widehat{\lambda}_{e,p}$ and $\widehat{q}_{e,1}, \ldots, \widehat{q}_{e,p}$ denote the eigenvalues and corresponding eigenvectors of $\widehat{\Sigma}_e$, in decreasing order. Then, for all $e \in \{1, \ldots, E\}$, let

$$\widehat{Y}_e^\star := \sum_{j=1}^{k} \sum_{\ell=k+1}^{p} \frac{n\, \widehat{\lambda}_{e,j} \widehat{\lambda}_{e,\ell}}{n_e (\widehat{\lambda}_{e,j} - \widehat{\lambda}_{e,\ell})^2} \left(c_j c_j^\top \otimes \widehat{C}_0^\top \widehat{q}_{e,\ell} \widehat{q}_{e,\ell}^\top \widehat{C}_0 \right),$$

where c_j is the jth standard basis vector in $\mathbb{R}^k$, $n := \sum_{e=1}^{E} n_e$, and $\otimes$ denotes the Kronecker product. The eigengap enters here through the factors $(\widehat{\lambda}_{e,j} - \widehat{\lambda}_{e,\ell})^{-2}$: if the gap is small, the estimated eigenspace becomes unstable. Now, for $h, i \in \{1, \ldots, E\}$, define the $k(p-s) \times k(p-s)$ matrix

$$\widehat{V}_{hi} := \sum_{f=1}^{E} (\widehat{C}_h^\top \widehat{\Pi}_\star^+ \widehat{C}_f \otimes I_{p-s}) \widehat{Y}_f^\star (\widehat{C}_f^\top \widehat{\Pi}_\star^+ \widehat{C}_i \otimes I_{p-s}) - (\widehat{C}_h^\top \widehat{\Pi}_\star^+ \widehat{C}_i \otimes I_{p-s}) \widehat{Y}_i^\star - \widehat{Y}_h^\star (\widehat{C}_h^\top \widehat{\Pi}_\star^+ \widehat{C}_i \otimes I_{p-s})$$

and let the $Ek(p-s) \times Ek(p-s)$ partitioned matrix whose (h,i)th block is $\widehat{V}_{hi}$ be denoted by $\widehat{V}$. Finally, let

$$\widehat{W}^\star := \operatorname{diag}(\widehat{Y}_1^\star, \ldots, \widehat{Y}_E^\star) + \widehat{V}.$$

Projector onto the leading eigenspace. Let $\widehat{F}$ be the eigenprojection of $\widehat{W}^\star$ corresponding to its $n_s = (Ek - s)(p - s)$ largest eigenvalues.

Reduced Wald statistic. Schott's reduced Wald statistic is then

$$T_s := n\, (v^\star)^\top (\widehat{F} \widehat{W}^\star \widehat{F})^+ v^\star,$$

where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse. Under the Gaussian assumptions and the eigengap condition, T_s is asymptotically $\chi^2_{n_s}$ under $H_{0,s}$.

B.3.2 Adaptation to $\mathcal{S}_\star$

For `FindS`$_\star$, we apply the same construction not to the top-k projectors Π_e, but to their orthogonal complements

$$\Pi_e^\perp := I_p - \Pi_e.$$

A vector belongs to $\mathcal{S}_\star$ if and only if it lies in the null space of every $\Pi_e^\perp$. Hence, $\mathcal{S}_\star$ is the common null space of $\Pi_1^\perp, \ldots, \Pi_E^\perp$, and the span of these complementary projectors has dimension $r := p - \dim(\mathcal{S}_\star)$; estimating $\dim(\mathcal{S}_\star)$ is therefore equivalent to estimating r.

In practice, this means that Schott's statistic is applied exactly as above, but with the selected eigenspaces replaced by the $(p-k)$-dimensional complementary eigenspaces, that is, by $\Pi_e^\perp$, and for $t = p - k, \ldots, p - m_{\min} - 1$, test

$$H_{0,t} : r = t \quad \text{versus} \quad H_{A,t} : r > t, \qquad r = p - \dim(\mathcal{S}_\star).$$

Let $\widehat{r}$ be the first t for which we do not reject. The estimated invariant dimension is then $\widehat{m} = p - \widehat{r}$. This is the same procedure as in the main text under the change of variables $s = p - t$. For the complementary-space test, $\widehat{C}_e$ is formed from the empirical eigenvectors $\widehat{q}_{e,k+1}, \ldots, \widehat{q}_{e,p}$. Accordingly, in the covariance formula the selected indices range over $j = k+1, \ldots, p$ and the non-selected indices over $\ell = 1, \ldots, k$; equivalently, one may apply Schott's formula after reordering the spectrum so that the complementary eigenspace is treated as the selected subspace.

Under the complement parametrization, the degrees of freedom become $n_t = (E(p-k) - t)(p - t)$, so T_t is asymptotically $\chi^2_{n_t}$ under $H_{0,t}$. The eigengap assumption $\lambda_k(\Sigma_e) > \lambda_{k+1}(\Sigma_e)$ separates the top-k and bottom-$(p-k)$ eigenspaces symmetrically, so Schott's regularity conditions apply verbatim to the complementary eigenspaces.

C Proofs

C.1 Additional intermediate results

Here we collect additional results used in the proofs of the main text results. We first introduce principal angles and the corresponding projector-distance identity, then restate the Davis–Kahan theorem, and finally record the three lemmas used in the consistency proof § C.4.

Definition C.1 (Principal angles)**.** Let $U, V \in \mathcal{O}_{p\times k}$ and let $\sigma_1 \geq \cdots \geq \sigma_k$ be the singular values of $U^\top V$. The *principal angles* between U and V are

$$\cos^{-1}(\sigma_1), \ldots, \cos^{-1}(\sigma_d).$$

We write $\Theta(U, V)$ for the diagonal matrix of principal angles and $\sin\Theta(U, V)$ for its entrywise sine.

Proposition C.2 (Projector distance and principal angles)**.** *Let $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^p$ be subspaces of equal dimension k, and let $U, V \in \mathbb{R}^{p\times k}$ have orthonormal columns spanning $\mathcal{U}$ and $\mathcal{V}$, respectively. Then*

$$\|\Pi_{\mathcal{U}} - \Pi_{\mathcal{V}}\|_{\mathrm{op}} = \|\sin\Theta(U, V)\|_{\mathrm{op}}. \tag{8}$$

Proof. This is the standard projector-distance identity for equal-dimensional subspaces; see, for example, [10, Thm. 2.1]. Since U and V have orthonormal columns, their orthogonal projectors are $\Pi_{\mathcal{U}} = UU^\top$ and $\Pi_{\mathcal{V}} = VV^\top$. □

Theorem C.3 (Davis–Kahan $\sin\Theta$ theorem [53])**.** *Let $\Sigma, \widehat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ and let $r, s \in \{1, \ldots, p\}$ such that $r \leq s$. Let V and $\widehat{V}$ contain orthonormal eigenvectors associated with the eigenvalues $\lambda_r, \ldots, \lambda_s$ and $\widehat{\lambda}_r, \ldots, \widehat{\lambda}_s$, respectively. Let*

$$\delta := \inf\{|\widehat{\lambda} - \lambda| : \lambda \in [\lambda_s, \lambda_r],\ \widehat{\lambda} \in (-\infty, \widehat{\lambda}_{s+1}] \cup [\widehat{\lambda}_{r-1}, \infty)\},$$

with $\widehat{\lambda}_0 := \infty$ and $\widehat{\lambda}_{p+1} := -\infty$. If $\delta > 0$, then

$$\left\|\sin\Theta(\widehat{V}, V)\right\|_{\mathrm{op}} \leq \frac{\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}}}{\delta}. \tag{9}$$

Lemma C.4 (Top-d eigenspace perturbation bound)**.** *Let $A, \widehat{A}$ be symmetric $p \times p$ matrices, and let $d \in \{1, \ldots, p-1\}$. Let $\Pi_d(A)$ and $\Pi_d(\widehat{A})$ denote the orthogonal projector onto the spans of the top d eigenvectors of A and $\widehat{A}$, respectively. If $\gamma_d(A) := \lambda_d(A) - \lambda_{d+1}(A) > 0$ and $\left\|\widehat{A} - A\right\|_{\mathrm{op}} < \frac{\gamma_d(A)}{2}$, then*

$$\left\|\Pi_d(\widehat{A}) - \Pi_d(A)\right\|_{\mathrm{op}} \leq \frac{2\left\|\widehat{A} - A\right\|_{\mathrm{op}}}{\gamma_d(A)}.$$

Proof. Let $\varepsilon := \left\|\widehat{A} - A\right\|_{\mathrm{op}}$. By Weyl's inequality [21, Theorem 4.3.7] in decreasing-order notation with indices $(d+1, 1)$ applied to $(A, \widehat{A} - A)$, $\lambda_{d+1}(\widehat{A}) \leq \lambda_{d+1}(A) + \varepsilon$, so

$$\lambda_d(A) - \lambda_{d+1}(\widehat{A}) \geq \lambda_d(A) - \lambda_{d+1}(A) - \varepsilon = \gamma_d(A) - \varepsilon > 0,$$

where the last inequality uses $\varepsilon < \gamma_d(A)/2$ which holds by assumption. Let V and $\widehat{V}$ have orthonormal columns spanning the top-d eigenspaces of A and $\widehat{A}$, respectively. Applying Thm. C.3 with $\Sigma = A$, $\widehat{\Sigma} = \widehat{A}$, $r = 1$, and $s = d$. The separation quantity is $\delta = \inf\{|\widehat{\lambda} - \lambda| : \lambda \in [\lambda_d(A), \lambda_1(A)], \widehat{\lambda} \in (-\infty, \lambda_{d+1}(\widehat{A})]\} = \lambda_d(A) - \lambda_{d+1}(\widehat{A}) \geq \gamma_d(A) - \varepsilon > \gamma_d(A)/2$. Therefore,

$$\left\|\Pi_d(\widehat{A}) - \Pi_d(A)\right\|_{\mathrm{op}} = \left\|\sin\Theta(\widehat{V}, V)\right\|_{\mathrm{op}} \leq \frac{\varepsilon}{\delta} \leq \frac{2\varepsilon}{\gamma_d(A)},$$

where the first equality follows from (8). □

Recall the following asymptotic notation. Let $(b_N)_{N\geq 1}$ be a deterministic positive sequence and let $(R_N)_{N\geq 1}$ be a sequence of random scalars or matrices. Fix a norm $\|\cdot\|$ on the relevant space; for scalars this is the absolute value. We write $R_N = O(b_N)$ almost surely if there exists an event of probability one on which, for each realization, there exist a finite constant $C < \infty$ and index $N_0 \in \mathbb{N}_+$ (both allowed to depend on the realization) such that $\|R_N\| \leq C b_N$ for all $N \geq N_0$. We write $R_N = o(b_N)$ almost surely if $\|R_N\|/b_N \to 0$ almost surely. For a matrix-valued R_N, an entrywise $O(b_N)$ (resp. $o(b_N)$) statement means that every entry $(R_N)_{ij}$ is itself $O(b_N)$ (resp. $o(b_N)$) almost surely, with the constant and index threshold allowed to depend on (i, j).

Lemma C.5 (Polynomial rate for the averaged projectors)**.** *Assume that, for all $e \in \mathcal{E}$, the observations $x_{e,1}, \ldots, x_{e,n_e}$ are i.i.d. copies of x_e with $\mathbb{E}[x_e] = 0$, $\mathbb{E}\|x_e\|_2^4 < \infty$, and $\lambda_k(\Sigma_e) > \lambda_{k+1}(\Sigma_e)$. Then, for all $\alpha \in (0, 1/2)$ and all $e \in \mathcal{E}$,*

$$\left\|\widehat{\Sigma}_e - \Sigma_e\right\|_{\mathrm{op}} = o(n_e^{-\alpha}) \qquad \text{and} \qquad \left\|\widehat{\Pi}_k^{(e)} - \Pi_k^{(e)}\right\|_{\mathrm{op}} = o(n_e^{-\alpha})$$

almost surely. Consequently, with $n := \min_{e\in\mathcal{E}} n_e$,

$$\left\|\widehat{\Sigma} - \overline{\Sigma}\right\|_{\mathrm{op}} = o(n^{-\alpha}) \qquad \text{and} \qquad \left\|\widehat{\Pi} - \overline{\Pi}\right\|_{\mathrm{op}} = o(n^{-\alpha})$$

almost surely.

Proof. **Step 1: convergence of coordinates in each domain.** Let $e \in \mathcal{E}$ and $\alpha \in (0, 1/2)$. In this proof only, let $N := n_e$ and $X_i := x_{e,i}$. Choose $s \in (1, 2)$ such that $\alpha < 1 - 1/s$. For all $a, b \in \{1, \ldots, p\}$ and all $i \in \{1, \ldots, N\}$, let $Y_i^{ab} := (X_i)_a (X_i)_b$ (where $(X_i)_a$ denotes the a-th coordinate of X_i). Since $s < 2$ and $\mathbb{E}\|x_e\|_2^4 < \infty$, we have

$$\mathbb{E}|Y_i^{ab}|^s = \mathbb{E}[|(X_i)_a|^s |(X_i)_b|^s] \leq \mathbb{E}\|X_i\|_2^{2s} < \infty$$

where the final bound uses $2s < 4$. By the Marcinkiewicz–Zygmund strong law of large numbers [4, Theorem 8.4.4, statement (ii)],

$$\frac{1}{N^{1/s}} \sum_{i=1}^{N} \left(Y_i^{ab} - \mathbb{E}Y_i^{ab}\right)$$

converges almost surely, and is therefore bounded almost surely. Dividing by $N^{1-1/s}$ gives

$$\frac{1}{N} \sum_{i=1}^{N} Y_i^{ab} - \mathbb{E}Y_i^{ab} = O\left(N^{-(1-1/s)}\right) = o(N^{-\alpha}) \qquad \text{almost surely,}$$

where the last equality uses $\alpha < 1 - 1/s$. Applying the same strong law to the coordinate variables $(X_i)_a$ and using $\mathbb{E}(X_i)_a = 0$, gives $\overline{X}_a := \frac{1}{N}\sum_{i=1}^{N}(X_i)_a = O(N^{-(1-1/s)})$ almost surely, and hence $\overline{X}_a \overline{X}_b = O(N^{-2(1-1/s)}) = o(N^{-\alpha})$ almost surely, since $2(1 - 1/s) > \alpha$.

Step 2: convergence of covariances in each domain. Let $A_N := N^{-1}\sum_{i=1}^{N} X_i X_i^\top$. Since $\widehat{\Sigma}_e = \frac{N}{N-1}\left(A_N - \overline{X}\,\overline{X}^\top\right)$, we have

$$\widehat{\Sigma}_e - \Sigma_e = (A_N - \Sigma_e) - \overline{X}\,\overline{X}^\top + \frac{1}{N-1}\left(A_N - \overline{X}\,\overline{X}^\top\right).$$

The first two terms are entrywise $o(N^{-\alpha})$ almost surely by the bounds of Step 1: for $A_N - \Sigma_e$ this uses $\mathbb{E}[X_i X_i^\top] = \Sigma_e$, which holds since $\mathbb{E}[X_i] = 0$, so that the entrywise bound on A_N from Step 1 is a bound on $A_N - \Sigma_e$. The last term is entrywise $O(N^{-1}) = o(N^{-\alpha})$ almost surely, because $A_N - \overline{X}\,\overline{X}^\top = O(1)$ entrywise almost surely and $\alpha < 1$. Since p is fixed, entrywise convergence at rate $o(N^{-\alpha})$ implies operator-norm convergence at the same rate. Therefore

$$\left\|\widehat{\Sigma}_e - \Sigma_e\right\|_{\mathrm{op}} = o(n_e^{-\alpha}) \qquad \text{almost surely.} \tag{10}$$

Step 3: convergence of projection matrices in each domain. Let $\gamma_k(\Sigma_e) := \lambda_k(\Sigma_e) - \lambda_{k+1}(\Sigma_e) > 0$, by assumption. By (10), for all sufficiently large n_e, $\left\|\widehat{\Sigma}_e - \Sigma_e\right\|_{\mathrm{op}} < \gamma_k(\Sigma_e)/2$ almost surely. Applying Lemma C.4 with $A = \Sigma_e$, $\widehat{A} = \widehat{\Sigma}_e$, and $d = k$ gives

$$\left\|\widehat{\Pi}_k^{(e)} - \Pi_k^{(e)}\right\|_{\mathrm{op}} \leq \frac{2}{\gamma_k(\Sigma_e)} \left\|\widehat{\Sigma}_e - \Sigma_e\right\|_{\mathrm{op}} = o(n_e^{-\alpha}) \qquad \text{almost surely,}$$

where the last equality follows from (10).

