PROVABLY IDENTIFYING LINEAR CONSTRAINTS UNDER UNKNOWN AND SINGULAR MEASUREMENT NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Many scientific and engineering systems satisfy unknown linear relations, but only noisy measurements of their variables are available. Four issues make these relations challenging. Noise variances can differ across variables, measurement errors can be correlated, and the noise covariance matrix can be singular. Moreover, both the noise covariance matrix and the number of relations may be unknown. We address all four challenges by proposing a novel generalized eigenvalue decomposition (GenEVD) based algorithm that avoids inverting the noise covariance matrix unlike existing methods. Iterative PCA and ordinary least squares arise as special cases of our method. Our first algorithm alternates this spectral update with an identifiable supported-moment estimator of the noise covariance. Our second algorithm estimates the number of relations through a descending unity-eigenvalue test. We establish high-probability recovery of the constraint subspace, noise covariance matrix, and number of relations. We also prove local geometric convergence and obtain a subspace rate with minimax-optimal dependence on sample size and dimension. Controlled synthetic experiments and empirical study on seven real-world datasets validate the corresponding theoretical claims.

1 INTRODUCTION

Linear relations arise in many scientific and engineered systems. Electrical, traffic, and distribution flows obey conservation laws. Steady-state reaction fluxes obey stoichiometric balances. Let $\mathbf{x}^{(t)} \in \mathbb{R}^d$ denote the latent vector for sample t. We assume that all samples satisfy the same m independent linear relations $\mathbf{A}\mathbf{x}^{(t)} = \mathbf{0}$, where $\mathbf{A} \in \mathbb{R}^{m\times d}$ is unknown. Given n i.i.d. noisy measurements $\mathbf{y}^{(t)} = \mathbf{x}^{(t)} + \boldsymbol{\epsilon}^{(t)}$, our goal is to estimate the constraint rowspace, the number of linear relations m, and the measurement-noise covariance $\boldsymbol{\Sigma}_\epsilon = \mathrm{Cov}(\boldsymbol{\epsilon}^{(t)})$. This problem has the following four challenging aspects.

Challenge 1: Heterogeneity. Measurement noise is heterogeneous when its variance changes across variables because sensors differ in precision, dynamic range, and calibration quality. As Table 1 shows, standard PCA does not account for this heterogeneity and can mistake high-variance noise for signal. Existing noise-aware methods address heterogeneity, but impose additional restrictions discussed below (Wentzell et al., 1997; Narasimhan & Shah, 2008; Zhang et al., 2022; Zhou & Chen, 2025).

Challenge 2: Non-Diagonal $\boldsymbol{\Sigma}_\epsilon$. In addition to being heterogeneous, measurement errors need not be independent. Sensors that share a reference, calibration unit, electronics, or environment can have correlated errors, making $\boldsymbol{\Sigma}_\epsilon$ non-diagonal. Zhang et al. (2022); Zhou & Chen (2025) assume diagonal covariance matrix and therefore cannot represent these correlations.

Challenge 3: Singularity. The problem becomes more difficult when $\boldsymbol{\Sigma}_\epsilon$ is singular. For instance, a source injection supplied as a known control input may be treated as exact and have zero measurement-error variance. Exact reference or derived variables have the same effect. When $\boldsymbol{\Sigma}_\epsilon \succ 0$, Wentzell et al. (1997); Narasimhan & Shah (2008) scale the data by $\boldsymbol{\Sigma}_\epsilon^{-1/2}$, but this inverse does not exist when $\boldsymbol{\Sigma}_\epsilon$ is singular.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison under $\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}$ and $\mathbf{A}\mathbf{x} = \mathbf{0}$, where m is the number of independent linear relations and $\boldsymbol{\Sigma}_\epsilon = \mathrm{Cov}(\boldsymbol{\epsilon})$. A check under **Allows** means that the method permits heterogeneous noise (coordinate-dependent variances), non-diagonal noise (correlated errors), or singular $\boldsymbol{\Sigma}_\epsilon$, respectively. Under **Estimates**, the $\boldsymbol{\Sigma}_\epsilon$ and m columns indicate whether the method returns an estimate rather than always requiring that quantity as an input. In our experiments, HeteroPCA and Deflated-HeteroPCA return diagonal residual covariance estimates. Their theory targets signal-subspace recovery under diagonal noise. IPCA estimates m through a candidate-order heuristic for nonsingular covariance iterates. **Statistical Guarantees** indicates a finite-sample recovery guarantee under the method's stated assumptions.

	Allows			**Estimates**		**Statistical**
Method	**Hetero.**	**Non-diag.**	**Singular**	$\boldsymbol{\Sigma}_\epsilon$	m	**Guarantees**
Standard PCA	✗	✗	✗	✗	✗	✓
Factor Analysis (EM)	✓	✗	✗	✓	✗	✗
MLPCA (Wentzell et al., 1997)	✓	✓	✗	✗	✗	✗
IPCA (Narasimhan & Shah, 2008)	✓	✓	✗	✓	✓	✗
HeteroPCA (Zhang et al., 2022)	✓	✗	✓	✓	✗	✓
Deflated-HeteroPCA (Zhou & Chen, 2025)	✓	✗	✓	✓	✗	✓
GenEVD (ours)	✓	✓	✓	✓	✓	✓

Challenge 4: Unknown Number of Linear Relations and Noise Covariance. Finally, both m and $\boldsymbol{\Sigma}_\epsilon$ may be unknown. The number of independent linear relations can change with topology, switch status, or active components. Most methods in Table 1 require m. IPCA instead proposes an outer sweep over candidate values and checks whether the converged number of unity singular values equals the candidate order (Narasimhan & Shah, 2008). This useful heuristic is restricted to nonsingular noise-covariance iterates and is not accompanied by a finite-sample order- recovery guarantee. Estimating $\boldsymbol{\Sigma}_\epsilon$ is a separate difficulty. MLPCA assumes it known. IPCA estimates it when its covariance iterates are nonsingular. The HeteroPCA variants yield diagonal residual estimates, but cannot represent correlated measurement errors (Wentzell et al., 1997; Narasimhan & Shah, 2008; Zhang et al., 2022; Zhou & Chen, 2025). Our goal is to estimate a possibly non-diagonal and singular $\boldsymbol{\Sigma}_\epsilon$ jointly with m, with finite-sample guarantees for both tasks.

Together, these challenges motivate the capabilities summarized in Table 1.

1.1 Our Approach Handles Unknown and Singular Noise

Handling Heterogeneous, Non-Diagonal, and Singular Noise Covariance. GenEVD solves $\boldsymbol{\Sigma}_y\mathbf{v} = \lambda\boldsymbol{\Sigma}_\epsilon\mathbf{v}$ without forming $\boldsymbol{\Sigma}_\epsilon^{-1}$. It therefore handles all three noise properties in Table 1, including singular $\boldsymbol{\Sigma}_\epsilon$, through the QZ algorithm (Moler & Stewart, 1973).

Estimating $\boldsymbol{\Sigma}_\epsilon$ and m. GenEVD alternates with a supported-moment update for the q free entries of $\boldsymbol{\Sigma}_\epsilon$, while a descending unity-eigenvalue sweep estimates m. Identifiability requires $q \le m(m+1)/2$ and a full-rank moment design. This structure is essential when $d > n$ and the sample covariance is rank deficient.

Statistical Guarantees. We guarantee recovery of the constraint rowspace, $\boldsymbol{\Sigma}_\epsilon$, and m once $n \gtrsim m^2(d + \log(1/\delta))$, up to gap and conditioning factors. The alternating iterates converge geometrically, and the subspace dependence on n and d is minimax optimal.

1.2 Contributions and Guarantees

Table 1 places our results in context.

1. **One formulation for all four challenges.** GenEVD handles heterogeneous, correlated, and singular noise while allowing both $\boldsymbol{\Sigma}_\epsilon$ and m to be unknown. IPCA is its positive-definite special case, and OLS is recovered when only the response coordinate is noisy.

2. **Our algorithmic contribution.** The first algorithm jointly estimates the constraint rowspace and supported $\mathbf{\Sigma}_\epsilon$. The second estimates m by a descending spectral sweep. Together they estimate $(m, \mathbf{A}, \mathbf{\Sigma}_\epsilon)$ with a local geometric-convergence guarantee.
3. **Our statistical contribution.** We give finite-sample oracle, order-recovery, and end-to-end guarantees. The subspace rate matches our minimax lower bound in its dependence on sample size and dimension under bounded signal-to-noise ratio.
4. **Empirical validation.** GenEVD improves constraint and noise-covariance recovery under correlated and singular noise. The misspecification experiments confirm the predicted unity-eigenvalue diagnostics for the covariance support and number of relations. GenEVD also outperforms the applicable baselines in semi-synthetic experiments constructed from seven real-world datasets. Together, these experiments validate our theoretical claims for constraint, noise-covariance, and relation-number recovery.