Step 4: convergence of the average covariance and average projection matrix. Since E is fixed and $n_e \geq n$ for all $e \in \mathcal{E}$, averaging over $e \in \mathcal{E}$ yields

$$\left\|\widehat{\Sigma} - \overline{\Sigma}\right\|_{\mathrm{op}} \leq \frac{1}{E}\sum_{e=1}^{E}\left\|\widehat{\Sigma}_e - \Sigma_e\right\|_{\mathrm{op}} = o(n^{-\alpha})$$

and

$$\left\|\widehat{\Pi} - \overline{\Pi}\right\|_{\mathrm{op}} \leq \frac{1}{E}\sum_{e=1}^{E}\left\|\widehat{\Pi}_k^{(e)} - \Pi_k^{(e)}\right\|_{\mathrm{op}} = o(n^{-\alpha})$$

almost surely. □

Lemma C.6 (Recovery of the eigenvalue blocks of $\overline{\Pi}$)**.** *Let $1 \geq \rho_1 > \cdots > \rho_L \geq 0$ be the distinct eigenvalues of $\overline{\Pi}$, and let Π_ℓ denote the orthogonal projector onto the eigenspace of ρ_ℓ, with multiplicity $d_\ell := \mathrm{rank}(\Pi_\ell)$. Listing the eigenvalues of $\overline{\Pi}$ in nonincreasing order with multiplicity, ρ_ℓ occupies a block of d_ℓ consecutive indices which we denote I_ℓ, so that $I_1, \ldots, I_L$ partition $\{1, \ldots, p\}$. Let $\gamma_{\overline{\Pi}}$ be the smallest gap between the distinct population eigenvalues,*

$$\gamma_{\overline{\Pi}} := \begin{cases} \min\{\rho - \rho' : \rho, \rho' \in \{\rho_1, \ldots, \rho_L\},\ \rho > \rho'\}, & L \geq 2, \\ +\infty, & L = 1. \end{cases}$$

Let $\widehat{\rho}_1 \geq \cdots \geq \widehat{\rho}_p$ be the eigenvalues of $\widehat{\Pi}$ (including duplicates). Suppose that $\|\widehat{\Pi} - \overline{\Pi}\|_{\mathrm{op}} < \mathrm{tol}_n/4$ and $\mathrm{tol}_n < \gamma_{\overline{\Pi}}/2$. Then the block-stabilized construction of `AnchorPCA`$_\infty$ *in § 5.1 partitions the indices $1, \ldots, p$ into empirical blocks that coincide with $I_1, \ldots, I_L$.*

Proof. Let $\ell \in \{1, \ldots, L-1\}$ and $c_\ell := \sum_{j<\ell} d_j$, so that $I_\ell = \{c_\ell + 1, \ldots, c_\ell + d_\ell\}$. Let $\varepsilon_n := \|\widehat{\Pi} - \overline{\Pi}\|_{\mathrm{op}}$. Recall that the construction in § 5.1 scans the ordered eigenvalues of $\widehat{\Pi}$ from largest to smallest to group them into empirical blocks: each block starts at an eigenvalue and includes subsequent eigenvalues as long as they are at most tol_n smaller than the first eigenvalue of the block.

Step 1: no block I_ℓ is split. By Weyl's inequality [21, Theorem 4.3.1] in decreasing-order notation applied to the two symmetric matrices $\overline{\Pi}$ and $\widehat{\Pi} - \overline{\Pi}$, we have that the eigenvalues of $\widehat{\Pi}$ and $\overline{\Pi}$, matched by rank order and counted with multiplicity, satisfy $|\lambda_j(\widehat{\Pi}) - \lambda_j(\overline{\Pi})| \leq \|\widehat{\Pi} - \overline{\Pi}\|_{\mathrm{op}} = \varepsilon_n$ for all $j \in \{1, \ldots, p\}$. Since $\lambda_j(\overline{\Pi}) = \rho_\ell$ for all $j \in I_\ell$, it follows that $|\widehat{\rho}_j - \rho_\ell| \leq \varepsilon_n$ for all $j \in I_\ell$. Hence, for all $j, j' \in I_\ell$,

$$|\widehat{\rho}_j - \widehat{\rho}_{j'}| \leq |\widehat{\rho}_j - \rho_\ell| + |\widehat{\rho}_{j'} - \rho_\ell| \leq 2\varepsilon_n < \mathrm{tol}_n/2 < \mathrm{tol}_n,$$

where the third inequality follows by the assumption $\varepsilon_n = \|\widehat{\Pi} - \overline{\Pi}\|_{\mathrm{op}} < \mathrm{tol}_n/4$. In particular, every eigenvalue in I_ℓ lies within tol_n of the first eigenvalue of the block, $\widehat{\rho}_{c_\ell+1}$, so the rule keeps all indices in I_ℓ in the same empirical block.

Step 2: adjacent blocks I_ℓ and $I_{\ell+1}$ are not merged. Applying Weyl's inequality [21, Theorem 4.3.1] in decreasing-order notation to $\overline{\Pi}$ and $\widehat{\Pi} - \overline{\Pi}$, for any index j we have $\lambda_j(\overline{\Pi}) + \lambda_p(\widehat{\Pi} - \overline{\Pi}) \leq \lambda_j(\widehat{\Pi}) \leq \lambda_j(\overline{\Pi}) + \lambda_1(\widehat{\Pi} - \overline{\Pi})$, and hence $\lambda_j(\overline{\Pi}) - \varepsilon_n \leq \lambda_j(\widehat{\Pi}) \leq \lambda_j(\overline{\Pi}) + \varepsilon_n$. Taking $j = c_\ell + 1$ and using $\lambda_{c_\ell+1}(\overline{\Pi}) = \rho_\ell$ gives $\widehat{\rho}_{c_\ell+1} = \lambda_{c_\ell+1}(\widehat{\Pi}) \geq \rho_\ell - \varepsilon_n$. Similarly, taking $j = c_\ell + d_\ell + 1$ and using $\lambda_{c_\ell+d_\ell+1}(\overline{\Pi}) = \rho_{\ell+1}$ gives $\widehat{\rho}_{c_\ell+d_\ell+1} = \lambda_{c_\ell+d_\ell+1}(\widehat{\Pi}) \leq \rho_{\ell+1} + \varepsilon_n$. Therefore,

$$\widehat{\rho}_{c_\ell+1} - \widehat{\rho}_{c_\ell+d_\ell+1} \geq \rho_\ell - \rho_{\ell+1} - 2\varepsilon_n \geq \gamma_{\overline{\Pi}} - 2\varepsilon_n > \gamma_{\overline{\Pi}} - \frac{\mathrm{tol}_n}{2} > \mathrm{tol}_n,$$

where the second last inequality uses that $\varepsilon_n < \mathrm{tol}_n/4$, and the final inequality uses $\mathrm{tol}_n < \gamma_{\overline{\Pi}}/2$ (both hold by assumption). Thus the first eigenvalue past I_ℓ lies more than tol_n below $\widehat{\rho}_{c_\ell+1}$, the first eigenvalue of the current block, so the rule closes that block at index $c_\ell + d_\ell$ and starts a new one.

Going through the eigenvalues from largest to smallest, Steps 1 and 2 show that each block starts at index $c_\ell + 1$ and ends at index $c_\ell + d_\ell$: Step 1 keeps a block from ending too early, and Step 2 forces it to end at the right place. The blocks the construction produces are therefore exactly $I_1, \ldots, I_L$. □

C.2 Proof of Thm. 3.1

Thm. 3.1 is a projector-level statement. We prove Thm. C.8, which is a stronger version that uses the ordered convention of § B.1 and additionally records what happens to the ordered directions. Part (i) concerns $\texttt{AnchorPCA}_\infty$, and Part (ii) concerns $\texttt{AnchorPCA}_\lambda$ as $\lambda \to \infty$; both are stated first at the projector level and then at the direction level under a simple-spectrum Assm. C.7. Together, the two parts imply Thm. 3.1. If $m = 0$, the statements below are vacuous, so we assume $m \geq 1$.

Assumption C.7 (Simple spectrum on the invariant block)**.** *The restriction of $\overline{\Sigma}_{\mathcal{S}_\star}$ to $\mathcal{S}_\star$ (i.e., $\overline{\Sigma}_{\mathcal{S}_\star} := \Pi_{\mathcal{S}_\star}\overline{\Sigma}\Pi_{\mathcal{S}_\star}$) has simple spectrum within $\mathcal{S}_\star$: there exist $\mu_1 > \cdots > \mu_m$ and an orthonormal basis $s_1, \ldots, s_m \in \mathcal{S}_\star$ such that $\overline{\Sigma}_{\mathcal{S}_\star} s_i = \mu_i s_i$ for all $i \in \{1, \ldots, m\}$.*

Theorem C.8 (Thm. 3.1, extended version)**.** *Assume $1 \leq m := \dim(\mathcal{S}_\star) \leq k < p$, and let $\overline{\Sigma}_{\mathcal{S}_\star} = \Pi_{\mathcal{S}_\star}\overline{\Sigma}\Pi_{\mathcal{S}_\star}$. Let $\Pi_k^{\perp,\infty}$ and $\Pi_k^{\perp,\lambda}$ be any rank-k solution projectors of $\texttt{AnchorPCA}_\infty$ and $\texttt{AnchorPCA}_\lambda$, respectively, and let*

$$W_k^{\perp,\infty} = (u_1, \ldots, u_k), \qquad W_k^{\perp,\lambda} = (w_1^\lambda, \ldots, w_k^\lambda)$$

be corresponding ordered bases (§ B.1).

(i) ($\texttt{AnchorPCA}_\infty$.) *The first m vectors $u_1, \ldots, u_m$ form an ordered orthonormal basis of $\mathcal{S}_\star$ that diagonalizes $\overline{\Sigma}_{\mathcal{S}_\star}$ within $\mathcal{S}_\star$; in particular, $\operatorname{span}(u_1, \ldots, u_m) = \mathcal{S}_\star$ and $\Pi_k^{\perp,\infty}\Pi_{\mathcal{S}_\star} = \Pi_{\mathcal{S}_\star}$. If additionally Assm. C.7 holds, then $u_i \in \{+s_i, -s_i\}$ for all $i \in \{1, \ldots, m\}$.*

(ii) ($\texttt{AnchorPCA}_\lambda$, $\lambda \to \infty$.) *Let $\Pi_m^{\perp,\lambda}$ denote the projector onto $\operatorname{span}(w_1^\lambda, \ldots, w_m^\lambda)$, that is, the leading rank-m spectral projector of M_λ. Then,*

$$\left\|\Pi_m^{\perp,\lambda} - \Pi_{\mathcal{S}_\star}\right\|_{\mathrm{op}} \to 0 \qquad \text{as } \lambda \to \infty.$$

If additionally Assm. C.7 holds, then, with a suitable choice of signs, as $\lambda \to \infty$,

$$\lambda_i(M_\lambda) - 2E\lambda \to \mu_i, \qquad w_i^\lambda \to s_i \quad \text{for all } i \in \{1, \ldots, m\}.$$

Proof. **Part (i)** $\texttt{AnchorPCA}_\infty$**.** Since $\mathcal{S}_\star$ is the eigenspace of $\overline{\Pi}$ corresponding to its largest eigenvalue 1, and $\dim(\mathcal{S}_\star) = m \leq k$, the ordering convention of § B.1 draws the first m directions $u_1, \ldots, u_m$ entirely from $\mathcal{S}_\star$. Hence

$$\operatorname{span}(u_1, \ldots, u_m) = \mathcal{S}_\star, \qquad \text{so} \qquad \Pi_k^{\perp,\infty}\Pi_{\mathcal{S}_\star} = \Pi_{\mathcal{S}_\star}.$$

Within this block, the convention orders directions by diagonalizing $\overline{\Sigma}$, equivalently $\overline{\Sigma}_{\mathcal{S}_\star}$, within $\mathcal{S}_\star$. Thus $u_1, \ldots, u_m$ is an orthonormal eigenbasis of $\overline{\Sigma}_{\mathcal{S}_\star}$ on $\mathcal{S}_\star$, ordered by decreasing eigenvalue. Under Assm. C.7, this eigenbasis is unique up to sign, so $u_i \in \{+s_i, -s_i\}$ for each $i \in \{1, \ldots, m\}$.

Part (ii) $\texttt{AnchorPCA}_\lambda$**.** We first show the claimed subspace convergence in steps 1–2 invoking the Davis–Kahan theorem (Thm. C.3). The direction-level convergence under Assm. C.7 is proved in steps 3–6.

Step 1: Uniqueness of the top-m eigenspace of $(2E\lambda)^{-1}M_\lambda$. Let $S \in \mathbb{R}^{p\times m}$ have orthonormal columns spanning $\mathcal{S}_\star$, and complete it to an orthogonal matrix (S, Z) with $Z \in \mathbb{R}^{p\times(p-m)}$ spanning $\mathcal{S}_\star^\perp$. Since $\mathcal{S}_\star$ is the eigenvalue-1 eigenspace of $\overline{\Pi}$, we have

$$(S, Z)^\top \overline{\Pi}(S, Z) = \begin{pmatrix} I_m & 0 \\ 0 & R \end{pmatrix}, \qquad R := Z^\top \overline{\Pi} Z.$$

Let $\widetilde{M}_\lambda := \frac{1}{2E\lambda}M_\lambda = \overline{\Pi} + \frac{1}{2E\lambda}\overline{\Sigma}$. By Weyl's inequality [21, Theorem 4.3.1] in decreasing-order notation applied to $(\overline{\Pi}, (2E\lambda)^{-1}\overline{\Sigma})$, we have, for each index j, $\lambda_j(\overline{\Pi}) - \|(2E\lambda)^{-1}\overline{\Sigma}\|_{\mathrm{op}} \leq \lambda_j(\widetilde{M}_\lambda) \leq \lambda_j(\overline{\Pi}) + \|(2E\lambda)^{-1}\overline{\Sigma}\|_{\mathrm{op}}$. Since $\|(2E\lambda)^{-1}\overline{\Sigma}\|_{\mathrm{op}} = (2E\lambda)^{-1}\|\overline{\Sigma}\|_{\mathrm{op}}$, applying this with $j = m$ and using $\lambda_m(\overline{\Pi}) = 1$ gives $\lambda_m(\widetilde{M}_\lambda) \geq 1 - (2E\lambda)^{-1}\|\overline{\Sigma}\|_{\mathrm{op}}$. Applying the same bound with $j = m + 1$ and using $\lambda_{m+1}(\overline{\Pi}) = \lambda_1(R)$ gives $\lambda_{m+1}(\widetilde{M}_\lambda) \leq \lambda_1(R) + (2E\lambda)^{-1}\|\overline{\Sigma}\|_{\mathrm{op}}$. Hence, for all $\lambda > \lambda_0 := \frac{\|\overline{\Sigma}\|_{\mathrm{op}}}{E(1-\lambda_1(R))}$, we have $\lambda_m(\widetilde{M}_\lambda) > \lambda_{m+1}(\widetilde{M}_\lambda)$. Thus, the leading rank-m spectral projector of M_λ is unique for all $\lambda > \lambda_0$.

Step 2: Apply Davis–Kahan to obtain subspace convergence. We have that

$$1 - \lambda_{m+1}(\widetilde{M}_\lambda) \geq 1 - \lambda_1(R) - \frac{1}{2E\lambda}\left\|\overline{\Sigma}\right\|_{\mathrm{op}} > 0, \tag{11}$$

where the first inequality follows from step 1 and the second holds for all $\lambda > \lambda_0$ as $\lambda_1(R) < 1$ (shown in step 2 in the proof of Prop. B.3). We apply Thm. C.3 to $\Sigma := \overline{\Pi}$ and $\widehat{\Sigma} := \widetilde{M}_\lambda$ with $r = 1$ and $s = m$. Hence, $V = S$, $\widehat{V} = (w_1^\lambda, \ldots, w_m^\lambda) =: W_m^\lambda$. With these choices, the separation parameter $\delta = \inf\{|\widehat{\lambda} - \lambda| : \lambda \in [\lambda_m(\overline{\Pi}) = 1, \lambda_1(\overline{\Pi}) = 1],\ \widehat{\lambda} \in (-\infty, \lambda_{m+1}(\widetilde{M}_\lambda)]\} = |1 - \lambda_{m+1}(\widetilde{M}_\lambda)|$, so the theorem states that

$$\left\|\sin\Theta(W_m^\lambda, S)\right\|_{\mathrm{op}} \leq \frac{\left\|\overline{\Pi} - \widetilde{M}_\lambda\right\|_{\mathrm{op}}}{|1 - \lambda_{m+1}(\widetilde{M}_\lambda)|}.$$

We apply (11) for all $\lambda > \lambda_0$ and then consider $\lambda \to \infty$, which gives

$$\left\|\sin\Theta(W_m^\lambda, S)\right\|_{\mathrm{op}} \leq \frac{(2E\lambda)^{-1}\left\|\overline{\Sigma}\right\|_{\mathrm{op}}}{1 - \lambda_1(R) - (2E\lambda)^{-1}\left\|\overline{\Sigma}\right\|_{\mathrm{op}}} \to 0.$$

Since $\Pi_m^{\perp,\lambda} = W_m^\lambda (W_m^\lambda)^\top$ and $\Pi_{\mathcal{S}_\star} = SS^\top$, by (8), as $\lambda \to \infty$,

$$\left\|\Pi_m^{\perp,\lambda} - \Pi_{\mathcal{S}_\star}\right\|_{\mathrm{op}} \to 0.$$

Step 3: Write the eigenvalue equation in $\mathcal{S}_\star \oplus \mathcal{S}_\star^\perp$ coordinates. Assume now, in addition, that Assm. C.7 holds. For the remainder of the proof, set $S := (s_1, \ldots, s_m)$ and complete it to an orthogonal matrix (S, Z). In the basis (S, Z),

$$(S, Z)^\top M_\lambda (S, Z) = \begin{pmatrix} 2E\lambda I_m + A & C \\ C^\top & 2E\lambda R + B \end{pmatrix},$$

where

$$A := S^\top \overline{\Sigma}_{\mathcal{S}_\star} S = S^\top \overline{\Sigma} S = \mathrm{diag}(\mu_1, \ldots, \mu_m), \qquad B := Z^\top \overline{\Sigma} Z, \qquad C := S^\top \overline{\Sigma} Z.$$

The equality $S^\top \overline{\Sigma}_{\mathcal{S}_\star} S = S^\top \overline{\Sigma} S$ follows from $\Pi_{\mathcal{S}_\star} S = S$.

Let $i \in \{1, \ldots, m\}$. Recall that $M_\lambda w_i^\lambda = \lambda_i(M_\lambda) w_i^\lambda$, i.e. w_i^λ is the ith leading eigenvector of M_λ. Since the columns of (S, Z) form an orthonormal basis of $\mathbb{R}^p$, the vector w_i^λ has unique coordinates $u_i^\lambda \in \mathbb{R}^m$ and $v_i^\lambda \in \mathbb{R}^{p-m}$ in this basis, so that

$$w_i^\lambda = S u_i^\lambda + Z v_i^\lambda.$$

We also let $\eta_i(\lambda) := \lambda_i(M_\lambda) - 2E\lambda$. The eigenvalue equation $M_\lambda w_i^\lambda = \lambda_i(M_\lambda) w_i^\lambda$ then is equivalent to

$$(A - \eta_i(\lambda) I_m) u_i^\lambda + C v_i^\lambda = 0, \tag{12}$$

$$C^\top u_i^\lambda - \underbrace{\left(2E\lambda(I_{p-m} - R) + \eta_i(\lambda) I_{p-m} - B\right)}_{=: L_i(\lambda)} v_i^\lambda = 0. \tag{13}$$

Step 4: Reduce to an eigenvalue equation in $\mathcal{S}_\star$. Rearranging (13) gives

$$L_i(\lambda) v_i^\lambda = C^\top u_i^\lambda.$$

Because $R \preceq \lambda_1(R) I_{p-m}$ with $\lambda_1(R) < 1$, we have $I_{p-m} - R \succeq (1 - \lambda_1(R)) I_{p-m}$. By Weyl's inequality [21, Theorem 4.3.1], used in decreasing-order notation and applied to $(2E\lambda\overline{\Pi},\ \overline{\Sigma})$, for each $i \leq m$ we have $\lambda_i(2E\lambda\overline{\Pi}) + \lambda_p(\overline{\Sigma}) \leq \lambda_i(M_\lambda) \leq \lambda_i(2E\lambda\overline{\Pi}) + \lambda_1(\overline{\Sigma})$. Since $S_\star$ is the eigenvalue-1 eigenspace of $\overline{\Pi}$, $\lambda_i(2E\lambda\overline{\Pi}) = 2E\lambda$ for $i \leq m$. Hence $\lambda_p(\overline{\Sigma}) \leq \lambda_i(M_\lambda) - 2E\lambda \leq \lambda_1(\overline{\Sigma})$, and therefore

$$|\eta_i(\lambda)| \leq \|\overline{\Sigma}\|_{\mathrm{op}}. \tag{14}$$

Combining this with $\|B\|_{\mathrm{op}} \le \left\|\overline{\Sigma}\right\|_{\mathrm{op}}$, we have

$$L_i(\lambda) = 2E\lambda(I_{p-m} - R) + \eta_i(\lambda) I_{p-m} - B \succeq \left(2E\lambda(1-\lambda_1(R)) - 2\left\|\overline{\Sigma}\right\|_{\mathrm{op}}\right) I_{p-m}.$$

Thus, for all $\lambda > \lambda_0$, the matrix $L_i(\lambda)$ is positive definite and therefore invertible. In addition, for all $\lambda > \lambda_0$,

$$\left\|L_i(\lambda)^{-1}\right\|_{\mathrm{op}} \le \frac{1}{2E\lambda(1-\lambda_1(R)) - 2\left\|\overline{\Sigma}\right\|_{\mathrm{op}}}. \tag{15}$$

Since $L_i(\lambda) v_i^\lambda = C^\top u_i^\lambda$ and $\|u_i^\lambda\|_2 \le 1$, it follows that

$$\|v_i^\lambda\|_2 \le \left\|L_i(\lambda)^{-1}\right\|_{\mathrm{op}} \|C\|_{\mathrm{op}} \|u_i^\lambda\|_2 \le \left\|L_i(\lambda)^{-1}\right\|_{\mathrm{op}} \|C\|_{\mathrm{op}}.$$

Hence, $\|v_i^\lambda\|_2 \to 0$ as $\lambda \to \infty$ by (15). Because $w_i^\lambda = S u_i^\lambda + Z v_i^\lambda$ has unit norm and (S, Z) is orthogonal, we have $\|u_i^\lambda\|_2^2 + \|v_i^\lambda\|_2^2 = 1$. Therefore

$$\|u_i^\lambda\|_2 \to 1 \tag{16}$$

as $\lambda \to \infty$.

Substituting $v_i^\lambda = L_i(\lambda)^{-1} C^\top u_i^\lambda$ into (12) gives $(A - \eta_i(\lambda) I_m) u_i^\lambda + C L_i(\lambda)^{-1} C^\top u_i^\lambda = 0$, or equivalently

$$(A + C L_i(\lambda)^{-1} C^\top) u_i^\lambda = \eta_i(\lambda) u_i^\lambda. \tag{17}$$

Step 5: Identify the limiting eigenvalues and directions along subsequences. Fix any sequence $\lambda_n \to \infty$. For all $i \in \{1, \dots, m\}$, the sequence $\eta_i(\lambda_n)$ is bounded (see (14)). By (16), after discarding finitely many terms we have $\|u_i^{\lambda_n}\|_2 > 0$ for all $i \in \{1, \dots, m\}$. For such n, define $\widehat{u}_i^{\lambda_n} := \frac{u_i^{\lambda_n}}{\|u_i^{\lambda_n}\|_2}$, which belongs to the compact unit sphere of $\mathbb{R}^m$. By the Bolzano–Weierstrass theorem, each of these sequences has a convergent subsequence. Passing to a common subsequence (which we continue to index by n) along which all $2m$ sequences converge, it holds that for all $i \in \{1, \dots, m\}$, there exist $\eta_i^\star \in \mathbb{R}$ and a unit vector $\widetilde{u}_i \in \mathbb{R}^m$ such that

$$\eta_i(\lambda_n) \to \eta_i^\star, \qquad \widehat{u}_i^{\lambda_n} \to \widetilde{u}_i.$$

Dividing the reduced eigenvalue equation (17) by $\|u_i^{\lambda_n}\|_2$ gives

$$\left(A + C L_i(\lambda_n)^{-1} C^\top\right) \widehat{u}_i^{\lambda_n} = \eta_i(\lambda_n) \widehat{u}_i^{\lambda_n}.$$

Passing to the limit $\lambda_n \to \infty$ along the chosen subsequence, the term $C L_i(\lambda_n)^{-1} C^\top$ vanishes since $\left\|C L_i(\lambda)^{-1} C^\top\right\|_{\mathrm{op}} \le \|C\|_{\mathrm{op}}^2 \left\|L_i(\lambda)^{-1}\right\|_{\mathrm{op}} \to 0$ by (15), and we obtain

$$A \widetilde{u}_i = \eta_i^\star \widetilde{u}_i.$$

This implies that $\widetilde{u}_i$ is a unit eigenvector of A with eigenvalue $\eta_i^\star$. Since $A = \mathrm{diag}(\mu_1, \dots, \mu_m)$ with $\mu_1 > \cdots > \mu_m$, there exists a unique index $j_i \in \{1, \dots, m\}$ such that

$$\eta_i^\star = \mu_{j_i}, \qquad \widetilde{u}_i = \pm e_{j_i}.$$

It remains to show that $j_i = i$, which we do using orthogonality and eigenvalue ordering. Let $i, \ell \in \{1, \dots, m\}$ with $i \neq \ell$. Orthogonality of $w_i^{\lambda_n}$ and $w_\ell^{\lambda_n}$ implies

$$0 = \langle w_i^{\lambda_n}, w_\ell^{\lambda_n} \rangle = \langle u_i^{\lambda_n}, u_\ell^{\lambda_n} \rangle + \langle v_i^{\lambda_n}, v_\ell^{\lambda_n} \rangle.$$

Since $v_i^{\lambda_n}, v_\ell^{\lambda_n} \to 0$ and $\|u_i^{\lambda_n}\|_2, \|u_\ell^{\lambda_n}\|_2 \to 1$, passing to the limit yields $\langle \widetilde{u}_i, \widetilde{u}_\ell \rangle = 0$. Hence $j_1, \dots, j_m$ is a permutation of $\{1, \dots, m\}$. On the other hand, by the ordering convention,

$$\lambda_1(M_{\lambda_n}) \ge \cdots \ge \lambda_m(M_{\lambda_n}) \quad \Longrightarrow \quad \eta_1(\lambda_n) \ge \cdots \ge \eta_m(\lambda_n).$$

Passing to the limit gives $\mu_{j_1} \ge \cdots \ge \mu_{j_m}$. Because $\mu_1 > \cdots > \mu_m$ are strictly decreasing, the only possible permutation $(j_1, \dots, j_m)$ of $\{1, \dots, m\}$ for which $\mu_{j_1} \ge \cdots \ge \mu_{j_m}$ holds is the identity. Therefore, for all $i \in \{1, \dots, m\}$

$$\eta_i^\star = \mu_i, \qquad \widetilde{u}_i = \pm e_i.$$

Step 6: Full convergence of eigenvalues and eigenvectors. Since the original sequence $\lambda_n \to \infty$ was arbitrary, every convergent subsequence of $\eta_i(\lambda)$ has limit μ_i. A bounded sequence whose convergent subsequences all share the same limit converges to that limit, so for all $i \in \{1, \dots, m\}$

$$\lambda_i(M_\lambda) - 2E\lambda = \eta_i(\lambda) \to \mu_i \qquad \text{as} \qquad \lambda \to \infty.$$

Similarly, every subsequential limit of $\widehat{u}_i^\lambda$ equals $\pm e_i$. We can choose the sign of w_i^λ so that $\langle w_i^\lambda, s_i \rangle \geq 0$ rules out the $-e_i$ limit and $\widehat{u}_i^\lambda \to e_i$. As $\|u_i^\lambda\|_2 \to 1$ by (16), this gives $u_i^\lambda \to e_i$. Since $v_i^\lambda \to 0$,

$$w_i^\lambda = S u_i^\lambda + Z v_i^\lambda \to S e_i = s_i \qquad \text{as} \qquad \lambda \to \infty,$$

which is the asserted direction-level convergence. □

C.3 Proof of Thm. 4.1

We first show (5) and then use it to prove statements (i) and (ii). Let $W \in \mathcal{O}_{p\times k}$ and $\Pi_W := WW^\top$. By definition of $\mathcal{C}_\lambda$, for every perturbed covariance Σ'_e, there exists $\Delta_e \in \mathbb{S}_+^p$ with $0 \preceq \Delta_e \preceq 2E\lambda\,\Pi_k^{(e)}$ such that $\Sigma'_e = \Sigma_e + \Delta_e$. Hence,

$$\max_{(\Sigma'_1,\dots,\Sigma'_E)\in\mathcal{C}_\lambda} \mathcal{R}(W;\Sigma'_1,\dots,\Sigma'_E) = \frac{1}{E}\sum_{e=1}^{E}\left[\operatorname{Tr}\big(\Sigma_e(I_p-\Pi_W)\big) + \max_{\substack{\Delta_e\in\mathbb{S}_+^p,\\ 0\preceq\Delta_e\preceq 2E\lambda\,\Pi_k^{(e)}}} \operatorname{Tr}\big(\Delta_e(I_p-\Pi_W)\big)\right].$$

Applying the trace-monotonicity inequality $\operatorname{Tr}(AC) \leq \operatorname{Tr}(BC)$ (valid for $A \preceq B$ and $C \succeq 0$; see, e.g., Boyd and Vandenberghe [7, Example 2.24]) with $A = \Delta_e$, $B = 2E\lambda\,\Pi_k^{(e)}$, and $C = I_p - \Pi_W$, we obtain

$$\operatorname{Tr}\big(\Delta_e(I_p-\Pi_W)\big) \leq \operatorname{Tr}\big(2E\lambda\,\Pi_k^{(e)}(I_p-\Pi_W)\big),$$

with equality for $\Delta_e^\star = 2E\lambda\,\Pi_k^{(e)}$. Therefore,

$$\begin{aligned}\max_{(\Sigma'_1,\dots,\Sigma'_E)\in\mathcal{C}_\lambda} \mathcal{R}(W;\Sigma'_1,\dots,\Sigma'_E) &= \frac{1}{E}\sum_{e=1}^{E}\operatorname{Tr}\big(\Sigma_e(I_p-\Pi_W)\big) + 2\lambda\sum_{e=1}^{E}\operatorname{Tr}\big(\Pi_k^{(e)}(I_p-\Pi_W)\big)\\ &= \frac{1}{E}\sum_{e=1}^{E}\operatorname{Tr}\big(\Sigma_e(I_p-\Pi_W)\big) + \lambda\sum_{e=1}^{E}\|\Pi_W-\Pi_k^{(e)}\|_{\mathrm{F}}^2,\end{aligned}$$

where the second line follows since Π_W and $\Pi_k^{(e)}$ are rank-k orthogonal projectors, so

$$2\operatorname{Tr}\big(\Pi_k^{(e)}(I_p-\Pi_W)\big) = 2k - 2\operatorname{Tr}(\Pi_W\Pi_k^{(e)}) = \operatorname{Tr}\big((\Pi_W-\Pi_k^{(e)})^2\big) = \|\Pi_W-\Pi_k^{(e)}\|_{\mathrm{F}}^2.$$

This proves (5).

Part (i). A direct computation gives $\frac{1}{E}\sum_{e=1}^{E}\operatorname{Tr}\big(\Sigma_e(I_p-\Pi_W)\big) = \operatorname{Tr}(\overline{\Sigma}) - \operatorname{Tr}(W^\top\overline{\Sigma}W)$ so minimizing the worst-case reconstruction error is equivalent to maximizing

$$\operatorname{Tr}(W^\top\overline{\Sigma}W) - \lambda\sum_{e=1}^{E}\|\Pi_W-\Pi_k^{(e)}\|_{\mathrm{F}}^2,$$

which is Definition 2.3.

Part (ii). Using the fact that $\|\Pi_W-\Pi_k^{(e)}\|_{\mathrm{F}}^2 = 2k - 2\operatorname{Tr}(W^\top\Pi_k^{(e)}W)$ for all $e \in \mathcal{E}$ and summing over e,

$$\sum_{e=1}^{E}\|\Pi_W-\Pi_k^{(e)}\|_{\mathrm{F}}^2 = 2Ek - 2E\operatorname{Tr}(W^\top\overline{\Pi}W).$$

Hence

$$\mathcal{B} = \operatorname*{arg\,min}_{\widetilde{W}\in\mathcal{O}_{p\times k}}\sum_{e=1}^{E}\|\Pi_{\widetilde{W}}-\Pi_k^{(e)}\|_{\mathrm{F}}^2 = \operatorname*{arg\,max}_{\widetilde{W}\in\mathcal{O}_{p\times k}}\operatorname{Tr}(\widetilde{W}^\top\overline{\Pi}\widetilde{W}).$$

Combining this with $\mathcal{R}(W;\Sigma_1,\dots,\Sigma_E) = \operatorname{Tr}(\overline{\Sigma}) - \operatorname{Tr}(W^\top\overline{\Sigma}W)$, minimizing $\mathcal{R}$ over $\mathcal{B}$ amounts to maximizing $\operatorname{Tr}(W^\top\overline{\Sigma}W)$ over the maximizers of $\operatorname{Tr}(W^\top\overline{\Pi}W)$, which is Definition 2.2.