Notation. Bold lowercase letters are vectors, bold uppercase letters matrices, and calligraphic letters subspaces, supports and model classes. A superscript (t) indexes samples and (k) indexes iterations. Hats denote sample quantities. For a positive integer p, $[p] := \{1, \dots, p\}$, and $\mathbf{e}_j$ is the j-th standard basis vector. For a matrix $\mathbf{B}$, $\mathbf{B}^\mathsf{T}$ is its transpose, $\mathbf{B}^+$ its Moore–Penrose pseudoinverse, $\mathrm{row}(\mathbf{B})$ its row space, $\mathrm{Null}(\mathbf{B})$ its null space and $\mathrm{rank}(\mathbf{B})$ its rank. We write $\|\cdot\|$, $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ for the Euclidean, spectral and Frobenius norms. If $\mathbf{U}$ has orthonormal columns spanning $\mathcal{U}$, then $\mathbf{P}_\mathcal{U} := \mathbf{U}\mathbf{U}^\mathsf{T}$ is the orthogonal projector onto $\mathcal{U}$. We write $\mathbb{S}^m := \{\mathbf{M} \in \mathbb{R}^{m\times m} : \mathbf{M} = \mathbf{M}^\mathsf{T}\}$ with its Frobenius inner product, $\mathcal{G}_{m,d}$ for the Grassmannian of m-dimensional subspaces of $\mathbb{R}^d$, and $\kappa(\mathbf{B}) := \lambda_{\max}(\mathbf{B})/\lambda_{\min}(\mathbf{B})$ for $\mathbf{B} \succ 0$. Finally, $\lesssim$, $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ and $\mathcal{O}_\mathbb{P}(\cdot)$ hide *absolute* constants only, never quantities that depend on the model. A full symbol table is given in Appendix A.

2 Problem Formulation

We first state the observation model and the assumptions used throughout the analysis. We then show that GenEVD contains iterative PCA (IPCA) and ordinary least squares (OLS) as exact special cases; these connections motivate using one generalized-eigenvalue formulation in the singular setting.

Observation model. We observe n i.i.d. samples $\mathbf{y}^{(t)} = \mathbf{x}^{(t)} + \boldsymbol{\epsilon}^{(t)} \in \mathbb{R}^d$, where $\mathbf{x}^{(t)}$ and $\boldsymbol{\epsilon}^{(t)}$ are zero mean, mutually uncorrelated, and i.i.d. across t. The latent signal satisfies $\mathbf{A}\mathbf{x}^{(t)} = \mathbf{0}$ for a full-row-rank $\mathbf{A} \in \mathbb{R}^{m\times d}$, and $\mathrm{Cov}(\boldsymbol{\epsilon}^{(t)}) = \mathbf{\Sigma}_\epsilon \succeq 0$. Only $\mathrm{row}(\mathbf{A})$ is determined by the model, so we fix an arbitrary representative with orthonormal rows, $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}_m$; this leaves the residual ambiguity $\mathbf{A} \mapsto \mathbf{R}\mathbf{A}$ for $\mathbf{R} \in \mathrm{O}(m)$, which is harmless because every bound below is stated for the projector $\mathbf{P}_\mathcal{C}$. We write $r := d - m \geq 1$, $\mathbf{\Sigma}_x := \mathrm{Cov}(\mathbf{x}^{(t)})$ with $\mathrm{rank}(\mathbf{\Sigma}_x) = r$, $\mathbf{\Sigma}_y := \mathbf{\Sigma}_x + \mathbf{\Sigma}_\epsilon$, and

$$\mathcal{C} := \mathrm{row}(\mathbf{A}) = \mathrm{Null}(\mathbf{\Sigma}_x), \qquad \mathcal{N} := \mathrm{Null}(\mathbf{\Sigma}_\epsilon), \qquad s := \dim(\mathcal{N}).$$

The sample covariance is $\mathbf{S}_y := \frac{1}{n}\sum_{t=1}^n \mathbf{y}^{(t)}\mathbf{y}^{(t)\mathsf{T}} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^\mathsf{T}$.

Problem of Interest: Given n samples i.i.d. samples $\mathbf{y}^{(t)}$ and support (indices of non-zero entries) of noise covariance matrix $\mathbf{\Sigma}_\epsilon$, we want to estimate m, $\mathbf{A}$, and $\mathbf{\Sigma}_\epsilon$ jointly.

Assumption 1 (Sub-Gaussian observations)**.** There is $K < \infty$ with $\big\|\langle \mathbf{y}^{(t)}, \mathbf{u}\rangle\big\|_{\psi_2} \leq K\sqrt{\mathbf{u}^\mathsf{T}\mathbf{\Sigma}_y\mathbf{u}}$ for every $\mathbf{u} \in \mathbb{R}^d$.

This holds with an absolute K when $\mathbf{y}^{(t)}$ is Gaussian and, more generally, whenever $\mathbf{x}^{(t)}$ and $\boldsymbol{\epsilon}^{(t)}$ are sub-Gaussian. We state it relative to $\mathbf{\Sigma}_y$ rather than after scaling by $\mathbf{\Sigma}_\epsilon^{-1/2}$ because that scaling is unavailable when $\mathbf{\Sigma}_\epsilon$ is singular, and the same assumption then serves both regimes. When $\mathbf{\Sigma}_\epsilon \succ 0$ it also implies the scaled form used in Appendix I.

Assumption 2 (Well-posedness)**.** $\mathbf{\Sigma}_y \succ 0$: every direction carries signal or noise.

Assumption 3 (Intrinsic signal gap)**.** The *intrinsic signal gap* $\sigma_\mathcal{C}^2 := \lambda_{\min}(\mathbf{\Sigma}_x|_{\mathcal{C}^\perp})$ is positive.

Positivity of $\sigma_\mathcal{C}^2$ is not an extra hypothesis: $\mathrm{rank}(\mathbf{\Sigma}_x) = r$ already forces the r nonzero eigenvalues of $\mathbf{\Sigma}_x$ to be positive.

Assumption 4 (Spectral noise cap for estimating m)**.** There is a known $\bar{\sigma}^2 < \sigma_{\mathcal{C}}^2$ such that $\mathbf{\Sigma}_\epsilon \preceq \bar{\sigma}^2 \mathbf{I}$.

Assumption 4 is needed only for estimating m. It is an SNR-type separation: admissible measurement noise must remain below the weakest nonzero signal direction, preventing an overspecified model from absorbing that direction into its noise covariance.

Identifiability. The subspace $\mathcal{C}$ is identifiable from $(\mathbf{\Sigma}_y, \mathbf{\Sigma}_\epsilon)$ because $\mathcal{C} = \mathrm{Null}(\mathbf{\Sigma}_y - \mathbf{\Sigma}_\epsilon)$. To recover $\mathbf{\Sigma}_\epsilon$ as well, let $\mathcal{S}$ be a symmetric support, a set of unordered index pairs with the convention that $(i, j) \in \mathcal{S}$ frees both $[\mathbf{\Sigma}_\epsilon]_{ij}$ and $[\mathbf{\Sigma}_\epsilon]_{ji}$, so that $q := |\mathcal{S}|$ counts free symmetric parameters. Collecting them into $\boldsymbol{\theta} \in \mathbb{R}^q$ and writing $\mathrm{mat}_{\mathcal{S}}(\boldsymbol{\theta})$ for the matrix they build, the constraint $\mathbf{A}\mathbf{x}^{(t)} = \mathbf{0}$ gives $\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^\mathsf{T} = \mathbf{0}$ and hence the moment equation $\mathbf{A}\mathbf{\Sigma}_y\mathbf{A}^\mathsf{T} = \mathbf{A}\mathbf{\Sigma}_\epsilon\mathbf{A}^\mathsf{T}$. This motivates the *moment map*

$$\Phi_{\mathbf{A},\mathcal{S}} : \mathbb{R}^q \to \mathbb{S}^m, \qquad \Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta}) := \mathbf{A}\,\mathrm{mat}_{\mathcal{S}}(\boldsymbol{\theta})\,\mathbf{A}^\mathsf{T}, \qquad \kappa_\Phi := 1/\sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}}), \tag{1}$$

which any estimator built from that equation must invert. Since the equation supplies $\dim(\mathbb{S}^m) = m(m+1)/2$ scalar constraints, the counting condition $q \le m(m+1)/2$ is necessary; Theorem 4 makes this precise. We also write

$$\mathcal{K} := \left\{\mathbf{M} \in \mathbb{S}^d : \mathbf{M} \succeq 0,\ \mathrm{supp}(\mathbf{M}) \subseteq \mathcal{S}\right\} \tag{2}$$

for the set of admissible error covariances and $\Pi_{\mathcal{K}}$ for the Euclidean projection onto it, which is well defined because $\mathcal{K}$ is closed and convex.