Corollary C.9 (Full-rank covariance perturbations). *Fix $\lambda \geq 0$ and $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_E)$ such that $\varepsilon_e \geq 0$ for all $e \in \{1, \ldots, E\}$. Define*

$$\mathcal{C}_{\lambda,\varepsilon} := \Big\{ (\Sigma'_e)_{e=1}^E : \ 0 \preceq \Sigma'_e \preceq \Sigma_e + 2E\lambda\Pi_k^{(e)} + \varepsilon_e I_p \text{ for all } e \in \mathcal{E} \Big\}.$$

For all $W \in \mathcal{O}_{p\times k}$,

$$\begin{aligned} &\max_{(\Sigma'_1,\ldots,\Sigma'_E)\in\mathcal{C}_{\lambda,\varepsilon}} \mathcal{R}(W; \Sigma'_1, \ldots, \Sigma'_E) \\ &= \mathcal{R}(W; \Sigma_1, \ldots, \Sigma_E) + \lambda \sum_{e=1}^E \left\| \Pi_W - \Pi_k^{(e)} \right\|_{\mathrm{F}}^2 + \frac{p-k}{E} \sum_{e=1}^E \varepsilon_e. \end{aligned} \tag{18}$$

As the additional term in the worst-case error is constant in W, the minimizers of (18) are the rank-k $\texttt{AnchorPCA}_\lambda$ solutions. However, the worst-case error increases with the noise levels ε_e.

Proof of Corollary C.9. Fix $W \in \mathcal{O}_{p\times k}$. Since $I_p - \Pi_W \succeq 0$, trace monotonicity (Boyd and Vandenberghe [7, Example 2.24]) gives for all $(\Sigma'_e)_{e=1}^E \in \mathcal{C}_{\lambda,\epsilon}$,

$$\operatorname{Tr}\big(\Sigma'_e(I_p - \Pi_W)\big) \leq \operatorname{Tr}\big((\Sigma_e + 2E\lambda\Pi_k^{(e)} + \varepsilon_e I_p)(I_p - \Pi_W)\big) \quad \forall e \in \{1, \ldots, E\}.$$

Equality holds for every domain at $\Sigma'_e = \Sigma_e + 2E\lambda\Pi_k^{(e)} + \varepsilon_e I_p$ for all $e \in \{1, \ldots, E\}$, which are positive semidefinite and admissible. Averaging over domains and using $2\operatorname{Tr}(\Pi_k^{(e)}(I_p - \Pi_W)) = \|\Pi_W - \Pi_k^{(e)}\|_F^2$ and $\operatorname{Tr}(I_p - \Pi_W) = p - k$ gives the stated worst-case value. Its added term does not depend on W, so the minimizers are unchanged. □

Remark C.10. The class $\mathcal{C}_{\lambda,\varepsilon}$ in particular covers bounded covariance-estimation error. Suppose each covariance $\Sigma_e + \Delta_e$, with $0 \preceq \Delta_e \preceq 2E\lambda\Pi_k^{(e)}$, is further perturbed by a symmetric error A_e satisfying $\|A_e\|_{\mathrm{op}} \leq \varepsilon_e$. Since $A_e \preceq \varepsilon_e I_p$, we have

$$\Sigma_e + \Delta_e + A_e \preceq \Sigma_e + 2E\lambda\Pi_k^{(e)} + \varepsilon_e I_p.$$

Thus $(\Sigma_e + \Delta_e + A_e)_{e=1}^E$ belongs to $\mathcal{C}_{\lambda,\varepsilon}$, provided each resulting covariance is positive semidefinite.

C.4 Consistency of Anchor PCA

We now state the consistency results of the empirical $\texttt{AnchorPCA}_\lambda$ estimator and of the block-stabilized empirical $\texttt{AnchorPCA}_\infty$ construction (§ 5.1), already stated in § 5.3 informally. Corollary C.13 then shows that, when $\mathcal{S}_\star \neq \{0\}$, the first block returned by $\texttt{AnchorPCA}_\infty$ consistently estimates $\mathcal{S}_\star$. Throughout, for all $e \in \mathcal{E}$, let $x_{e,1}, x_{e,2}, \ldots, x_{e,n_e}$ be i.i.d. copies of x_e, and assume $n := \min_{e\in\mathcal{E}} n_e \to \infty$.

Theorem C.11 (Consistency of the empirical projector-agreement estimators). *Assume that, for all $e \in \mathcal{E}$, $\mathbb{E}[x_e] = 0$, $\mathbb{E}\|x_e\|_2^4 < \infty$, and $\lambda_k(\Sigma_e) > \lambda_{k+1}(\Sigma_e)$. Then the following statements hold.*

(i) Let $\lambda \geq 0$, and let $\widehat{\Pi}_k^{\perp,\lambda}$ be an empirical rank-k $\texttt{AnchorPCA}_\lambda$ solution projector. If $\lambda_k(M_\lambda) > \lambda_{k+1}(M_\lambda)$ then

$$\left\| \widehat{\Pi}_k^{\perp,\lambda} - \Pi_k^{\perp,\lambda} \right\|_{\mathrm{op}} \to 0 \qquad \text{almost surely as } n \to \infty.$$

(ii) Let $\widehat{\Pi}_k^{\perp,\infty}$ be the block-stabilized $\texttt{AnchorPCA}_\infty$ estimator with tolerance $\mathrm{tol}_n > 0$. Assume that there exists $\alpha \in (0, 1/2)$ and $c > 0$ such that for all n, $\mathrm{tol}_n = cn^{-\alpha}$ and the population rank-k $\texttt{AnchorPCA}_\infty$ solution projector is unique, then

$$\left\| \widehat{\Pi}_k^{\perp,\infty} - \Pi_k^{\perp,\infty} \right\|_{\mathrm{op}} \to 0 \qquad \text{almost surely as } n \to \infty.$$

Remark C.12. The uniqueness assumption in part (ii) concerns the population projector, i.e. the optimal subspace, not any particular basis. For example, it holds automatically if there is a $\overline{\Pi}$-eigenvalue block boundary at rank-k. Otherwise, it requires an eigengap condition (23).

To prove Thm. C.11, we use three auxiliary lemmas collected in § C.1: Lemma C.4 transfers operator-norm perturbation bounds to leading eigenspace projectors, Lemma C.5 gives polynomial almost-sure rates for $\widehat{\Sigma}$ and $\widehat{\Pi}$, and Lemma C.6 guarantees recovery of the population spectral blocks of $\overline{\Pi}$ by the empirical block clustering.

Proof of Thm. C.11. Choose any $\alpha_0 \in (0, 1/2)$. By Lemma C.5,

$$\left\|\widehat{\Sigma} - \overline{\Sigma}\right\|_{\mathrm{op}} \to 0 \qquad \text{and} \qquad \left\|\widehat{\Pi} - \overline{\Pi}\right\|_{\mathrm{op}} \to 0 \qquad \text{almost surely.} \tag{19}$$

Statement (i): $\texttt{AnchorPCA}_\lambda$. By (19), there is an event Ω_0 with $\mathbb{P}(\Omega_0) = 1$ on which $\left\|\widehat{M}_\lambda - M_\lambda\right\|_{\mathrm{op}} \le \left\|\widehat{\Sigma} - \overline{\Sigma}\right\|_{\mathrm{op}} + 2E\lambda \left\|\widehat{\Pi} - \overline{\Pi}\right\|_{\mathrm{op}} \to 0$ almost surely, where $\widehat{M}_\lambda = \widehat{\Sigma} + 2E\lambda\widehat{\Pi}$, $M_\lambda = \overline{\Sigma} + 2E\lambda\overline{\Pi}$. Let $\gamma_k(M_\lambda) := \lambda_k(M_\lambda) - \lambda_{k+1}(M_\lambda)$; by assumption, $\gamma_k(M_\lambda) > 0$. Fix a realization in Ω_0. Then $\left\|\widehat{M}_\lambda - M_\lambda\right\|_{\mathrm{op}} < \gamma_k(M_\lambda)/2$ for all $n \ge N$ (with N depending on the realization), so Lemma C.4 (with $A = M_\lambda$, $\widehat{A} = \widehat{M}_\lambda$, $d = k$) gives

$$\left\|\widehat{\Pi}_k^{\perp,\lambda} - \Pi_k^{\perp,\lambda}\right\|_{\mathrm{op}} \le \frac{2}{\gamma_k(M_\lambda)} \left\|\widehat{M}_\lambda - M_\lambda\right\|_{\mathrm{op}} \to 0.$$

As this holds on Ω_0, the convergence is almost sure.

Statement (ii): $\texttt{AnchorPCA}_\infty$. Recall that the block-stabilized construction of the solution to $\texttt{AnchorPCA}_\infty$ (§ 5.1) scans the ordered eigenvalues of $\widehat{\Pi}$ from largest to smallest to group them into empirical blocks: each empirical block starts at an eigenvalue and includes subsequent eigenvalues as long as they are at most tol_n smaller than the first eigenvalue of the block.

Let $1 \ge \rho_1 > \cdots > \rho_L \ge 0$ be the distinct eigenvalues of $\overline{\Pi}$, and let Π_ℓ denote the orthogonal projector onto the eigenspace corresponding to ρ_ℓ. Let $d_\ell := \mathrm{rank}(\Pi_\ell)$ and $c_\ell := \sum_{j<\ell} d_j$. Listing the eigenvalues of $\overline{\Pi}$ in nonincreasing order with multiplicity, ρ_ℓ occupies a block of d_ℓ consecutive indices $I_\ell := \{c_\ell + 1, \ldots, c_\ell + d_\ell\}$, so that $I_1, \ldots, I_L$ partition $\{1, \ldots, p\}$. Let $\varepsilon_n := \|\widehat{\Pi} - \overline{\Pi}\|_{\mathrm{op}}$ and $\xi_n := \|\widehat{\Sigma} - \overline{\Sigma}\|_{\mathrm{op}}$.

Step 1: Match the indices of the empirical and population blocks. By Lemma C.5, with the same α as in $\mathrm{tol}_n = cn^{-\alpha}$, we have $\varepsilon_n = o(n^{-\alpha})$ and $\xi_n = o(n^{-\alpha})$ almost surely. Since $\mathrm{tol}_n = cn^{-\alpha}$ with $c > 0$, this implies $\varepsilon_n = o(\mathrm{tol}_n)$ almost surely. We now work on an event Ω_1 with $\mathbb{P}(\Omega_1) = 1$ on which these convergences hold. Fix a realization in Ω_1; every convergence in the remainder of the proof is along this realization, and each "for all sufficiently large n" refers to a realization-dependent threshold. Then, for all sufficiently large n,

$$\varepsilon_n < \mathrm{tol}_n/4. \tag{20}$$

Moreover, since tol_n decreases monotonically to 0 deterministically, for all sufficiently large n we also have

$$\mathrm{tol}_n < \gamma_{\overline{\Pi}}/2, \tag{21}$$

where $\gamma_{\overline{\Pi}}$ is the smallest gap between distinct eigenvalues of $\overline{\Pi}$ as defined in Lemma C.6. Hence Lemma C.6 applies, and the block-stabilized construction in § 5.1 partitions the indices $1, \ldots, p$ into empirical blocks that coincide with $I_1, \ldots, I_L$. For each $\ell \in \{1, \ldots, L\}$, let $\widehat{\Pi}_\ell$ denote the orthogonal projector onto the span of the eigenvectors of $\widehat{\Pi}$ with indices in I_ℓ, i.e. onto $\widehat{\mathrm{Eig}}_\ell$ in the notation of § 5.1; $\widehat{\mathrm{Eig}}_\ell$ estimates $\mathrm{Im}(\Pi_\ell)$, and it has dimension d_ℓ.

Step 2: Show consistency of each block projector. We show $\left\|\widehat{\Pi}_\ell - \Pi_\ell\right\|_{\mathrm{op}} \to 0$ for all $\ell \in \{1, \ldots, L\}$. If $L = 1$, then $\overline{\Pi}$ has a single eigenvalue, so $\Pi_1 = I_p$. For all sufficiently large n, the empirical block construction also returns a single block, so $\widehat{\Pi}_1 = I_p = \Pi_1$, which proves the claim. Assume now that $L \ge 2$, and let $\ell \in \{1, \ldots, L\}$. Apply Thm. C.3 with $\Sigma = \overline{\Pi}$, $\widehat{\Sigma} = \widehat{\Pi}$, $r = c_\ell + 1$, and $s = c_\ell + d_\ell$: then V spans $\mathrm{Im}(\Pi_\ell)$, $\widehat{V}$ spans $\mathrm{Im}(\widehat{\Pi}_\ell)$, and $[\lambda_s(\overline{\Pi}), \lambda_r(\overline{\Pi})] = \{\rho_\ell\}$. It remains

to lower bound the separation δ between ρ_ℓ and the empirical eigenvalues outside I_ℓ (defined in Thm. C.3). By Weyl's inequality [21, Theorem 4.3.1] in decreasing-order notation applied to $\overline{\Pi}$ and $\widehat{\Pi} - \overline{\Pi}$ at indices c_ℓ and $c_\ell + d_\ell + 1$,

$$\widehat{\rho}_{c_\ell} \geq \rho_{\ell-1} - \varepsilon_n \qquad \text{and} \qquad \widehat{\rho}_{c_\ell + d_\ell + 1} \leq \rho_{\ell+1} + \varepsilon_n,$$

where for $\ell = 1$ there is no eigenvalue above the block and for $\ell = L$ none below, so only the relevant one-sided bound is used. In either case, the empirical eigenvalues outside I_ℓ lie at distance at least $\gamma_{\overline{\Pi}} - \varepsilon_n$ from ρ_ℓ, so $\delta \geq \gamma_{\overline{\Pi}} - \varepsilon_n > 0$, where the last inequality follows from $\varepsilon_n < \mathrm{tol}_n/2 < \gamma_{\overline{\Pi}}/4$ using (20) and (21). Thus, we get

$$\left\|\widehat{\Pi}_\ell - \Pi_\ell\right\|_{\mathrm{op}} = \left\|\sin\Theta(\widehat{V}, V)\right\|_{\mathrm{op}} \leq \frac{\varepsilon_n}{\gamma_{\overline{\Pi}} - \varepsilon_n} \to 0,$$

where we additionally used Prop. C.2 for the first equality as $\widehat{\Pi}_\ell$ and Π_ℓ have equal rank d_ℓ (Step 1).

Step 3: Assemble the rank-k projector. Let $r \in \{1, \ldots, L\}$ be the unique index such that $c_r < k \leq c_r + d_r$, and let $q := k - c_r$, so, in the population case, blocks $1, \ldots, r-1$ are taken in full and q directions are taken from block r.