From IPCA to the generalized eigenvalue formulation. Narasimhan & Shah (2008) proposed IPCA, which alternates between scaling the data by the inverse Cholesky factor of the current estimate $\mathbf{\Sigma}_\epsilon^{(k)}$ and extracting the bottom-m eigenspace, and updating $\mathbf{\Sigma}_\epsilon$ by maximum likelihood on the constraint residuals; Appendix B records the algorithm. The following lemma shows that its scaling step is a special case of our formulation. Narasimhan & Shah (2008) call this step scaling. The scaled observations $\mathbf{L}^{-1}\mathbf{y}^{(t)}$ have covariance $\mathbf{M} = \mathbf{I} + \mathbf{L}^{-1}\mathbf{\Sigma}_x\mathbf{L}^{-\mathsf{T}}$, so only their noise component, not the data, has identity covariance.

Lemma 1 (IPCA as a special case of GenEVD)**.** *Let $\mathbf{\Sigma}_\epsilon \succ 0$ with Cholesky factor $\mathbf{\Sigma}_\epsilon = \mathbf{L}\mathbf{L}^\mathsf{T}$. The scaled eigenvalue problem $\mathbf{M}\mathbf{w} = \mu\mathbf{w}$, where $\mathbf{M} = \mathbf{L}^{-1}\mathbf{\Sigma}_y\mathbf{L}^{-\mathsf{T}}$, is equivalent to $\mathbf{\Sigma}_y\mathbf{v} = \lambda\mathbf{\Sigma}_\epsilon\mathbf{v}$ under $\mathbf{w} = \mathbf{L}^\mathsf{T}\mathbf{v}$ and $\mu = \lambda$.*

The proof is given in Appendix C.

Lemma 1 shows that GenEVD preserves the established IPCA estimator whenever the inverse-based formulation is valid. Its role is therefore not merely algebraic: it verifies backward compatibility while identifying the generalized pencil as the extension that remains available when $\mathbf{\Sigma}_\epsilon$ loses rank. The proof is the invertible change of coordinates $\mathbf{w} = \mathbf{L}^\mathsf{T}\mathbf{v}$.

Ordinary least squares (OLS) minimizes squared response residuals when the regressors are treated as exact. The next result shows that this familiar regression problem is another boundary case of the same generalized pencil.

Theorem 2 (Informal: GenEVD recovers OLS)**.** *When the regressors have full column rank and only the response is noisy, GenEVD has a unique finite generalized eigenvector and recovers exactly the OLS coefficient vector.*

Theorem 2 connects the constraint formulation to regression: the generalized eigenvector encodes the OLS coefficient vector when only the response is noisy. More broadly, total least squares (TLS) permits error in every variable, while generalized TLS (GTLS) incorporates a structured error covariance (Fuller, 1987; Markovsky & Van Huffel, 2007). GenEVD lies in this errors-in-variables family but selects the finite eigenvalues closest to one, which is the spectral signature of the linear constraints in our model. The proof eliminates the regressor block of the pencil; its Schur complement is exactly the OLS residual sum of squares.

The generalized formulation never forms $\mathbf{L}^{-1}$ and stays well defined when $\mathbf{\Sigma}_\epsilon$ is singular, where it is solved by the QZ algorithm (Moler & Stewart, 1973). Section 3 presents the resulting algorithm.

When $\mathbf{\Sigma}_\epsilon \succ 0$, Lemma 1 also yields the standard scaled subspace, order, and noise-covariance guarantees; their formal statements and proofs are entirely in Appendix I. Thus Section 2 supplies

both the statistical model and the classical special cases; we now develop the estimator and guarantees for the general case $\mathbf{\Sigma}_\epsilon \succeq 0$.

3 Recovery with Singular Noise Covariance

GenEVD applies to every $\mathbf{\Sigma}_\epsilon \succeq 0$ because it never forms $\mathbf{\Sigma}_\epsilon^{-1}$. It alternates a QZ-based generalized-eigenspace step with the supported moment update of Eq. (1); when $\mathbf{\Sigma}_\epsilon \succ 0$, Lemma 1 recovers IPCA's constraint step. Algorithm 1 states the estimator analyzed below; Appendices F–K give the formal statements and proofs.

The analysis follows the tasks posed in the introduction. We first recover the constraint rowspace when $\mathbf{\Sigma}_\epsilon$ is supplied, then characterize and estimate an unknown supported $\mathbf{\Sigma}_\epsilon$, and finally estimate m. A minimax lower bound establishes statistical optimality, after which the end-to-end theorem combines the two oracle steps. A *local basin* is a neighborhood of the true $\mathbf{\Sigma}_\epsilon$ in which one alternating update is contractive; a basin initialization means that $\mathbf{\Sigma}_\epsilon^{(1)}$ lies in this neighborhood. The appendix gives its explicit radius ρ.

Algorithm 1 GenEVD: subspace recovery via generalized eigenvalue decomposition

Input: Data $\mathbf{Y} \in \mathbb{R}^{d\times n}$, number of linear relations m, support $\mathcal{S}$ with $q \le m(m+1)/2$, iteration budget T

Output: Constraint matrix $\mathbf{A}^{(T)}$ with orthonormal rows, covariance $\mathbf{\Sigma}_\epsilon^{(T+1)} \succeq 0$ supported on $\mathcal{S}$

1: $\mathbf{S}_y \leftarrow \frac{1}{n}\mathbf{Y}\mathbf{Y}^\mathsf{T}$; initialize $\mathbf{\Sigma}_\epsilon^{(1)} \succeq 0$ on $\mathcal{S}$
2: **for** $k = 1, 2, \ldots, T$ **do**
3: Solve the generalized eigenproblem for $(\mathbf{S}_y, \mathbf{\Sigma}_\epsilon^{(k)})$ by QZ (Moler & Stewart, 1973); let $\mathbf{A}^{(k)}$ have orthonormalized rows spanning the m eigenvectors whose finite eigenvalues are closest to 1, and set $\mathcal{C}^{(k)} \leftarrow \mathrm{row}(\mathbf{A}^{(k)})$
4: $\boldsymbol{\theta}^{(k+1)} \leftarrow \Phi^{+}_{\mathbf{A}^{(k)},\mathcal{S}}\big(\mathbf{A}^{(k)}\mathbf{S}_y(\mathbf{A}^{(k)})^\mathsf{T}\big)$; $\mathbf{\Sigma}_\epsilon^{(k+1)} \leftarrow \Pi_{\mathcal{K}}\big(\mathrm{mat}_{\mathcal{S}}(\boldsymbol{\theta}^{(k+1)})\big)$
5: **end for**
6: **return** $\mathbf{A}^{(T)}$, $\mathbf{\Sigma}_\epsilon^{(T+1)}$

3.1 Estimating the Constraint Matrix under Singular Noise

Because only the rowspace of $\mathbf{A}$ is identifiable, we estimate $\mathcal{C} = \mathrm{row}(\mathbf{A}) = \mathrm{Null}(\mathbf{\Sigma}_x)$. Let $\mathbf{V} \in \mathbb{R}^{d\times m}$ have orthonormal columns spanning $\mathcal{C}$, and let $\widehat{\mathbf{V}}$ span the bottom-m eigenspace of $\widehat{\mathbf{\Sigma}}_x := \mathbf{S}_y - \mathbf{\Sigma}_\epsilon$. Thus $\mathbf{P}_{\mathbf{V}} := \mathbf{V}\mathbf{V}^\mathsf{T}$ and $\mathbf{P}_{\widehat{\mathbf{V}}} := \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\mathsf{T}$. The eigengap is the intrinsic signal gap $\sigma^2_{\mathcal{C}}$, which remains positive when $\mathbf{\Sigma}_\epsilon$ is singular.

Theorem 3 (Informal: estimating $\mathbf{A}$)**.** *Under a fixed intrinsic gap and bounded conditioning, if $n \gtrsim d + \log(1/\delta)$, then for every $\mathbf{\Sigma}_\epsilon \succeq 0$, with probability at least $1-\delta$,*

$$\big\|\mathbf{P}_{\widehat{\mathbf{V}}} - \mathbf{P}_{\mathbf{V}}\big\|_{\mathrm{op}} \lesssim \sqrt{\frac{d+\log(1/\delta)}{n}}. \tag{3}$$

The hidden factor is proportional to $\|\mathbf{\Sigma}_y\|_{\mathrm{op}}/\sigma^2_{\mathcal{C}}$ and therefore does not diverge as $\lambda_{\min}(\mathbf{\Sigma}_\epsilon) \to 0$. The proof combines operator-norm concentration of $\mathbf{S}_y$ with a Davis–Kahan perturbation bound for $\mathbf{S}_y - \mathbf{\Sigma}_\epsilon$. Thus singularity itself causes no blow-up; only the intrinsic separation between the constraint and signal subspaces matters. Appendix F.1 gives the formal argument. This oracle result identifies what can be recovered when $\mathbf{\Sigma}_\epsilon$ is known; the next subsection asks when $\mathbf{\Sigma}_\epsilon$ can itself be learned from the constraints.

3.2 Estimating the Unknown Noise Covariance

The moment equation can determine $\mathbf{\Sigma}_\epsilon$ only when its supported parameterization is identifiable. We therefore separate the population identifiability question from the finite-sample estimation rate.

Theorem 4 (Informal: identifiability of $\mathbf{\Sigma}_\epsilon$)**.** *For a fixed support $\mathcal{S}$, the unrestricted supported parameter vector is uniquely determined by the constraint moments for every right-hand side exactly when the moment map has full column rank; necessarily,*

$$\operatorname{rank}(\Phi_{\mathbf{A},\mathcal{S}}) = q, \qquad q \le \frac{m(m+1)}{2}. \tag{4}$$

The proof views the moment equation as a linear system in the q free covariance parameters. A nontrivial kernel produces indistinguishable supported covariances, whereas full column rank gives a unique solution. The count is therefore necessary but not sufficient: the geometry of $\mathbf{A}$ and $\mathcal{S}$ must also make the moment columns independent. With uniqueness established, concentration controls the error of the corresponding finite-sample solve.

Theorem 5 (Informal: estimating $\mathbf{\Sigma}_\epsilon$)**.** *If $q \le m(m+1)/2$, the moment design has full column rank and is well conditioned, and $n \gtrsim d + \log(1/\delta)$, then given the true rowspace, with probability at least $1-\delta$,*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2 \lesssim \sqrt{\frac{m(d+\log(1/\delta))}{n}}. \tag{5}$$

The proof applies the pseudoinverse of the moment map to the projected sample-covariance error; $\sqrt{m}$ converts its operator norm to the Frobenius norm of the $m \times m$ moment matrix. Hence κ_Φ measures the statistical price of a nearly nonidentifiable support. Appendices H and F.2 give the formal statements. We can now estimate both $\mathbf{A}$ and $\mathbf{\Sigma}_\epsilon$ for a supplied m; the remaining structural task is to estimate m itself.

3.3 Estimating the Number of Linear Relations

Algorithm 1 takes m as input. To estimate it, we use the cap in Assumption 4 and replace $\mathcal{K}$ in each candidate run by

$$\mathcal{K}_{\mathrm{cap}} := \left\{\mathbf{M} \in \mathcal{K} : \mathbf{M} \preceq \bar{\sigma}^2 \mathbf{I}\right\}. \tag{6}$$

For a candidate k, define $c_k := \#\{j : \lambda_j^{[k]} \text{ is finite}, |\lambda_j^{[k]} - 1| \le \tau\}$. The sweep tests the identifiable candidate set $\mathcal{I}_q := \{k \in [d-1] : q \le k(k+1)/2\}$.

Algorithm 2 Estimating the number of linear relations

Input: Data $\mathbf{Y}$, support $\mathcal{S}$ with $q = |\mathcal{S}|$, cap $\bar{\sigma}^2$, threshold τ, budget T
Output: Estimated number of linear relations $\widehat{m}$

1: **for** $k \in \mathcal{I}_q$ in descending order **do**
2: Run Algorithm 1 with k linear relations, replacing $\Pi_{\mathcal{K}}$ by $\Pi_{\mathcal{K}_{\mathrm{cap}}}$
3: Set $c_k \leftarrow \#\{j : \lambda_j^{[k]} \text{ finite}, \ |\lambda_j^{[k]} - 1| \le \tau\}$ for the final pencil
4: **if** $c_k \ge k$, **return** $\widehat{m} \leftarrow k$
5: **end for**

Theorem 6 (Informal: estimating m)**.** *Under the spectral-cap, separation, identifiability, and basin conditions, define $f_n(d,m,\delta) := (d+\log(1/\delta))\max\{m^2, K^4\kappa(\mathbf{\Sigma}_y)^2(1+m\kappa_\Phi^2)/\tau^2\}$. If $n \gtrsim f_n(d,m,\delta)$, then with probability at least $1-\delta$,*

$$c_k \le m < k \quad (k = m+1, \dots, d-1), \qquad c_m = m, \qquad \widehat{m} = m. \tag{7}$$

The proof has two parts. Courant–Fischer and the spectral cap leave at most m eigenvalues near one, so every $k > m$ fails; the end-to-end error bound keeps the true m eigenvalues inside the threshold, so m passes. The descending sweep therefore avoids making any claim about underspecified candidates. Appendix K gives the formal threshold and proof. Having established recovery of all three unknown objects, we next ask whether the subspace rate can be improved by any estimator.

3.4 The Minimax Limit Matches the Estimation Rate

An upper bound alone does not show that the dependence on n and d is unavoidable. We therefore compare Theorem 3 with the worst-case risk over a controlled singular-noise class.

Theorem 7 (Informal: minimax lower bound)**.** *At bounded signal-to-noise ratio, the singular class $\mathcal{F}^s_{d,m,\kappa}$ in the appendix satisfies*

$$\inf_{\widehat{\mathcal{C}}} \sup_{\mathbb{P}\in\mathcal{F}^s_{d,m,\kappa}} \mathbb{E}_{\mathbb{P}} \left\|\mathbf{P}_{\widehat{\mathcal{C}}} - \mathbf{P}_{\mathcal{C}}\right\|_F^2 \gtrsim \min\left\{1, \frac{m(d-m-s)}{n}\right\}. \tag{8}$$

Thus the n and d dependence of Theorem 3 is minimax sharp up to the stated dimension factor. The proof constructs separated constraint subspaces in the Grassmannian and applies Fano's inequality, with a two-point Le Cam argument for the smallest dimensions. Appendix G specifies the model classes, SNR dependence, and dimensional restrictions. This lower bound validates the oracle statistical target; the final theorem shows that the alternating algorithm reaches it without receiving either $\mathbf{A}$ or $\mathbf{\Sigma}_\epsilon$.

3.5 The Alternating Algorithm Achieves End-to-End Recovery

The preceding results analyze one subspace step with $\mathbf{\Sigma}_\epsilon$ supplied and one covariance step with $\mathbf{A}$ supplied. We now couple them and show that a basin initialization is sufficient for the actual alternating iterates to reach the same statistical scale.

Theorem 8 (Informal: end-to-end recovery)**.** *Under fixed conditioning and a basin initialization, if $n \gtrsim m^2(d + \log(1/\delta))$, then with probability at least $1-\delta$,*

$$\max\left\{\left\|\mathbf{\Sigma}_\epsilon^{(T+1)} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}}, \left\|\mathbf{P}_{\mathcal{C}^{(T)}} - \mathbf{P}_{\mathcal{C}}\right\|_{\mathrm{op}}\right\} \lesssim 2^{-T}\rho + \sqrt{\frac{m(d+\log(1/\delta))}{n}}. \tag{9}$$

The contraction comes from a second-order insensitivity. Since $\mathbf{\Sigma}_x\mathbf{A}^\top = \mathbf{0}$, writing $\mathbf{F} = \bar{\mathbf{A}} - \mathbf{R}\mathbf{A}$ for an aligned constraint estimate gives

$$\bar{\mathbf{A}}\,\mathbf{\Sigma}_x\,\bar{\mathbf{A}}^\top = \mathbf{F}\,\mathbf{\Sigma}_x\,\mathbf{F}^\top. \tag{10}$$

Thus a subspace error of size e enters the covariance update as $\mathcal{O}(e^2)$. The resulting guarantee is local. Appendix J gives the basin, confidence, conditioning, and iteration-budget details suppressed here. Together, the results of this section show that GenEVD estimates $\mathbf{A}$, $\mathbf{\Sigma}_\epsilon$, and m, attains the minimax subspace scaling, and converges geometrically inside its basin. The next section tests these predictions and the spectral diagnostics empirically.