Case 1: $q = d_r$. The first r blocks are fully taken for the population solution. By Steps 1a and 1b in § B.1.2, the population $\texttt{AnchorPCA}_\infty$ projector is $\Pi_k^{\perp,\infty} = \sum_{\ell=1}^{r} \Pi_\ell$. For all sufficiently large n, the empirical construction selects the first r empirical blocks, so $\widehat{\Pi}_k^{\perp,\infty} = \sum_{\ell=1}^{r} \widehat{\Pi}_\ell$. By Step 2,

$$\left\|\widehat{\Pi}_k^{\perp,\infty} - \Pi_k^{\perp,\infty}\right\|_{\mathrm{op}} \leq \sum_{\ell=1}^{r} \left\|\widehat{\Pi}_\ell - \Pi_\ell\right\|_{\mathrm{op}} \to 0.$$

Case 2: $q < d_r$. Blocks $1, \ldots, r-1$ are taken in full, and the secondary variance criterion selects q directions from block r. Let $B_r := \Pi_r \overline{\Sigma} \Pi_r$ and $\widehat{B}_r := \widehat{\Pi}_r \widehat{\Sigma} \widehat{\Pi}_r$, and let Q_r and $\widehat{Q}_r$ denote the orthogonal projectors onto the spans of the top q eigenvectors of B_r and $\widehat{B}_r$, respectively. We now show that $\widehat{B}_r \to B_r$ in operator norm. Note that

$$\widehat{B}_r - B_r = \widehat{\Pi}_r(\widehat{\Sigma} - \overline{\Sigma})\widehat{\Pi}_r + (\widehat{\Pi}_r - \Pi_r)\overline{\Sigma}\widehat{\Pi}_r + \Pi_r\overline{\Sigma}(\widehat{\Pi}_r - \Pi_r),$$

together with $\left\|\widehat{\Pi}_r\right\|_{\mathrm{op}} = \|\Pi_r\|_{\mathrm{op}} = 1$ and (19) gives

$$\|\widehat{B}_r - B_r\|_{\mathrm{op}} \leq \xi_n + 2\|\overline{\Sigma}\|_{\mathrm{op}}\|\widehat{\Pi}_r - \Pi_r\|_{\mathrm{op}} \to 0. \tag{22}$$

By Steps 1a and 1b in § B.1.2, the population $\texttt{AnchorPCA}_\infty$ projector is $\Pi_k^{\perp,\infty} = \sum_{\ell=1}^{r-1} \Pi_\ell + Q_r$, where the first sum is a fixed projector independent of the selection within block r. Hence $Q_r = \Pi_k^{\perp,\infty} - \sum_{\ell=1}^{r-1} \Pi_\ell$ is determined by $\Pi_k^{\perp,\infty}$, so the uniqueness assumption forces the top-q eigenspace of B_r to be unique; equivalently, B_r has a positive eigengap at rank q:

$$\eta_r := \lambda_q(B_r) - \lambda_{q+1}(B_r) > 0. \tag{23}$$

Thus, for all sufficiently large n, $\|\widehat{B}_r - B_r\|_{\mathrm{op}} < \eta_r/2$. Applying Lemma C.4 with $A = B_r$, $\widehat{A} = \widehat{B}_r$, and $d = q$, and then using (22) gives

$$\left\|\widehat{Q}_r - Q_r\right\|_{\mathrm{op}} \leq \frac{2\|\widehat{B}_r - B_r\|_{\mathrm{op}}}{\eta_r} \to 0.$$

For all sufficiently large n, the empirical and population $\texttt{AnchorPCA}_\infty$ projectors can be written as

$$\widehat{\Pi}_k^{\perp,\infty} = \sum_{\ell=1}^{r-1} \widehat{\Pi}_\ell + \widehat{Q}_r, \qquad \Pi_k^{\perp,\infty} = \sum_{\ell=1}^{r-1} \Pi_\ell + Q_r.$$

Hence

$$\left\|\widehat{\Pi}_k^{\perp,\infty} - \Pi_k^{\perp,\infty}\right\|_{\mathrm{op}} \leq \sum_{\ell=1}^{r-1} \left\|\widehat{\Pi}_\ell - \Pi_\ell\right\|_{\mathrm{op}} + \left\|\widehat{Q}_r - Q_r\right\|_{\mathrm{op}} \to 0.$$

All displayed convergences hold for the fixed realization in Ω_1. Since $\mathbb{P}(\Omega_1) = 1$, it follows that $\left\|\widehat{\Pi}_k^{\perp,\infty} - \Pi_k^{\perp,\infty}\right\|_{\mathrm{op}} \to 0$ almost surely. □

Corollary C.13 (Consistency of the empirical invariant block)**.** *Assume the moment and eigengap conditions of Thm. C.11 and a tolerance* $\mathrm{tol}_n = cn^{-\alpha}$ *with* $\alpha \in (0, 1/2)$, $c > 0$ *(as in part (ii)); suppose further that* $\mathcal{S}_\star \neq 0$*. Let* $\widehat{\Pi}_1$ *denote the first empirical block returned by* $\texttt{AnchorPCA}_\infty$*, i.e., the projector onto* $\widehat{\mathrm{Eig}}_1$ *Then*

$$\left\|\widehat{\Pi}_1 - \Pi_{\mathcal{S}_\star}\right\|_{\mathrm{op}} \to 0 \qquad \textit{almost surely.}$$

Equivalently, $\widehat{\mathrm{Eig}}_1$ *consistently estimates* $\mathcal{S}_\star$*.*

Proof. The eigenvalue-one block of $\overline{\Pi}$ is $\mathcal{S}_\star$, so $\Pi_1 = \Pi_{\mathcal{S}_\star}$. Step 2 in the proof of Thm. C.11 *(ii)* gives $\left\|\widehat{\Pi}_\ell - \Pi_\ell\right\|_{\mathrm{op}} \to 0$ almost surely for every ℓ; the case $\ell = 1$ is the claim. □

D Details on experiments

This appendix provides additional details for the motivating example in § 1.2, the synthetic $\mathcal{S}_\star$ recovery simulations in § 6.1, and the gas-sensor drift application in § 6.2. Code to reproduce all simulations and figures is publicly available at https://github.com/benesei/anchor-pca.

D.1 Details of the motivating example

This appendix records the explicit construction used in the motivating example (§ 1.2) which is also used for the perturbation-path plot Fig. 2.

Construction. Recalling § 1.2, let $c_1, \dots, c_4$ be the canonical basis vectors of $\mathbb{R}^4$, and set $a := c_1, b := c_2, u := c_3$. We further define two directions in the c_3–c_4 plane,

$$v := \cos(50^\circ)c_3 + \sin(50^\circ)c_4, \qquad w := \cos(100^\circ)c_3 + \sin(100^\circ)c_4,$$

and let $u^\perp, v^\perp, w^\perp$ denote the unit directions in the c_3–c_4 plane that are orthogonal to u, v, w, respectively. The population covariance matrices of the three domains are then defined as

$$\begin{aligned}
\Sigma_1 &= 220\, uu^\top + 140\, aa^\top + 90\, bb^\top + 25\, u^\perp u^{\perp\top},\\
\Sigma_2 &= 120\, aa^\top + 90\, vv^\top + 70\, bb^\top + 10\, v^\perp v^{\perp\top},\\
\Sigma_3 &= 320\, ww^\top + 120\, bb^\top + 80\, aa^\top + 10\, w^\perp w^{\perp\top};
\end{aligned}$$

these are shown in Table 1.

Recovered subspaces. The solution of poolPCA spans $\mathrm{span}(a, c_3, c_4)$ and therefore drops b. Both $\texttt{AnchorPCA}_{25}$ and $\texttt{AnchorPCA}_\infty$ recover $\mathrm{span}(a, b)$. Their difference lies only in the third direction: $\texttt{AnchorPCA}_\infty$ selects v, and $\texttt{AnchorPCA}_{25}$ selects $q = 0.246c_3 + 0.969c_4$, a direction capturing slightly more variance on average than v (126.47 vs. 111.22), but is less invariant ($q^\top \overline{\Pi} q = 0.57 < 0.61 = v^\top \overline{\Pi} v$).

D.2 Details on the perturbation-path illustration

To visualize Thm. 4.1, we reuse the population motivating example and plot, for several methods (fit on the population source covariances), the worst-case average reconstruction error over $\mathcal{C}_\lambda$ as the perturbation strength $\rho := 2E\lambda$ increases. By Eq. (5), for any rank-k projector Π, this worst-case error is

$$\frac{1}{E}\sum_{e=1}^{E} \mathrm{Tr}\big((I_p - \Pi)\Sigma_e\big) + \frac{\rho}{2E}\sum_{e=1}^{E}\left\|\Pi - \Pi_k^{(e)}\right\|_{\mathrm{F}}^2,$$

so each method traces a line in ρ: the intercept is its average reconstruction error on the source covariances, and the slope is its disagreement with the local top-k projectors. The oracle minimizer at perturbation strength ρ is $\texttt{AnchorPCA}_{\rho/(2E)}$. We plot these lines for poolPCA, $\texttt{AnchorPCA}_{25}$, $\texttt{AnchorPCA}_\infty$, and the wcPCA baselines maxRCS, maxRegret, and norm-maxRegret from [17].

Fig. 2 is an unsupervised analog of the perturbation-path plot in Anchor Regression [37]: `poolPCA` is optimal only near $\rho = 0$, a fixed finite-penalty $\texttt{AnchorPCA}_\lambda$ is optimal over a broad middle range of perturbations centered around its penalty, and $\texttt{AnchorPCA}_\infty$ is the right choice once robustness to top-k variance inflation matters more than reconstruction accuracy. The figure additionally shows that, in this example, the wcPCA baselines `norm-maxRegret` and `maxRCS` perform best only on short transition windows near the origin, though they target a different form of distribution shift: shifts within the convex hull of the source covariances.

D.3 Details on the simulations

We now describe the details of the experiment described in § 6.1.

D.3.1 Data-generating process

Notation. We write $E = |\mathcal{E}|$ for the number of domains, p for the number of covariates. Each domain $e \in \mathcal{E}$ is associated with a covariance Σ_e, whose construction we detail in the following paragraphs. Let k be the target rank for the low-dimensional representation, and $m = \dim(\mathcal{S}_\star)$ be the dimension of the invariant subspace $\mathcal{S}_\star$. Let

$$q := p - k, \qquad d := p - m.$$

So, q is the dimension of the complement of each top-k subspace, and d is the dimension of $\mathcal{S}_\star^\perp$.

Throughout this section, a *configuration* corresponds to one choice of (E, p, k, m); for a given configuration and eigenvalue regime (see § D.3.1 below), a *distribution draw* is a multi-domain distributional instance, namely the full collection of covariances $(\Sigma_e)_{e\in\mathcal{E}}$ with its own $\mathcal{S}_\star$ (if we talk about multiple distribution draws, we assume they are chosen i.i.d.); a *sample* is one finite dataset drawn from a fixed distribution draw; and individual variables $X_{e,i}$ are called observations or data points.

Generating population covariance matrices. Each population covariance is constructed so that the top-k eigenspaces of the E covariances have a nontrivial maximal intersection of dimension m. We build this in three steps.

1. *Invariant subspace.* We draw a Haar orthogonal matrix $U = (u_1, \dots, u_p) \in \mathbb{R}^{p\times p}$ and split its columns into
$$S := (u_1, \dots, u_m), \qquad R := (u_{m+1}, \dots, u_p),$$
so that S is an orthonormal basis of $\mathcal{S}_\star := \mathrm{Im}(S)$ and R is an orthonormal basis of $\mathcal{S}_\star^\perp$. The remaining construction works in R-coordinates inside $\mathcal{S}_\star^\perp$.
2. *Bottom subspace per domain.* For each domain $e \in \mathcal{E}$, we draw a bottom subspace of dimension q inside $\mathcal{S}_\star^\perp$: sample a $d \times q$ matrix with i.i.d. standard Gaussian entries, take the Q factor of its reduced QR decomposition to obtain an orthonormal matrix $H_e \in \mathbb{R}^{d\times q}$, and set $B_e := RH_e$. If the resulting collection $(B_1, \dots, B_E)$ does not jointly span $\mathcal{S}_\star^\perp$, we discard this collection and redraw all E bottom subspaces until
$$\dim \mathrm{span}(B_1, \dots, B_E) = p - m,$$
which is what makes the intersection of the top-k subspaces equal to $\mathcal{S}_\star$ exactly, rather than something larger.
3. *Top subspace per domain.* The remaining $k - m$ top directions in domain $e \in \mathcal{E}$ lie in the complement of B_e inside $\mathcal{S}_\star^\perp$. Let $G_e \in \mathbb{R}^{d\times(k-m)}$ be an orthonormal basis of $\ker(H_e^\top)$ and set $C_e := RG_e$. The top-k eigenspace for all $e \in \mathcal{E}$ is then
$$\mathcal{U}_e := \mathrm{Im}(S, C_e) = \mathcal{S}_\star \oplus \mathrm{Im}(C_e).$$
By construction, $\bigcap_{e\in\mathcal{E}} \mathcal{U}_e = \bigcap_{e\in\mathcal{E}} \mathrm{Im}(B_e)^\perp = \big(\bigcup_{e\in\mathcal{E}} \mathrm{Im}(B_e)\big)^\perp = (\mathcal{S}_\star^\perp)^\perp = \mathcal{S}_\star$.

For all $e \in \mathcal{E}$, given the orthonormal basis (S, C_e, B_e), we define the covariance matrix as

$$\Sigma_e := S\,\mathrm{diag}(\theta_{e,\star})S^\top + C_e\,\mathrm{diag}(\theta_{e,\mathrm{env}})C_e^\top + B_e\,\mathrm{diag}(\eta_e)B_e^\top,$$

where $\theta_{e,\star} \in \mathbb{R}^m$ are the eigenvalues on the invariant directions, $\theta_{e,\mathrm{env}} \in \mathbb{R}^{k-m}$ on the domain-specific top directions, and $\eta_e \in \mathbb{R}^q$ on the bottom directions. The eigenvalue regimes below ensure a local k versus $k + 1$ eigengap in every domain.

Easy and hard eigenvalue regimes. We use the same subspace construction with two different choices of eigenvalues. In the *easy* regime, the invariant and domain-specific top eigenvalues are well separated from the bottom eigenvalues:

$$\theta_{e,\star,j} \sim U[5,8], \qquad \theta_{e,\mathrm{env},j} \sim U[5,8], \qquad \eta_{e,j} \sim U[0.5,3].$$

In the *hard* regime, the invariant directions still sit in each top-k subspace but carry less variance than the domain-specific top directions:

$$\theta_{e,\star,j} \sim U[3.2,4.2], \qquad \theta_{e,\mathrm{env},j} \sim U[6,9], \qquad \eta_{e,j} \sim U[0.5,2.5].$$

All distribution draws are generated with master seed 42. For each configuration (p,k,m,E) and each eigenvalue regime, we generate 100 independent distribution draws; $\mathcal{S}_\star$ is redrawn for every replicate.

Gaussian sampling. Given a distribution draw and sample size N, we sample N data points independently,

$$X_{e,i} \sim \mathcal{N}(0,\Sigma_e), \qquad i \in \{1,\dots,N\}, \qquad e \in \mathcal{E}.$$

For each distribution draw and sample size N, we generate 20 independent samples, each consisting of N observations per domain. The standard sample-size grid is

$$N \in \{50,100,200,500,1000,2000,5000\}.$$

For the small-E configuration described below, we extend the grid to include $\{10000, 30000, 100000\}$.

Non-Gaussian covariance-preserving sampling. As a robustness check, we also use a symmetric two-component Gaussian mixture with the same mean and covariance as the Gaussian model. For each domain, let r_e be the leading eigenvector of Σ_e and set $\mu_e = 0.75\sqrt{\lambda_{\max}(\Sigma_e)}\, r_e$; if needed, this vector is shrunk to ensure that $\Sigma_e - \mu_e\mu_e^\top$ is positive semidefinite. We then sample for $i \in \{1,\dots,N\}, e \in \mathcal{E}$,

$$X_{e,i} = Z_{e,i} + \xi_{e,i}\mu_e, \qquad Z_{e,i} \sim \mathcal{N}(0, \Sigma_e - \mu_e\mu_e^\top), \qquad \mathbb{P}(\xi_{e,i}=1) = \mathbb{P}(\xi_{e,i}=-1) = 1/2.$$

Thus

$$\mathbb{E}X_{e,i} = 0, \qquad \mathrm{Cov}(X_{e,i}) = \Sigma_e.$$

The mixture preserves the population eigenspaces but is not Gaussian.

D.3.2 Estimators

In the experiments, we consider $\texttt{AnchorPCA}_\infty$ and $\texttt{FindS}_\star$. $\texttt{AnchorPCA}_\infty$ is fit with the default block tolerance of the package[3]. $\mathrm{tol}_N := \min\{0.05,\ 0.5\, n_{\min}^{-0.4}\}$. In these experiments, each domain has the same sample size, so $n_{\min} = N$. The first empirical block contains all eigenvalues $\widehat{\rho}_j$ satisfying $\widehat{\rho}_1 - \widehat{\rho}_j \le \mathrm{tol}_N$, where $\widehat{\rho}_i$ is the ith largest eigenvalue of $\widehat{\Pi}$. Thus the oracle block corresponding to $\mathcal{S}_\star$ is separated whenever $\widehat{\rho}_1 - \widehat{\rho}_m \le \mathrm{tol}_N < \widehat{\rho}_1 - \widehat{\rho}_{m+1}$. $\texttt{FindS}_\star$ is used with significance level $\alpha = 0.05$.

D.3.3 Recovery configurations

We consider three configurations (E,p,k,m).

1. *Main random-subspace configuration:* $(E,p,k,m) = (5,10,5,2)$, shown in the main text.
2. *Small-m configuration:* $(E,p,k,m) = (5,8,5,1)$ shown in Fig. 6.
3. *Small-E configuration:* $(E,p,k,m) = (2,8,5,2)$ shown in Fig. 7. This configuration uses the minimal feasible invariant dimension. Indeed, since each domain has a $q = p-k = 3$ dimensional bottom space and the bottom spaces must span $\mathcal{S}_\star^\perp$, feasibility requires

$$p - m \le Eq, \qquad \text{equivalently} \qquad m \ge p - E(p-k) = 2.$$

[3] https://github.com/benesei/anchor-pca

D.3.4 Recovering the invariant subspace $\mathcal{S}_\star$

We now repeat the experiments of Fig. 3 for non-Gaussian data, small-m, and small-E configurations.

Performance quantities and aggregation. For each distribution draw and sample size N, we compute two quantities over the 20 independent samples: the dimension-recovery indicator, equal to 1 if the estimated first-block dimension equals the true m, and the operator-norm projector error for the estimated invariant subspace. Within each distribution draw, we average the 20 dimension-recovery indicators and take the median of the 20 projector errors. The displayed dimension-recovery probability is the mean of these distribution-draw-level averages across distribution draws, while the displayed subspace error is the median of the distribution-draw-level median projector errors. Shaded bands, when shown, indicate the 10%–90% quantiles across distribution draws.