4 Empirical Evaluation Validates the Theoretical Claims

We evaluate constraint and noise-covariance recovery under correlated, possibly singular noise, and test whether the generalized-eigenvalue spectrum identifies a misspecified number of linear relations.

4.1 Experimental Design Tests Correlated and Singular Noise

Data generating process. We use a Gaussian flow-network model with $(d, m) = (5, 3)$. For $s = 0$, the identifiable non-diagonal noise support contains all five diagonal entries and $(1, 3)$, giving six free parameters. For $s = 1$ we make coordinate 5 noise-free, and for $s = 2$ we make coordinates 4 and 5 noise-free; their rows and columns are then removed from the support. The minimum coordinate-wise SNR over the noisy variables is 10. We vary $n \in \{500, 1000, 2000, 5000\}$; Appendix L gives the matrices and generator.

Baselines and metric. We compare against OLS (Fuller, 1987) and Standard PCA (Rao, 1964). The robust low-rank baselines are RPCA (Candès et al., 2011) and GoDec (Zhou & Tao, 2011). The noise-aware baselines are HeteroPCA (Zhang et al., 2022), Deflated-HeteroPCA (Zhou & Chen, 2025), and IPCA (Narasimhan & Shah, 2008). We omit IPCA for $s \geq 1$ because its Cholesky step

Table 2: Principal angle in degrees on the five-variable system, mean $\pm$ one standard deviation over 100 runs. IPCA requires $\mathbf{\Sigma}_\epsilon \succ 0$; bold is best. We observe that GenEVD (Ours) performs better than existing methods.

s	n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	IPCA	GenEVD (Ours)
0	500	13.9 ± 1.5	3.0 ± 0.8	2.7 ± 0.8	3.4 ± 1.2	66.0 ± 11.4	88.7 ± 1.1	$\mathbf{1.4 \pm 0.6}$	$\mathbf{1.4 \pm 0.6}$
0	1000	13.9 ± 1.3	2.8 ± 0.6	2.6 ± 0.6	3.2 ± 1.1	62.9 ± 16.8	89.2 ± 0.6	$\mathbf{1.1 \pm 0.4}$	$\mathbf{1.1 \pm 0.4}$
0	2000	13.9 ± 1.1	2.7 ± 0.5	2.6 ± 0.5	3.1 ± 1.0	63.6 ± 15.8	89.3 ± 0.5	$\mathbf{0.8 \pm 0.3}$	$\mathbf{0.8 \pm 0.3}$
0	5000	14.0 ± 0.9	2.6 ± 0.3	2.6 ± 0.3	2.9 ± 0.7	64.1 ± 14.4	89.6 ± 0.3	$\mathbf{0.5 \pm 0.2}$	$\mathbf{0.5 \pm 0.2}$
1	500	13.7 ± 1.4	3.5 ± 0.7	3.0 ± 0.8	4.0 ± 1.2	66.1 ± 11.5	88.8 ± 0.8	n/a	$\mathbf{1.2 \pm 0.6}$
1	1000	14.0 ± 1.1	3.4 ± 0.5	3.2 ± 0.5	3.6 ± 0.8	64.9 ± 13.0	89.2 ± 0.6	n/a	$\mathbf{0.9 \pm 0.4}$
1	2000	14.0 ± 1.1	3.4 ± 0.4	3.2 ± 0.4	3.5 ± 0.8	64.2 ± 14.4	89.4 ± 0.4	n/a	$\mathbf{0.6 \pm 0.3}$
1	5000	14.3 ± 0.7	3.3 ± 0.3	3.3 ± 0.3	3.5 ± 0.8	64.7 ± 12.9	89.6 ± 0.3	n/a	$\mathbf{0.4 \pm 0.2}$
2	500	12.6 ± 1.0	3.3 ± 0.9	2.7 ± 0.9	3.4 ± 1.1	64.3 ± 14.4	89.1 ± 0.6	n/a	$\mathbf{1.1 \pm 0.5}$
2	1000	12.5 ± 0.7	3.4 ± 0.5	3.1 ± 0.5	3.5 ± 0.9	63.5 ± 15.7	89.4 ± 0.4	n/a	$\mathbf{0.7 \pm 0.3}$
2	2000	12.6 ± 0.5	3.3 ± 0.4	3.2 ± 0.4	3.3 ± 0.7	64.6 ± 12.9	89.6 ± 0.3	n/a	$\mathbf{0.5 \pm 0.2}$
2	5000	12.5 ± 0.3	3.3 ± 0.2	3.2 ± 0.2	3.3 ± 0.4	61.4 ± 18.9	89.7 ± 0.2	n/a	$\mathbf{0.3 \pm 0.1}$

requires $\mathbf{\Sigma}_\epsilon \succ 0$. All methods receive the same target rank, and only applicable methods receive the support. We report the maximum principal angle in degrees over 100 replicates. Appendix L gives implementations and metrics.

This design first tests recovery under the correct model specification and then changes only m or the covariance support in controlled misspecification experiments. We begin with constraint recovery.

4.2 GenEVD Improves Constraint Recovery

We compare how the principal-angle error changes with sample size in the invertible and singular regimes, and then test whether the generalized spectrum reveals order or support misspecification.

Invertible $\mathbf{\Sigma}_\epsilon$. GenEVD and IPCA agree, as established by Lemma 1, and both improve steadily with sample size. GenEVD also outperforms every generic baseline.

Singular $\mathbf{\Sigma}_\epsilon$. GenEVD remains well defined, is best throughout, and its error decreases with sample size while the generic baselines plateau. This behavior agrees with the singular-noise recovery guarantee.

Number-of-relations misspecification with fixed covariance support. We generate data with $d = 5$, true $m = 3$, one noise-free coordinate, and a fixed five-entry covariance support. The candidate runs for $m = 3$ and $m = 4$ receive the same samples and the same support, so only the supplied number of relations changes. The true order produces the required three finite generalized eigenvalues near one, whereas the overspecified fit does not produce a fourth finite unity eigenvalue. Thus the spectrum detects overspecification without access to the true subspace. The baseline subspace estimators require the target rank as input and do not return an analogous order diagnostic; this is a difference in functionality rather than an unfavorable accuracy comparison. Appendix L.1 gives the complete setup and results.

Covariance-support misspecification. We next keep the data fixed but replace the true off-diagonal support entry $(1, 3)$ by $(1, 2)$, retaining the same number of free covariance parameters. The incorrect support prevents the moment update from matching the covariance structure and moves the candidate unity eigenvalues away from one. Hence the generalized spectrum also provides a practical warning that the supplied support is incompatible with the data. The other subspace baselines do not take a covariance support as input, so their outputs cannot distinguish these two support hypotheses. Full results are in Appendix L.2.

Recovery with unknown diagonal support. To test whether a noise-free coordinate must be supplied in advance, we generate diagonal singular $\mathbf{\Sigma}_\epsilon$ with $[\mathbf{\Sigma}_\epsilon]_{55} = 0$ but allow all five diagonal entries in the candidate support. With the correct $m = 3$, GenEVD recovers the withheld zero variance and retains the expected unity cluster, whereas the overspecified fit does not. The method

Table 3: Semi-synthetic principal angles in degrees at the largest sample size, mean $\pm$ one standard deviation. IPCA is inapplicable; bold is best. Full grids are in Appendix M.2. We observe that GenEVD (Ours) performs better than existing methods.

Dataset	n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
Energy	768	17.0 ± 0.6	13.7 ± 0.4	13.4 ± 0.4	13.8 ± 0.9	80.6 ± 3.6	12.6 ± 0.5	$\mathbf{1.5 \pm 0.6}$
Air Quality	8991	55.8 ± 3.5	41.3 ± 0.8	40.6 ± 0.8	33.3 ± 4.8	89.8 ± 0.2	51.6 ± 0.7	$\mathbf{0.8 \pm 0.3}$
Appliances	19735	13.1 ± 0.3	19.9 ± 0.5	20.2 ± 0.5	32.7 ± 9.0	86.1 ± 0.2	50.7 ± 0.4	$\mathbf{0.5 \pm 0.2}$
Gas Drift	13910	9.9 ± 0.1	9.0 ± 0.2	9.1 ± 0.2	11.7 ± 1.8	76.5 ± 0.6	13.3 ± 0.2	$\mathbf{0.4 \pm 0.1}$
MHEALTH	10000	13.2 ± 0.2	9.5 ± 0.2	9.4 ± 0.2	9.2 ± 1.2	63.2 ± 0.2	78.3 ± 0.3	$\mathbf{0.4 \pm 0.2}$
Power Plant	9568	30.2 ± 0.3	21.8 ± 0.4	21.7 ± 0.4	22.5 ± 10.6	87.8 ± 0.6	70.9 ± 0.6	$\mathbf{0.5 \pm 0.2}$
Household	20000	18.1 ± 0.2	13.2 ± 0.4	13.5 ± 0.4	50.9 ± 25.5	75.8 ± 0.7	85.1 ± 1.6	$\mathbf{0.3 \pm 0.2}$

can therefore recover the nearly zero variance and reject the larger order jointly. HeteroPCA-type methods accommodate diagonal heteroskedasticity through diagonal imputation, but they do not jointly estimate the number of relations through a generalized-eigenvalue test. OLS, Standard PCA, RPCA, and GoDec do not estimate $\mathbf{\Sigma}_\epsilon$. Appendix L.3 reports all diagnostics.

Number-of-relations misspecification in Case Study 2. Finally, we repeat the relation-number experiment in the higher-dimensional setting $(d, m, s) = (10, 5, 3)$. Candidate orders 4, 5, and 6 use the same samples and a common ten-parameter support. The correct fifth eigenvalue concentrates near unity, whereas the overspecified candidate is rejected, so the descending procedure stops at the true value. Existing baselines are evaluated after being supplied the target rank and thus do not perform this order-estimation task. Appendix M.1 reports the complete results.

These controlled synthetic experiments support the three mechanisms used by the theory: consistency with IPCA in the invertible case, stable recovery under singular noise, and spectral rejection of incompatible orders or supports. We next test whether the same behavior persists on references derived from measured sensor data.

4.3 Semi-Synthetic Experiments on Real-World Datasets

We use five-variable subsets of the UCI Energy Efficiency, Air Quality, Appliances Energy, Gas Sensor Array Drift, MHEALTH, Combined Cycle Power Plant, and Household Electric Power Consumption datasets (Kelly et al., 2023). After dataset-specific cleaning and standardization, each dataset matrix is projected onto a rank-two reference subspace and corrupted by rank-three correlated Gaussian measurement noise with two noise-free coordinates. Every method receives the same target rank and samples; GenEVD receives the six-entry covariance support but not its values. We report principal-angle error over independent noise realizations, with complete preprocessing, sample grids, seeds, and covariance constructions in Appendix M.2.

Across all seven datasets, GenEVD gives the smallest principal-angle error and remains stable as the sample size grows. Standard PCA, RPCA, and GoDec cannot separate signal from correlated heterogeneous noise; OLS exploits only a response-noise special case; and the HeteroPCA variants estimate diagonal residual noise but cannot model the injected correlations. The consistent gain across distinct sensor domains supports the central claim that estimating a structured non-diagonal, singular $\mathbf{\Sigma}_\epsilon$ improves constraint recovery. These experiments evaluate the specified correlated, singular measurement-noise model while preserving known constraint and covariance ground truth.

5 Conclusion

We introduced GenEVD for recovering linear relations when measurement noise is heterogeneous, correlated, possibly singular, and unknown together with the number of relations. The algorithm avoids $\mathbf{\Sigma}_\epsilon^{-1}$ by alternating a generalized-eigenspace update with an identifiable supported-moment update for $\mathbf{\Sigma}_\epsilon$, while a descending unity-eigenvalue procedure estimates m; IPCA and OLS are recovered as special cases. Our analysis gives finite-sample guarantees for recovering the constraint subspace, $\mathbf{\Sigma}_\epsilon$, and m, proves local geometric convergence of the joint estimator, and establishes minimax-optimal dependence on sample size and dimension for subspace recovery. Controlled synthetic experiments and experiments constructed from seven real-world datasets support the predicted behavior under correlated and singular measurement noise.

AI USE STATEMENT

We used a large language model for drafting, copy-editing, proof auditing, literature review, and numerical checks. It did not originate the problem, estimator, or experiments. The authors verified and take responsibility for all content.

ETHICS STATEMENT

This theoretical work uses synthetic and public sensor data, with no human subjects, identifying information, or new data release.

REPRODUCIBILITY STATEMENT

The appendix restates and proves every result and records all notation, generators, supports, sample-size grids, preprocessing, and implementation details.

REFERENCES

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011. doi: 10.1145/1970392.1970395.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Wayne A. Fuller. *Measurement Error Models*. Wiley, New York, 1987.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. `https://archive.ics.uci.edu`, 2023.

Ivan Markovsky and Sabine Van Huffel. Overview of total least-squares methods. *Signal Processing*, 87(10):2283–2302, 2007.

Cleve B. Moler and G. W. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM Journal on Numerical Analysis*, 10(2):241–256, 1973.

Shankar Narasimhan and Sirish L. Shah. Model identification and error covariance matrix estimation from noisy data using pca. *Control Engineering Practice*, 16(2):146–155, 2008.

C. R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya, Series A*, 26:329–358, 1964.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Paul D. Wentzell, Dorothy T. Andrews, Donald C. Hamilton, Klaas Faber, and Bruce R. Kowalski. Maximum likelihood principal component analysis. *Journal of Chemometrics*, 11(4):339–366, 1997.

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.

Anru R. Zhang, T. Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *Annals of Statistics*, 50(1):53–80, 2022. doi: 10.1214/21-AOS2074.

Tianyi Zhou and Dacheng Tao. GoDec: Randomized low-rank and sparse matrix decomposition in noisy case. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 33–40, 2011.

Yuchen Zhou and Yuxin Chen. Deflated HeteroPCA: Overcoming the curse of ill-conditioning in heteroskedastic PCA. *The Annals of Statistics*, 53(1):91–116, 2025. doi: 10.1214/24-AOS2456.

Supplementary Material: Provably Identifying Linear Constraints under Unknown and Singular Measurement Noise

This supplement is organized as follows. Appendices A and B fix notation and record the original IPCA algorithm. Appendices C–E collect the tools every later proof uses: the setup lemmas, the covariance concentration bound that defines Δ_n, and the perturbation toolbox. Appendices F–H prove the singular-case results of Section 3, Appendix I gives the specialization to invertible $\mathbf{\Sigma}_\epsilon$, and Appendix J proves the end-to-end guarantee of Section 3. Appendix K proves recovery of the number of linear relations for the descending sweep. Appendices L–M give the experimental details.

A Notation **13**

B The IPCA Algorithm **14**

C Proofs for Section 2 **14**

D Proof of Lemma 12 **16**

E Perturbation Toolbox **17**

F Proofs for Section 3 (Singular Noise Covariance) **18**

F.1 Subspace recovery 18

F.2 Noise covariance estimation 19

G Proof of Theorem 7 **20**

H Proof of Theorem 4 **22**

I Invertible Noise Covariance **23**

I.1 Subspace consistency 24

I.2 Rank consistency 24

I.3 Noise covariance estimation, scaled form 24

J Proof of Theorem 8 **25**

J.1 The selection rule of line 3 25

K Estimating the Number of Linear Relations **32**

L Case Study 1: Five-Variable Identifiable System **34**

L.1 Misspecifying the Number of Linear Relations with Fixed Support 34

L.2 Incorrect Off-Diagonal Support 35

L.3 Unknown Diagonal Support 37

M Additional Experimental Results **37**

M.1 Relation-Number Misspecification in Case Study 2 38
M.2 Semi-Synthetic Experiments on Real-World Datasets 39

A NOTATION

We list the symbols specific to this paper. Standard objects (norms, Orlicz norms, orthogonal projectors, trace, rank, null space, Gaussian and Kullback-Leibler notation, and asymptotic symbols) follow the usual conventions and are not repeated here. Symbols local to a single proof are defined inside that proof.