Agreement-separation gap of the configurations. For a distribution draw, we define the agreement-separation gap as

$$\gamma_\Pi := \rho_m(\overline{\Pi}) - \rho_{m+1}(\overline{\Pi}) = \rho_1(\overline{\Pi}) - \rho_{m+1}(\overline{\Pi}) = 1 - \rho_{m+1}(\overline{\Pi}),$$

where $\rho_i(\overline{\Pi})$ denotes the ith largest eigenvalue of $\overline{\Pi}$ and the second and third equality use $\rho_1(\overline{\Pi}) = \cdots = \rho_m(\overline{\Pi}) = 1$ on $\mathcal{S}_\star$. The gap γ_Π measures how cleanly $\mathcal{S}_\star$ separates from the rest of the agreement spectrum and reflects one aspect of difficulty of the configuration. Pooling the 100 easy and 100 hard distribution draws for each configuration, the median γ_Π values, with 10–90% ranges, are

configuration	median γ_Π (10%–90%)
main random-subspace	0.277 (0.212–0.343)
small-m	0.119 (0.072–0.184)
small-E	$9.8 \cdot 10^{-3}$ ($1.8 \cdot 10^{-4}$–$5.5 \cdot 10^{-2}$)

The gap decreases across the three configurations, illustrating increased hardness. For the small-E configuration, the agreement gap is often very small, so this configuration is best viewed as a stress test rather than a typical regime.

At the empirical level, we analogously track $\widehat{\gamma}_\Pi := \rho_1(\widehat{\Pi}) - \rho_{m+1}(\widehat{\Pi})$. This is the quantity relevant for the block-stabilized estimator: the first empirical block has correct dimension m if

$$\rho_1(\widehat{\Pi}) - \rho_m(\widehat{\Pi}) \le \mathrm{tol}_N < \rho_1(\widehat{\Pi}) - \rho_{m+1}(\widehat{\Pi}).$$

Non-Gaussianity. Fig. 5 repeats the main-text experiment from Fig. 3 ($E = 5$, $p = 10$, $k = 5$, $m = 2$), but draws data from the covariance-preserving Gaussian mixtures introduced in § D.3.1 instead of Gaussians. The recovery curves are close to the Gaussian ones in Fig. 3. This is in line with § 5.3, which does not require Gaussianity for $\texttt{AnchorPCA}_\infty$ under its moment assumptions. The dashed curves show performance of $\texttt{FindS}_\star$ and are included to test its robustness to violations of its Gaussian data assumption.

Small-m configuration. Fig. 6 shows Gaussian results for the small-m configuration ($E = 5$, $p = 8$, $k = 5$, $m = 1$). This configuration has a smaller population agreement-separation gap than the main configuration, but the finite-sample recovery curves are comparable on the standard grid. In both eigenvalue regimes, $\texttt{AnchorPCA}_\infty$ reaches correct first-block dimension recovery by the largest standard-grid sample sizes.

Small-E configuration. For the small-E configuration ($E = 2$, $p = 8$, $k = 5$, $m = 2$), we use the extended sample-size grid up to $N = 100000$. This configuration has a much smaller population agreement-separation gap than the other configurations, so we expect it to be a particularly difficult case. Fig. 7 shows the same quantities as Fig. 5 but also shows the tolerance diagnostic. In the top row, the first-block dimension is correct at $N = 100000$ in about 0.68 of easy-regime runs and 0.57 of hard-regime runs. $\texttt{FindS}_\star$ improves dimension recovery in the easy regime, reaching about 0.96 correct recovery, but in the hard regime it tends to overestimate the dimension, with median $\widehat{m} = 3$. Intuitively, this is indeed a difficult case for $\texttt{FindS}_\star$ because it estimates m indirectly, by testing $r = \dim(\mathrm{span}\{I_p - \Pi_1^{(k)}, I_p - \Pi_2^{(k)}\}) = \dim(\mathcal{S}_\star^\perp) = p - m$ for all $p - k$ to $p - m_{\min}$; see § B.3. Thus, since $p - m = 6$ and $\mathrm{rank}(I_p - \Pi_e^{(k)}) = 3$, this configuration has the smallest

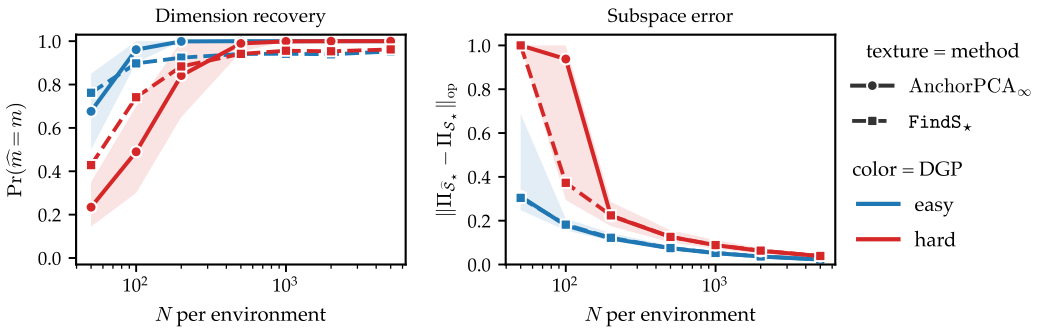

Figure 5: **Main random-subspace configuration with Gaussian mixtures.** Same distribution draws and aggregation as in Fig. 3, but observations are drawn from a symmetric two-component Gaussian mixture preserving each covariance Σ_e. The panels show the fraction of samples for which m was estimated correctly and median projector error. Solid lines use the first grouped eigenspace from AnchorPCA$_\infty$, while dashed lines show FindS$_\star$ using significance level 0.05. Shaded bands are 10–90% quantiles across distribution draws and are shown for AnchorPCA$_\infty$. As suggested by the consistency results (§ 5.3), for AnchorPCA$_\infty$ the estimated recovery probability converges to one and the subspace error to zero and, for FindS$_\star$, the estimated recovery probability approaches 0.95. The close agreement with the Gaussian curves in Fig. 3 suggests that, in this configuration, the AnchorPCA$_\infty$ block-recovery mechanism and FindS$_\star$ are robust to this covariance-preserving non-Gaussian sampling perturbation.

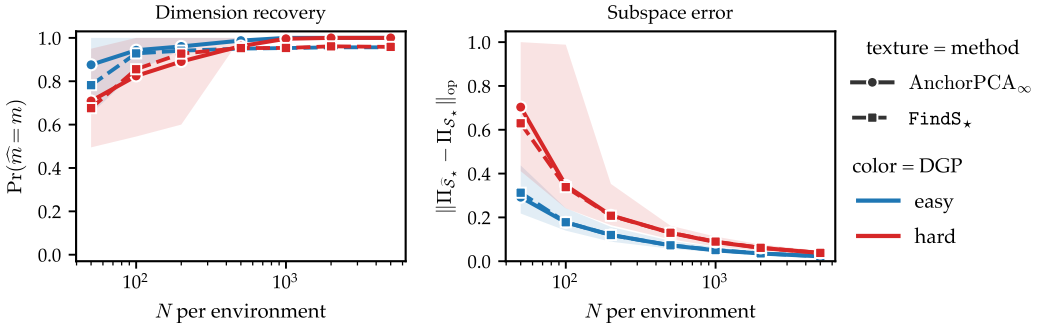

Figure 6: **Small-m configuration.** Plots show the same as Fig. 5 for the small-m configuration. Again, the results are as suggested by § 5.3: e.g., in the left plot, the errors of AnchorPCA$_\infty$ and FindS$_\star$ approach zero and 0.05, respectively.

possible $m = m_{\min} = 2$; equivalently, $r = p - m$ is maximal, so the correct dimension $m = 2$ is reached only if the test rejects all smaller candidate dimensions $r = 3, 4, 5$, leading to a small empirical power of the test. If sampling noise makes it fail to reject $r = 5$, it stops there and returns $\widehat{m} = p - 5 = 3$. AnchorPCA$_\infty$ uses a different criterion: it thresholds the empirical eigenvalues directly using tol_N, so it can still return $m = 2$ whenever the first two eigenvalues form a separated block, even when sampling noise makes $\mathcal{S}_\star^\perp$ appear five dimensional. The bottom row of Fig. 7 illustrates why the small-E configuration is also hard for AnchorPCA$_\infty$: $\widehat{\gamma}_\Pi$ is much smaller than in the main configuration and is smaller than the default block tolerance over much of the extended grid. We hypothesize that the performance of AnchorPCA$_\infty$ could be improved by a more involved choice of the block tolerance.

D.3.5 Explained variance on an unseen domain.

For the simulations, we additionally measure explained variance on an unseen domain. For each of the easy and hard settings, we generate 50 independent distribution draws, each with six domains sharing the same $m = 2$-dimensional invariant subspace $\mathcal{S}_\star$ but otherwise having different covariances. Five

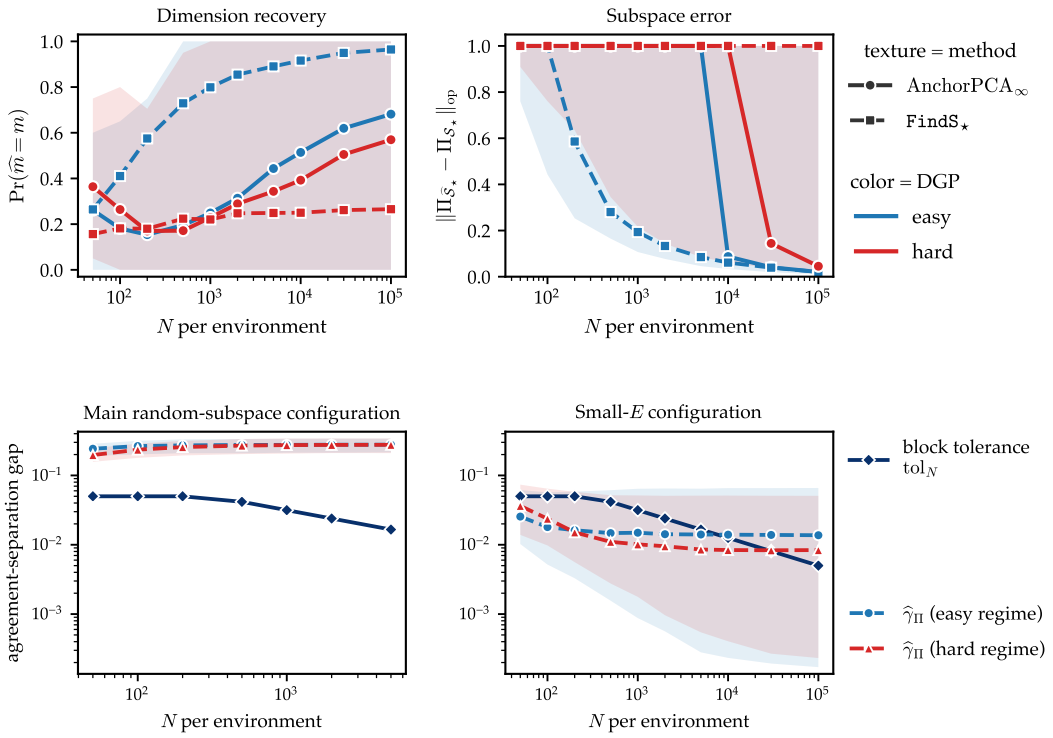

Figure 7: **Small-E configuration and agreement-separation gap.** The top row shows the same as Fig. 5 for the small-E configuration. As expected (see the discussion around γ_Π), this is a particularly hard case; indeed, the error of $\texttt{AnchorPCA}_\infty$ decreases but this time more slowly. Even for large N, some of the distribution draws are difficult in that for none of the 20 samples the estimated m is correct, which results in wide shaded areas. The bottom row (log scale) compares the default block tolerance tol_N (dark blue) with the median $\widehat{\gamma}_\Pi$ in the easy and hard regimes for the main random-subspace configuration (left) and the small-E configuration (right). The plot shows that the small-E setting is difficult because $\widehat{\gamma}_\Pi$ is much smaller than in the main configuration and, for many of the considered N, is smaller than tol_N.

domains are used as sources and the sixth is held out for evaluation. We use the covariance-preserving Gaussian mixtures from § D.3.1, with $p = 10$ and $k = 5$. We draw 500 observations per source domain and repeat the source-sample fit five times per distribution draw. We evaluate rank-k EV using the target population covariance.

To assess recovery of $\mathcal{S}_\star$, we also measure the operator-norm distance between the projector onto each method's first $m = 2$ directions and $\Pi_{\mathcal{S}_\star}$. Here we use the known dimension m, whereas the recovery experiment above estimates the dimension of the first agreement block. For both metrics, we first average over the five datasets within each distribution draw and the 50 draws in each setting. Table 2 shows that the two methods retain nearly identical target EV using all $k = 5$ directions, while the first $m = 2$ directions of $\texttt{AnchorPCA}_\infty$ recover $\mathcal{S}_\star$ more accurately than those of `poolPCA`.

D.4 Gas sensor array drift details and additional analyses

Data and domains. The UCI gas-sensor drift archive contains 13,910 measurements from 16 metal-oxide chemical sensors exposed to six gases at multiple concentration levels [15, 46], see Fig. 8. Each measurement comes from a 16-channel metal-oxide gas sensor array and is summarized by eight features per channel, giving $p = 128$ features in total. Two are steady-state features: the maximal resistance change DR with respect to the baseline, and a normalized version of DR, obtained by dividing DR by the sensor resistance measured under gas exposure. The remaining six are transient features computed from rising and decaying exponential moving averages of the signal. The archive is organized into ten temporal batches: B1 contains months 1–2, B2 months 3,4,8,9,10, B3 months

Table 2: Evaluation of Gaussian mixtures on an unseen domain. Target EV evaluates all $k = 5$ directions using the target population covariance (higher is better). Projector error measures recovery of $\mathcal{S}_\star$ using each method's first $m = 2$ directions (lower is better). Results are averaged over five datasets per distribution draw and 50 draws per setting.

Setting	Method	Target EV (%)	Projector error
Easy	`poolPCA`	57.81	0.226
Easy	$\texttt{AnchorPCA}_\infty$	57.75	0.078
Hard	`poolPCA`	49.44	0.999
Hard	$\texttt{AnchorPCA}_\infty$	49.33	0.133

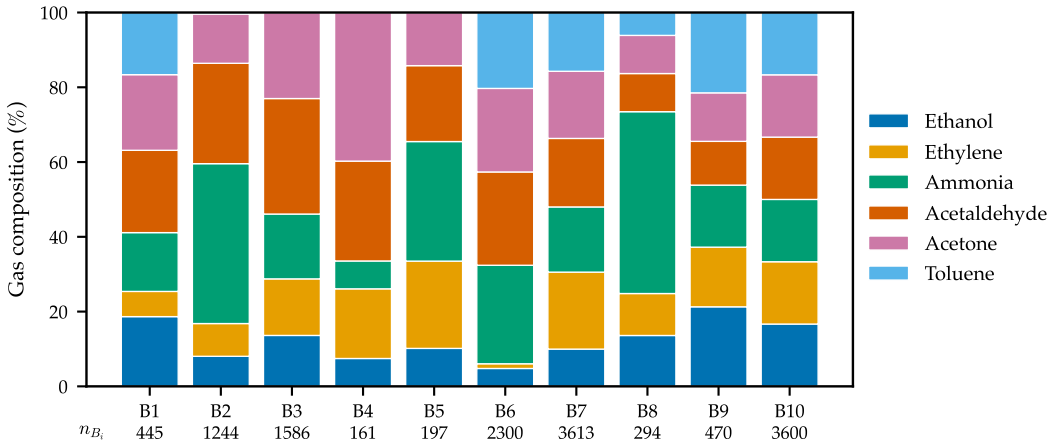

Figure 8: **Gas composition by temporal batch.** Stacked bars show the percentage of each gas class in batches B1–B10; numbers below the bars give the batch sizes n_{B_i}. The batches are strongly imbalanced and their gas mixtures vary over time, with some gases absent in several early batches and B10 balanced across the six gases.

11–13, B4 months 14–15, B5 month 16, B6 months 17–20, B7 month 21, B8 months 22–23, B9 months 24 and 30, and B10 month 36. We use these batches as domains. Data are available from the UCI Machine Learning Repository [46] at `https://doi.org/10.24432/C5MK6M` and are licensed under CC BY 4.0.