Symbol	Meaning
Model	
$d, n, m, r := d - m$	Measured variables, samples, constraints, signal rank
$s := \dim(\mathcal{N})$	Number of noise-free directions
$\mathbf{y}^{(t)} = \mathbf{x}^{(t)} + \boldsymbol{\epsilon}^{(t)}$	Observation, latent signal, measurement error at sample t
$\mathbf{A} \in \mathbb{R}^{m \times d}$	Constraint matrix, $\mathbf{A}\mathbf{x}^{(t)} = \mathbf{0}$, normalized so $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}_m$
$\mathbf{a}_k$	k-th column of $\mathbf{A}$
$\mathcal{C} := \mathrm{row}(\mathbf{A}) = \mathrm{Null}(\boldsymbol{\Sigma}_x)$	Constraint subspace, $\dim \mathcal{C} = m$
$\mathcal{N} := \mathrm{Null}(\boldsymbol{\Sigma}_\epsilon)$	Noise-free subspace, $\dim \mathcal{N} = s$; $\mathbf{P}_{\mathcal{N}^\perp} := \mathbf{P}_{\mathcal{N}^\perp}$
$\mathcal{G}_{m,d}$	Grassmannian manifold of m-dimensional subspaces of $\mathbb{R}^d$
Covariances	
$\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_\epsilon$	Signal, error and observation covariance
$\mathbf{S}_y := \frac{1}{n}\mathbf{Y}\mathbf{Y}^\mathsf{T}$	Sample covariance
$\widehat{\boldsymbol{\Sigma}}_x := \mathbf{S}_y - \widehat{\boldsymbol{\Sigma}}_\epsilon$	Plug-in signal covariance used by the subspace step
$\Delta_n := \lVert \mathbf{S}_y - \boldsymbol{\Sigma}_y \rVert_{\mathrm{op}}$	Sample covariance error, Eq. (24)
U_n	$K^2 \lVert \boldsymbol{\Sigma}_y \rVert_{\mathrm{op}} \sqrt{(d + \log(1/\delta))/n}$, the concentration scale; $\Delta_n \le C\, U_n$ on $\mathcal{E}_\delta$
$\mathbf{L}$	Cholesky factor of $\boldsymbol{\Sigma}_\epsilon$, defined only when $\boldsymbol{\Sigma}_\epsilon \succ 0$
$\mathbf{M} := \mathbf{L}^{-1}\boldsymbol{\Sigma}_y\mathbf{L}^{-\mathsf{T}}, \widehat{\mathbf{M}}$	Scaled population and sample covariance, eigenvalues μ_j, $\hat{\mu}_j$
Noise parameterization	
$\mathcal{S}, q := \lvert\mathcal{S}\rvert$	Symmetric support, pairs (i, j) with $i \le j$; number of free parameters
q_s	Free parameters supported on the noisy coordinates (singular covariance)
$\boldsymbol{\theta} \in \mathbb{R}^q, \mathrm{mat}_\mathcal{S}(\boldsymbol{\theta})$	Free parameters of $\boldsymbol{\Sigma}_\epsilon$ and the matrix they build
$\Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta}) := \mathbf{A}\,\mathrm{mat}_\mathcal{S}(\boldsymbol{\theta})\mathbf{A}^\mathsf{T}$	Moment map into $\mathbb{S}^m$
$\kappa_\Phi := 1/\sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}})$	Identifiability constant of the moment map
$\mathbf{a}_i \odot \mathbf{a}_j$	$\mathbf{a}_i\mathbf{a}_j^\mathsf{T} + \mathbf{a}_j\mathbf{a}_i^\mathsf{T}$ for $i < j$, and $\mathbf{a}_i\mathbf{a}_i^\mathsf{T}$ for $i = j$
$\mathcal{K}$	$\{\mathbf{M} \succeq 0 : \mathrm{supp}(\mathbf{M}) \subseteq \mathcal{S}\}$, the projection set in line 4
$\mathcal{K}_{\mathrm{cap}}, \bar{\sigma}^2$	Capped projection set for estimating the number of linear relations and its spectral noise cap
Constants and iterates	
K	Sub-Gaussian constant of Assumption 1
$\sigma_\mathcal{C}^2 := \lambda_{\min}(\boldsymbol{\Sigma}_x\vert_{\mathcal{C}^\perp})$	Intrinsic signal gap of Assumption 3
$\beta_\star := 6\sqrt{3}\,\kappa(\boldsymbol{\Sigma}_y)^{3/2}/\sigma_\mathcal{C}^2$	Deterministic constant of Lemma 22, Eq. (79)
$\varepsilon := \mu_r - 1$	Spectral gap of the scaled covariance, defined only when $\boldsymbol{\Sigma}_\epsilon \succ 0$
$\kappa, \kappa(\mathbf{S}_y)$	Signal-to-noise ratio in the minimax classes; condition number of $\mathbf{S}_y$
$\cos\vartheta$	Cosine of the largest principal angle between $\mathcal{C}$ and $\mathcal{N}^\perp$, Eq. (18)
ρ	Basin radius of Theorem 8, Eq. (94)

Symbol	Meaning
$\mathbf{F}_k := \mathbf{A}^{(k)} - \mathbf{R}_k\mathbf{A}$	Aligned constraint-matrix error at iteration k; $\mathbf{R}_k$ is the Procrustes rotation
$b_k := \left\|\mathbf{\Sigma}_\epsilon^{(k)} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}}$, $e_k := \|\mathbf{F}_k\|_{\mathrm{op}}$	Covariance and constraint-matrix error at iteration k
$T, T_\star(\delta) := \log_2(\rho/U_n)$	Iteration budget of Algorithm 1 and the budget the guarantee requires
τ, τ_n	Fixed and sample-dependent thresholds for estimating the number of linear relations
c_k	Number of finite generalized eigenvalues within τ of one for candidate value k

B THE IPCA ALGORITHM

For completeness we record the original IPCA algorithm of Narasimhan & Shah (2008). Given $\mathbf{\Sigma}_\epsilon^{(k)}$, the constraint step computes the Cholesky factor $\mathbf{L}^{(k)}$, forms $\widehat{\mathbf{M}}^{(k)} := (\mathbf{L}^{(k)})^{-1}\mathbf{S}_y(\mathbf{L}^{(k)})^{-\mathsf{T}}$, and extracts the bottom-m eigenvectors $\mathbf{U}_m^{(k)}$ to form $\mathbf{A}^{(k)} = (\mathbf{U}_m^{(k)})^{\mathsf{T}}(\mathbf{L}^{(k)})^{-1}$. The noise step maximizes the Gaussian log-likelihood of the residuals $\mathbf{r}^{(k)}(t) = \mathbf{A}^{(k)}\mathbf{y}^{(t)}$ over the free entries of $\mathbf{\Sigma}_\epsilon$ on $\mathcal{S}$.

Algorithm 3 IPCA (Narasimhan & Shah, 2008)

Input: Data $\mathbf{Y} \in \mathbb{R}^{d\times n}$, number of linear relations m, tolerance τ, support $\mathcal{S}$
Output: Estimated $\mathbf{A}$ and $\mathbf{\Sigma}_\epsilon$

1: Compute $\mathbf{S}_y \leftarrow \frac{1}{n}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}$ and initialize $\mathbf{\Sigma}_\epsilon^{(1)} \succeq 0$ on $\mathcal{S}$.
2: **repeat**
3: Cholesky: $\mathbf{L}^{(k)}(\mathbf{L}^{(k)})^{\mathsf{T}} = \mathbf{\Sigma}_\epsilon^{(k)}$, then scale: $\widehat{\mathbf{M}}^{(k)} \leftarrow (\mathbf{L}^{(k)})^{-1}\mathbf{S}_y(\mathbf{L}^{(k)})^{-\mathsf{T}}$.
4: Eigendecompose $\widehat{\mathbf{M}}^{(k)}$ and set $\mathbf{A}^{(k)} \leftarrow (\mathbf{U}_m^{(k)})^{\mathsf{T}}(\mathbf{L}^{(k)})^{-1}$.
5: Update $\mathbf{\Sigma}_\epsilon^{(k+1)} \leftarrow \arg\min_{\mathbf{\Sigma}_\epsilon \text{ on } \mathcal{S}} \big[n \log|\mathbf{A}^{(k)}\mathbf{\Sigma}_\epsilon(\mathbf{A}^{(k)})^{\mathsf{T}}| + \sum_t \mathbf{r}^{(k)}(t)^{\mathsf{T}}(\mathbf{A}^{(k)}\mathbf{\Sigma}_\epsilon(\mathbf{A}^{(k)})^{\mathsf{T}})^{-1}\mathbf{r}^{(k)}(t)\big]$.
6: $k \leftarrow k+1$.
7: **until** $|\lambda^{(k)} - \lambda^{(k-1)}|/\max\{1, |\lambda^{(k-1)}|\} < \tau$, where $\lambda^{(k)} = \sum_{j=r+1}^{d} \hat{\mu}_j^{(k)}$.