Baselines and hyperparameters. We compare `poolPCA`, fitted on the averaged source covariances; $\texttt{AnchorPCA}_{\widehat{\lambda}}$, using the selection procedure in § D.4; and $\texttt{AnchorPCA}_\infty$, using the block-stabilized estimator from § 5.1 with default tolerance. Lastly, we consider the worst-case PCA variant `norm-maxRegret` [17] which is fit with 10 restarts, 2000 iterations, learning rate 0.01, and deterministic seed $12020 + 10000s + 101k$.

Protocol. For each last source batch $s \in \{3, \dots, 8\}$, and for each $k \in \{10, 20, 30\}$, we fit the rank-k methods on B1–Bs and evaluate them on both the source batches and the held-out target batches B$(s+1)$–B10. For each split, feature standardization is applied once using the pooled source observations from B1–Bs only, and the resulting transformation is applied unchanged to all source and target batches. After this preprocessing, each batch covariance $\widehat{\Sigma}_b$ is computed from that batch's observations centered by their own empirical batch mean. For a representation $W \in \mathcal{O}_{p\times k}$ and batch covariance $\widehat{\Sigma}_b$, the reported percentage of explained variance (EV) is

$$\mathrm{EV}_b(W) := 100\,\frac{\mathrm{Tr}(W^\top \widehat{\Sigma}_b W)}{\mathrm{Tr}(\widehat{\Sigma}_b)}.$$

Source-only penalty selection. For each $s \in \{3, \dots, 8\}$ and each fixed $k \in \{10, 20, 30\}$, we select $2E\lambda$ by the leave-one-source-domain-out procedure in § 5.1, giving 18 configurations. For all

Table 3: B9 diagnostic after fitting on B1–B8. EV and PC1 capture are distinct percentages, as defined in the text.

k	Method	B9 EV (%)	B10 EV (%)	B9 PC1 capture (%)
20	poolPCA	94.84	92.72	95.51
20	AnchorPCA$_\infty$	68.09	95.85	66.00
30	poolPCA	99.06	96.43	99.59
30	AnchorPCA$_\infty$	73.32	97.99	71.57

penalty-selection experiments in this paper, we use the grid

$$2E\lambda \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, \infty\}.$$

For each choice of $2E\lambda$, we cross-validate over the source domains, holding each one out in turn. We estimate feature means and standard deviations from the remaining source observations, apply this transformation to the held-out domain, and fit AnchorPCA$_\lambda$ using the remaining domains' centered covariances. We score the fraction of total variance explained in the held-out source domain and select the penalty with the highest average score. Finally, we refit standardization and AnchorPCA$_\lambda$ on all source domains and evaluate on the target domains.

D.4.1 Rolling temporal splits.

Fig. 9 repeats the source–target evaluation of Fig. 4 across all source–target splits $s \in \{3, \dots, 8\}$. The top row shows mean source EV, the middle row mean target EV; shaded bands are min–max ranges across batches (shown only for poolPCA and AnchorPCA$_\infty$). The bottom row reports the selected penalties of AnchorPCA$_{\widehat{\lambda}}$.

On source batches, poolPCA has the highest EV as expected by its objective of maximizing average unnormalized explained variance. On target batches, AnchorPCA$_\infty$ has higher mean target EV than poolPCA in 16 of the 18 configurations. The two exceptions occur when fitting on B1–B8 with $k = 20$ or 30. In the following paragraph, we show that this is driven by B9, whose dominant principal direction is captured more strongly by poolPCA. AnchorPCA$_{\widehat{\lambda}}$ improves over poolPCA in 14 configurations, ties in two, and loses in two. It selects ∞ four times, a positive finite penalty twelve times, and zero (poolPCA) twice. Cross-validation with only three to eight source domains is uncertain, so this is a practical selection rule and not a guarantee of the best target performance. The overall pattern matches the main B1–B6/B7–B10 result in the main text.

Why poolPCA performs better on B9. At $s = 8$, $k = 20, 30$, the methods are fitted on B1–B8 and evaluated on B9 and B10. poolPCA outperforms AnchorPCA$_\infty$ because of B9: its first principal component alone explains 92.43% of its variance. For the corresponding unit eigenvector v_1, its squared projection $100\|\Pi_W v_1\|_2^2$ onto the AnchorPCA$_\infty$ solution subspace is 66.00% and 71.57% at $k = 20$ and $k = 30$, respectively. The first PC of B9 thus does not seem to be in the invariant subspace $\mathcal{S}_\star$. In contrast, its squared projection onto the poolPCA solution subspace is 95.51% and 99.59%. Thus, poolPCA captures this dominant target direction more strongly. Still, AnchorPCA$_\infty$ has higher EV on B10 (Table 3). This does not contradict Thm. 4.1, whose robustness result applies to a specified class of structured covariance perturbations, not to every unseen-domain shift. Anchor PCA is not guaranteed to improve EV on every unseen domain, especially when important target directions lie outside $\mathcal{S}_\star$.

Explained variance of estimated $\mathcal{S}_\star$. The AnchorPCA$_\infty$ solution has rank k, but its first grouped eigenspace is a finite-sample estimate of the invariant subspace $\mathcal{S}_\star$ with dimension $\widehat{m}$ (Thm. C.11). For $s \in \{3, \dots, 8\}$ and $k \in \{10, 20, 30\}$, Fig. 10 compares the percentage of explained variance contributed by this estimated block with source-only poolPCA using the same dimension $\widehat{m}$ on B9 and B10. The comparison is evaluated separately on target batches B9 and B10. Across $s \in \{3, \dots, 8\}$, the estimated dimensions are $\widehat{m} = (3, 3, 3, 2, 2, 2)$ for $k = 10$, $\widehat{m} = (5, 5, 5, 5, 5, 5)$ for $k = 20$, and $\widehat{m} = (9, 8, 7, 7, 7, 6)$ for $k = 30$. On the target batch B9, the estimated block explains more variance than the same-dimensional poolPCA baseline for $k = 20$ and $k = 30$. On B10, both subspaces explain high variance and are often close.

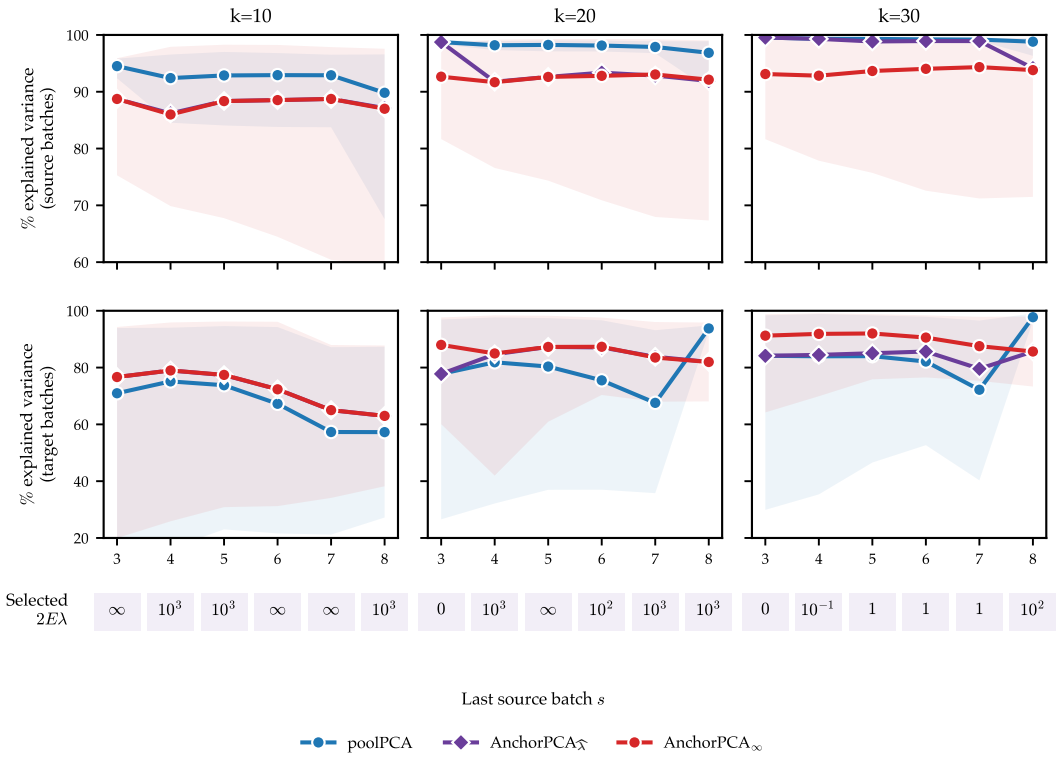

Figure 9: **Rolling source–target splits on gas-sensor drift data.** Each column fixes k; the horizontal axis is the last source batch s. Lines show mean EV over source batches (top row) or later target batches (middle row). The purple curve is $\texttt{AnchorPCA}_{\widehat{\lambda}}$, using § 5.1; the bottom strip gives its selected $2E\lambda$ for each split. Shaded bands retain the min–max ranges across batches for `poolPCA` and $\texttt{AnchorPCA}_{\infty}$.

D.4.2 Domain weighting and evaluation.

The experiments above give each source domain equal weight, regardless of sample size. Here we examine whether the results change when domains are weighted in proportion to sample size. We replace the equal weights $w_e = 1/E$ by $w_e = n_e / \sum_j n_j$ in both empirical averages, $\widehat{\Sigma} = \sum_e w_e \widehat{\Sigma}_e$ and $\widehat{\Pi} = \sum_e w_e \widehat{\Pi}_k^{(e)}$. The finite-penalty matrix is $\widehat{M}_\lambda = \sum_e w_e \widehat{\Sigma}_e + 2E\lambda \sum_e w_e \widehat{\Pi}_k^{(e)}$. At the population level, the corresponding reconstruction objective is

$$\sum_e w_e \operatorname{Tr}\big(\Sigma_e(I_p - \Pi_W)\big) + E\lambda \sum_e w_e \|\Pi_W - \Pi_k^{(e)}\|_{\mathrm{F}}^2.$$

This is also the worst-case reconstruction error over the original covariance class $\mathcal{C}_\lambda$ in Thm. 4.1, now weighted by sample-size.

Table 4 compares the two fitting schemes using the same gas-sensor preprocessing and EV metric as in § 6.2. For each split $s \in \{3, \dots, 8\}$, we fit on B1–Bs and evaluate on the later batches. We average target EV equally over later batches within each split, then equally over $s \in \{3, \dots, 8\}$. For both weighting schemes, $\texttt{AnchorPCA}_{\infty}$ has higher mean target EV than `poolPCA` at all three ranks, though neither weighting scheme is uniformly better than the other.

D.4.3 Gas-type classification.

As a downstream task, we classify the six gas types using the split in Fig. 4: we fit on batches B1–B6 (sources) and evaluate on B7–B10 (targets). Source labels are used for training and model selection, and target labels for evaluation.

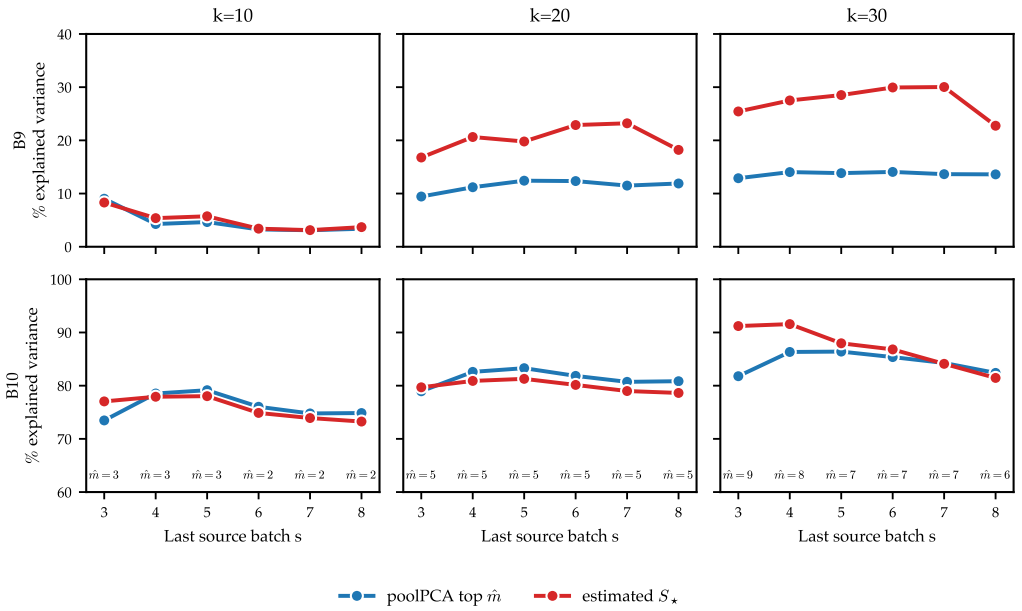

Figure 10: **Estimated invariant block vs. same-dimensional** `poolPCA`**.** For each (s, k), the red curve evaluates the first grouped eigenspace (estimated top eigenspace of $\overline{\Pi}$, equivalently an estimate of $\mathcal{S}_\star$) selected by $\texttt{AnchorPCA}_\infty$; the blue curve evaluates the top $\widehat{m}$ source-only `poolPCA` directions, where $\widehat{m}$ is the estimated dimension of $\mathcal{S}_\star$. Values are shown separately for target batches B9 and B10.

Table 4: The rolling experiment in Fig. 9 weights source batches equally when fitting; here we also fit both methods using weights proportional to batch sample sizes. Entries are mean target EV differences, $\texttt{AnchorPCA}_\infty$ minus `poolPCA`, in percentage points (values above zero imply that $\texttt{AnchorPCA}_\infty$ explains more variance than `poolPCA`).

k	Uniform-domain fitting	Sample-size-weighted fitting
10	5.29	5.07
20	6.03	11.19
30	5.78	4.50

Methods. We fit L2-regularized logistic regression on the 20-dimensional representations from `poolPCA`, $\texttt{AnchorPCA}_{\widehat{\lambda}}$, and $\texttt{AnchorPCA}_\infty$, and on all 128 unreduced features. This gives all four representations the same classifier family. Before fitting the representations, we rescale each feature using statistics from the pooled source observations, compute each batch's covariance around its own mean, and weight all source batches equally. We then standardize the classifier inputs using source observations only. As further baselines on the unreduced features, we fit L1-regularized logistic regression, linear GroupDRO [39], which minimizes the largest source-batch loss, and linear V-REx [24], which minimizes the mean source-batch loss plus a penalty on its variance across batches. So that large classes do not dominate the loss, logistic regression weights observations by inverse class frequency, and GroupDRO and V-REx use class-balanced cross-entropy within each batch. L2 logistic regression fits the classes jointly (multinomial) and L1 logistic regression fits each class against the rest.

Hyperparameter selection. We use leave-one-source-batch-out cross-validation. For each candidate, we refit the feature transformations, the representation (where applicable), and the classifier on five source batches, and score the held-out sixth. We repeat this for all six sources and select the candidate with the highest mean held-out balanced accuracy, weighting the six folds equally. Balanced accuracy averages the accuracy within each class present, so every class counts equally. Every logistic classifier selects the inverse penalty strength $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, where a smaller C means stronger regularization. $\texttt{AnchorPCA}_{\widehat{\lambda}}$ selects C jointly with $2E\lambda$ from the

Table 5: Gas classification: mean accuracy and balanced accuracy on the target batches, with each batch weighted equally. L1 and L2 denote the penalty of the logistic-regression classifier; the last two rows use linear GroupDRO and V-REx classifiers, both with L2 regularization.

Pipeline	Dim.	Selected values	Accuracy (%)	Balanced accuracy (%)
`poolPCA` + L2	20	$C = 10$	68.58	67.29
`AnchorPCA`$_{\widehat{\lambda}}$ + L2	20	$2E\lambda = 10$ $C = 0.1$	74.61	75.64
`AnchorPCA`$_{\infty}$ + L2	20	$C = 0.01$	**75.56**	**76.79**
Unreduced + L2	128	$C = 0.1$	75.05	75.08
Unreduced + L1	128	$C = 0.1$	74.44	76.16
Unreduced + GroupDRO	128	L2: 0.001	74.98	75.34
Unreduced + V-REx	128	$\beta = 1$ L2: 0.001	74.31	74.08

Table 6: Gas classification accuracy (%) by target batch, using the methods in Table 5.

Pipeline	B7	B8	B9	B10
`poolPCA` + L2	80.51	79.59	65.32	48.89
`AnchorPCA`$_{\widehat{\lambda}}$ + L2	80.68	88.10	65.96	63.69
`AnchorPCA`$_{\infty}$ + L2	79.19	80.95	82.98	59.14
Unreduced + L2	86.13	85.37	64.47	64.22
Unreduced + L1	89.21	70.41	71.28	66.89
Unreduced + GroupDRO	88.49	90.48	65.32	55.64
Unreduced + V-REx	86.35	89.12	65.11	56.67

grid in § D.4. GroupDRO selects its L2 coefficient from $\{10^{-4}, 10^{-3}, 10^{-2}\}$, where a larger value means stronger regularization. V-REx selects its L2 coefficient from the same grid jointly with the weight $\beta \in \{0, 0.1, 1, 10, 100, 1000\}$ on its loss-variance penalty, where $\beta = 0$ removes the penalty. We then refit each selected pipeline on all six sources.

Results. Table 5 averages accuracy and balanced accuracy over the four target batches, and Table 6 reports accuracy per batch. `AnchorPCA`$_{\widehat{\lambda}}$ is more accurate than `poolPCA` on all four target batches. `AnchorPCA`$_{\infty}$ has the highest mean accuracy and balanced accuracy of the seven pipelines. With only four target batches, we do not test these differences for significance. Further, unpenalized logistic regression performed worse on all four representations, so we omit it.

D.5 Fama–French industry portfolios

Summary of the experiment We evaluate Anchor PCA on a second real-world temporal study using financial-market data from the Fama–French 49 Industry Portfolios dataset [11]. Each observation represents one trading day with $p = 49$ features corresponding to the value-weighted daily returns of different industry portfolios. We use non-overlapping five-year periods as domains: 1970–1994 gives five source domains and 1995–2024 gives six target domains.

As the first two principal components capture roughly two-thirds of the variance in most source domains (Fig. 11), we first present results for $k = 2$. At $k = 2$, `AnchorPCA`$_{\widehat{\lambda}}$ chooses $\lambda = \infty$, i.e. `AnchorPCA`$_{\infty}$. Its mean target EV is 55.67%, compared with 51.80% for `poolPCA` and 52.79% for `norm-maxRegret` [17]. Relative to `poolPCA`, the mean gain is 3.87 percentage points, with gains of 1.94–5.55 points across all six future periods; it also improves over `norm-maxRegret` in all six periods. It also raises the worst-period EV from 40.18% (`poolPCA`) and 41.44% (`norm-maxRegret`) to 44.44%. Across the larger ranks $k \in \{3, \dots, 10\}$, `AnchorPCA`$_{\widehat{\lambda}}$ differs from `poolPCA` by at most 0.25 percentage points in absolute mean target EV.

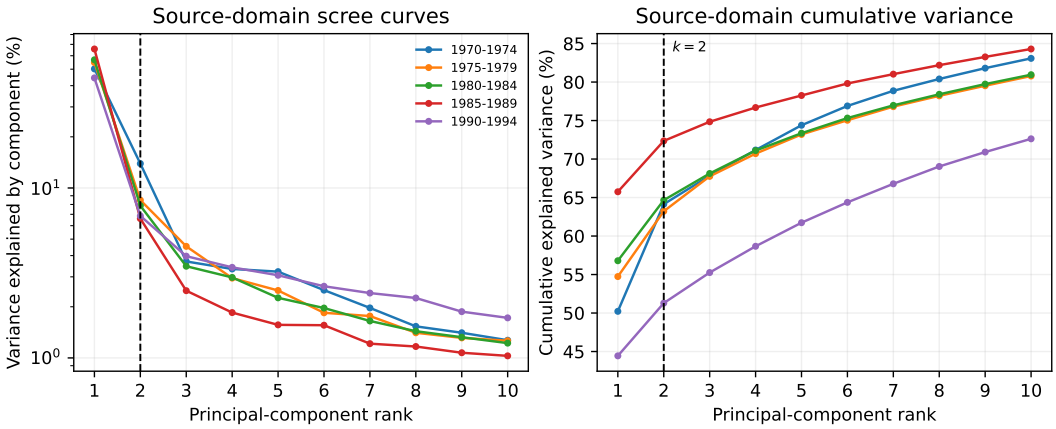

Figure 11: Fama–French source scree plots. Colors show the five-year source periods shown in the legend. Left: per-component EV (log scale); right: cumulative EV. Only source-period covariances enter these rank diagnostics.

Protocol. Following Kenneth French's Data Library, we treat entries of -99.99 and -999 as missing and drop days with any missing return. This removes one day in 1970–1974 and none elsewhere. We center returns within each period and compute their sample covariance without feature standardization. Source periods receive equal fitting weights. For each fixed $k \in \{2, \dots, 10\}$, we select $2E\lambda$ by § 5.1 with the grid of § D.4, and then refit on all five sources. At $k = 2$, we also fit `norm-maxRegret`.

Rank choice and results. All five source scree plots show two dominant components followed by a flatter tail (Fig. 11); the second eigenvalue is 1.73–3.77 times the third. This supports examining $k = 2$. Table 7 reports target EV for each period at $k = 2$, and Table 8 gives results for $k = 3, \dots, 10$. For $k = 3$ to 6, cross-validation selects $2E\lambda = 0$, which recovers `poolPCA`.

Table 7: Fama–French target EV (%) at $k = 2$. Source-only cross-validation selects $\texttt{AnchorPCA}_\infty$. The mean gives each target period equal weight; the minimum is across periods.

Target period	Days	`poolPCA`	`norm-maxRegret`	$\texttt{AnchorPCA}_\infty$
1995–1999	1263	40.18	41.44	45.73
2000–2004	1256	41.63	42.15	44.44
2005–2009	1259	61.31	62.33	65.15
2010–2014	1258	64.72	65.91	68.75
2015–2019	1258	48.07	49.42	53.14
2020–2024	1258	54.88	55.50	56.82
Mean	–	51.80	52.79	55.67
Minimum	–	40.18	41.44	44.44

Table 8: Fama–French rank sensitivity. EV is averaged equally over the six target periods. The final column gives $\texttt{AnchorPCA}_{\widehat{\lambda}}$ minus poolPCA in percentage points.

k	Selected $2E\lambda$	poolPCA EV (%)	$\texttt{AnchorPCA}_{\widehat{\lambda}}$ EV (%)	Difference (points)
2	∞	51.80	55.67	+3.873
3	0	57.12	57.12	0.000
4	0	62.20	62.20	0.000
5	0	64.69	64.69	0.000
6	0	66.04	66.04	0.000
7	1	67.94	68.19	+0.248
8	1	69.79	69.89	+0.101
9	1	70.97	71.06	+0.095
10	0.01	72.07	72.07	−0.003

E Nonlinear extensions

In this appendix, we discuss nonlinear extensions of our approach to finding rank-k robust or invariant subspaces via Anchor PCA translated to autoencoder-based approaches.

Autoencoders (AEs) are a popular approach for nonlinear dimension reduction [20, 38]. As mentioned in § 1.1, the projection matrix $\Pi_W = WW^\top$ can be interpreted as a linear autoencoder which first projects a given x to a k-dimensional embedding via $W^\top \in \mathbb{R}^{k\times p}$ and then reconstructs x by mapping back to the original p-dimensional space via $W \in \mathbb{R}^{p\times k}$. An autoencoder functions analogously, but with possibly nonlinear encoder and decoder, typically parameterized as (deep) neural networks and optimized via stochastic gradient descent [35].

E.1 Vanilla autoencoder (AE)

An autoencoder (f, g) consists of an encoder $f : \mathbb{R}^p \to \mathbb{R}^k$, which produces k-dimensional embeddings $z = f(x)$, and a decoder $g : \mathbb{R}^k \to \mathbb{R}^p$, which produces reconstructions $\hat{x} = g(z) = g(f(x))$. A standard autoencoder is trained to minimize the mean squared reconstruction error

$$\mathcal{L}_{\text{AE}}(f, g; \mathbb{P}) = \mathbb{E}_{x\sim\mathbb{P}}\left[\|x - g(f(x))\|_2^2\right],$$

where, in practice, $\mathbb{P}$ is replaced by its empirical distribution.

Given data from multiple domains, one approach to learning a shared nonlinear k-dimensional embedding is to train an autoencoder on the pooled data, similar to poolPCA in the linear setting. We refer to this approach as poolAE. However, like poolPCA, this approach is susceptible to focusing on spurious directions that exhibit high variation in only a few domains.

E.2 Anchor autoencoder (AnchorAE)

We now discuss a method we call AnchorAE as a more robust alternative. Let $\{(f_e, g_e)\}_{e=1}^E$ be domain-specific AEs, i.e.,

$$(f_e, g_e) \in \arg\min_{f,g} \mathcal{L}_{\text{AE}}(f, g; \mathbb{P}_e).$$

In analogy to Eq. (2), we define $\texttt{AnchorAE}_\lambda$ as the following unconstrained optimization problem

$$\min_{f,g} \mathbb{E}_{x\sim\bar{\mathbb{P}}}\left[\|x - g(f(x))\|_2^2\right] + \lambda \sum_{e=1}^{E} \mathcal{R}(g \circ f, g_e \circ f_e) \tag{24}$$

where $\bar{\mathbb{P}} = \frac{1}{E}\sum_{e=1}^E \mathbb{P}_e$ denotes the (empirical) pooled distribution and $\mathcal{R}$ is a suitable invariance penalty between the shared and domain-specific AEs (discussed in more detail below).

The first term in Eq. (24) is the pooled reconstruction error, which can be viewed as the AE-equivalent or proxy of explained variance in PCA. The second term is meant to capture disagreement among the

nonlinear projections produced by the shared $g \circ f$ and domain-specific $g_e \circ f_e$ AEs, similar to the disagreement term

$$\left\| \Pi_W - \Pi_k^{(e)} \right\|_{\mathrm{F}}^2$$

between projection matrices in $\texttt{AnchorPCA}_\lambda$ in Eq. (2).

On the choice of invariance penalty. Whereas in the linear case, it is possible to compare the projection matrices directly, it is less clear how to assess agreement between nonlinear functions, as these are infinite-dimensional objects whose global behavior is not determined by their local behavior. In principle, we can consider measuring invariance w.r.t. domain-specific nonlinear projections through suitable norms $\|(g \circ f) - (g_e \circ f_e)\|$ in function space. However, each AE is unconstrained outside the support of the (empirical) distribution used to train it, i.e., $\mathbb{P}_e$ for (f_e, g_e) and $\bar{\mathbb{P}}$ for (f, g). Hence, we should evaluate disagreement only on the respective shared supports. We therefore propose the following invariance penalty

$$\mathcal{R}\left(g \circ f, g_e \circ f_e\right) = \mathbb{E}_{x \sim \mathbb{P}_e}\left[\left\| g(f(x)) - g_e(f_e(x)) \right\|_2^2\right],$$

which can be viewed as the regret incurred by using the shared instead of the domain-specific AE for reconstructing data from domain e [2, 16].

Limitations. `AnchorAE` requires training $E + 1$ separate AEs, since to the best of our knowledge — and unlike in the linear case—the problem cannot be reduced to a modified single learning objective for (f, g). This approach may thus be computationally expensive, especially when the number of domains is large. In the following subsections, we discuss an approach for training a single, distributional AE from multi-domain data.

E.3 Distributional principal autoencoder (DPA)

The distributional principal autoencoder (DPA) [43] is a nonlinear dimension reduction approach, which can be interpreted as a nonlinear variant of PCA that, in contrast to standard autoencoders, emphasizes distributional rather than mean reconstruction. We first review DPA in the single-domain setting and then discuss a possible extension for our multi-domain setting.

Given a number $k \leq p$ of components, DPA consists of an encoder $f : \mathbb{R}^p \to \mathbb{R}^k$, which produces embeddings $z = f(x)$, and a stochastic decoder or generator $g : \mathbb{R}^k \times \mathbb{R}^{p-k} \to \mathbb{R}^p$, which, in addition to z, takes $(p-k)$-dimensional noise $\epsilon \sim \mathcal{N}(0, I)$ as input and allows for sampling reconstructions $\hat{x} = g(z, \epsilon)$. The DPA objective is constructed such that, for a fixed encoder f, the optimal DPA decoder g^* maps a given embedding z to the distribution of X given $f(X) = z$,

$$g^*(z, \epsilon; f) \overset{d}{=} (X | f(X) = z), \tag{25}$$

where $\overset{d}{=}$ denotes equality in distribution. In other words, the optimal decoder evaluated at z matches the distribution of realizations of X that are mapped by the encoder to z. At the same time, the DPA encoder minimizes the variability in the distributions in (25) by encoding the first k 'principal' components. Both of these goals are achieved by minimizing the following DPA objective

$$\begin{aligned}
&\mathcal{L}_{\mathrm{DPA}}^{\beta}(f, g; \mathbb{P}) \\
&\quad = \mathbb{E}_{X \sim \mathbb{P}, \epsilon \sim \mathcal{N}(0,I)} \left\| X - g(f(X), \epsilon) \right\|_2^{\beta} - \frac{1}{2} \mathbb{E}_{X \sim \mathbb{P}, \epsilon, \epsilon' \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,I)} \left\| g(f(X), \epsilon) - g(f(X), \epsilon') \right\|_2^{\beta},
\end{aligned}$$

where $\beta \in (0, 2)$. This objective corresponds to the negative expected energy score [18] between X and the corresponding stochastic decoder output, conditional on the encoding $f(X)$ of X. Due to this conditioning, the DPA objective differs from an energy distance by a normalization constant which depends on the encoder and encourages capturing principal (i.e., variation-minimizing) components.

E.4 Multi-domain DPA

One could extend the DPA approach to our task of learning invariant subspaces from multi-domain data as follows. To map data from different distributions $\{\mathbb{P}_e\}_{e \in \mathcal{E}}$ to an invariant subspace, we consider a single shared encoder $f : \mathbb{R}^p \to \mathbb{R}^k$, whereas to allow for accurate distributional reconstruction across domains, we consider domain-specific decoders $\left\{g_e : \mathbb{R}^k \times \mathbb{R}^{p-k} \to \mathbb{R}^p\right\}_{e \in \mathcal{E}}$.

We could now train E domain-specific DPAs. However, in contrast to $\texttt{AnchorPCA}_\lambda$ and AnchorAE, it is unclear how to then compare or penalize disagreement between a shared DPA and domain-specific ones since, by construction, each DPA can fully reconstruct any distribution if given sufficient capacity in the form of additional $p-k$-dimensional noise as input to the stochastic decoder. Moreover, as discussed above, training a separate model for each domain is computationally demanding.

This motivates considering the following objective for training a single multi-domain DPA:

$$\mathcal{L}^{\beta}_{\text{MD-DPA}}\left(f, \{g_e\}_{e\in\mathcal{E}}; \{\mathbb{P}_e\}_{e\in\mathcal{E}}\right) := \sum_{e\in\mathcal{E}} \mathcal{L}^{\beta}_{\text{DPA}}(f, g_e; \mathbb{P}_e). \tag{26}$$

Since domain-specific decoders are used, this is different from a distributional, nonlinear version of poolPCA. Intuitively, the additional reconstruction capacity from domain-specific decoders and noise may enable the encoder to focus on shared directions, rather than explaining pooled variance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by a theoretical study of the proposed methods in §§ 3 to 5 and the corresponding appendices as well as by synthetic and real-world experiments in § 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the presented work are discussed in the discussion and future-work paragraph (§ 7).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in the corresponding main text sections or (if necessary) in more explicit form in § C. Complete proofs are given in § C, with additional supporting results in § B.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experimental protocols are described in §§ D and 6, including synthetic data generation, sampling grids, aggregation rules, and the gas-sensor and Fama–French temporal splits. The supplementary code archive contains experiment scripts, seeds, dependencies, and reproduction commands.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary code archive, also available at https://github.com/benesei/anchor-pca, provides the package, experiment scripts, dependencies, and reproduction commands. Synthetic data are generated from fixed seeds, and the scripts download the public gas-sensor and Fama–French datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The finite-sample implementation is specified in § 5.1, and experimental details are given in §§ D.3 to D.5. The READMEs in the supplementary code archive provide reproduction commands and protocol details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The random-subspace figures define medians/means and 10–90% quantile bands across population-design summaries in their captions and in § D.3. The gas-sensor rolling plot explicitly reports min–max ranges over batches rather than confidence intervals.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The `experiments/README.md` in the supplementary code archive reports the CPU and RAM used, states that no GPU was used, and gives approximate wall-clock times for the main experiment groups and the full reproduction run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses synthetic data and public gas-sensor and financial datasets without human-subject research, respects documented licenses and terms of use, and provides reproduction code in the supplementary code archive.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work is methodological research on covariance-based dimension reduction and does not introduce a deployed decision system or high-risk application. The experiments use synthetic data and public gas-sensor and financial datasets.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not pose such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The gas-sensor and Fama–French datasets are credited in §§ D.4 and D.5. The supplementary code archive documents their licenses or terms of use and the external `minPCA` dependency in `ASSETS.md`, and includes our MIT license and pinned dependency versions.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The supplementary code archive contains the Anchor PCA implementation, documented by the top-level `README.md`, experiment READMEs, and tests. Generated synthetic designs are reproducible outputs of the scripts and are not presented as a separately released dataset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. The empirical work uses synthetic data and public gas-sensor and financial datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core present research does not involve LLMs as an important, non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.