For diagonal $\mathbf{\Sigma}_\epsilon$ the noise step is not a coordinate-wise ratio. The identity $\mathbf{A}\,\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)\mathbf{A}^{\mathsf{T}} = \sum_{j=1}^{d} \sigma_j^2\, \mathbf{a}_j\mathbf{a}_j^{\mathsf{T}}$ couples the d variances through $m(m+1)/2$ equations. When that system is consistent and the nonnegativity constraints are slack it reduces to a linear solve, and otherwise it is the constrained likelihood problem of line 5 above. The constraint step solves $\mathbf{S}_y\mathbf{v} = \lambda\mathbf{\Sigma}_\epsilon^{(k)}\mathbf{v}$, which is what motivates the generalized formulation.

C PROOFS FOR SECTION 2

Proof of Lemma 1. If $\mathbf{M}\mathbf{w} = \mu\mathbf{w}$, substitute $\mathbf{w} = \mathbf{L}^{\mathsf{T}}\mathbf{v}$ to obtain $\mathbf{L}^{-1}\mathbf{\Sigma}_y\mathbf{v} = \mu\mathbf{L}^{\mathsf{T}}\mathbf{v}$, and hence $\mathbf{\Sigma}_y\mathbf{v} = \mu\mathbf{L}\mathbf{L}^{\mathsf{T}}\mathbf{v} = \mu\mathbf{\Sigma}_\epsilon\mathbf{v}$. The map $\mathbf{v} \mapsto \mathbf{L}^{\mathsf{T}}\mathbf{v}$ is bijective because $\mathbf{L}$ is invertible, so it preserves eigenvalue multiplicities and maps the eigenvalue-1 eigenspace of $\mathbf{M}$ to that of the pencil. Moreover, $\mathbf{M} = \mathbf{I} + \mathbf{L}^{-1}\mathbf{\Sigma}_x\mathbf{L}^{-\mathsf{T}}$, where the second term is positive semidefinite with rank $r = d - m$. Hence $\mathbf{M}$ has exactly m eigenvalues equal to one and all remaining eigenvalues exceed one, so this is its bottom-m eigenspace. □

Theorem 2 (GenEVD recovers OLS). *Let $\mathbf{Y}^{\mathsf{T}} = [\mathbf{Z}\ \mathbf{r}]$, where $\mathbf{Z} \in \mathbb{R}^{n\times(d-1)}$ has full column rank, and let $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^{\mathsf{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{r}$. For $c > 0$, the pencil $(\mathbf{S}_y, c\,\mathbf{e}_d\mathbf{e}_d^{\mathsf{T}})$ has exactly one finite generalized*

eigenvalue, and its eigenpair satisfies

$$\widehat{\lambda} = \frac{\left\|\mathbf{r} - \mathbf{Z}\widehat{\boldsymbol{\beta}}\right\|^2}{nc}, \qquad \widehat{\mathbf{v}} \propto \begin{bmatrix} -\widehat{\boldsymbol{\beta}} \\ 1 \end{bmatrix}. \tag{11}$$

Proof. Partition the sample covariance as

$$\mathbf{S}_y = \begin{bmatrix} \mathbf{S}_{zz} & \mathbf{s}_{zr} \\ \mathbf{s}_{zr}^\mathsf{T} & s_{rr} \end{bmatrix} := \frac{1}{n} \begin{bmatrix} \mathbf{Z}^\mathsf{T}\mathbf{Z} & \mathbf{Z}^\mathsf{T}\mathbf{r} \\ \mathbf{r}^\mathsf{T}\mathbf{Z} & \mathbf{r}^\mathsf{T}\mathbf{r} \end{bmatrix}. \tag{12}$$

For a finite generalized eigenpair, write $\mathbf{v} = [\mathbf{u}^\mathsf{T}\ \alpha]^\mathsf{T}$. The upper block of $\mathbf{S}_y\mathbf{v} = \lambda c\, \mathbf{e}_d\mathbf{e}_d^\mathsf{T}\mathbf{v}$ is $\mathbf{S}_{zz}\mathbf{u} + \mathbf{s}_{zr}\alpha = \mathbf{0}$. Since $\mathbf{S}_{zz} \succ 0$, a nonzero eigenvector must have $\alpha \neq 0$, and after setting $\alpha = 1$ we obtain

$$\mathbf{u} = -\mathbf{S}_{zz}^{-1}\mathbf{s}_{zr} = -(\mathbf{Z}^\mathsf{T}\mathbf{Z})^{-1}\mathbf{Z}^\mathsf{T}\mathbf{r} = -\widehat{\boldsymbol{\beta}}. \tag{13}$$

The lower block then gives

$$\lambda c = s_{rr} - \mathbf{s}_{zr}^\mathsf{T}\mathbf{S}_{zz}^{-1}\mathbf{s}_{zr} = \frac{1}{n}\left\|\mathbf{r} - \mathbf{Z}\widehat{\boldsymbol{\beta}}\right\|^2, \tag{14}$$

which proves Eq. (11). Finally, the Schur-complement identity yields

$$\det\left(\mathbf{S}_y - \lambda c\, \mathbf{e}_d\mathbf{e}_d^\mathsf{T}\right) = \det(\mathbf{S}_{zz})\left(s_{rr} - \mathbf{s}_{zr}^\mathsf{T}\mathbf{S}_{zz}^{-1}\mathbf{s}_{zr} - \lambda c\right). \tag{15}$$

This regular determinant polynomial has degree one, hence exactly one finite generalized eigenvalue; the remaining $d-1$ generalized eigenvalues are infinite. Therefore the finite-eigenvalue selection for $m = 1$ returns the OLS relation. □

Lemma 9 (The constraint subspace is the eigenspace of the scaled covariance at eigenvalue one.)**.** *Let $\boldsymbol{\Sigma}_\epsilon \succ 0$ with Cholesky factor $\boldsymbol{\Sigma}_\epsilon = \mathbf{L}\mathbf{L}^\mathsf{T}$ and let $\mathbf{M} := \mathbf{L}^{-1}\boldsymbol{\Sigma}_y\mathbf{L}^{-\mathsf{T}}$. Then $\mathbf{M} = \mathbf{I} + \mathbf{L}^{-1}\boldsymbol{\Sigma}_x\mathbf{L}^{-\mathsf{T}}$, the eigenspace of $\mathbf{M}$ at eigenvalue 1 equals $\mathbf{L}^\mathsf{T}\mathcal{C}$ and has dimension m, and every other eigenvalue of $\mathbf{M}$ exceeds 1. Writing $\mu_1 \geq \cdots \geq \mu_d$ for its eigenvalues and $\varepsilon := \mu_r - 1$, we have $\varepsilon > 0$, $\mu_r \geq 1 + \varepsilon$ and $\mu_{r+1} = \cdots = \mu_d = 1$.*

Proof of Lemma 9. Since $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_\epsilon$ and $\boldsymbol{\Sigma}_\epsilon = \mathbf{L}\mathbf{L}^\mathsf{T}$, the scaling splits $\mathbf{M}$ into an identity part and a signal part:

$$\mathbf{M} \;=\; \mathbf{L}^{-1}\left(\boldsymbol{\Sigma}_x + \mathbf{L}\mathbf{L}^\mathsf{T}\right)\mathbf{L}^{-\mathsf{T}} \;=\; \mathbf{I} + \mathbf{N}, \qquad \mathbf{N} \;:=\; \mathbf{L}^{-1}\boldsymbol{\Sigma}_x\mathbf{L}^{-\mathsf{T}} \;\succeq\; 0. \tag{16}$$

The eigenvalue-1 eigenspace of $\mathbf{M}$ is thus the null space of $\mathbf{N}$, which we identify using $\mathbf{N} \succeq 0$ (so $\mathbf{N}\mathbf{w} = \mathbf{0}$ iff $\mathbf{w}^\mathsf{T}\mathbf{N}\mathbf{w} = 0$ iff $\boldsymbol{\Sigma}_x\mathbf{L}^{-\mathsf{T}}\mathbf{w} = \mathbf{0}$):

$$\mathbf{M}\mathbf{w} = \mathbf{w} \iff \mathbf{N}\mathbf{w} = \mathbf{0} \iff \mathbf{L}^{-\mathsf{T}}\mathbf{w} \in \mathrm{Null}(\boldsymbol{\Sigma}_x) = \mathcal{C} \iff \mathbf{w} \in \mathbf{L}^\mathsf{T}\mathcal{C}. \tag{17}$$

This space has dimension m because $\mathbf{L}^\mathsf{T}$ is invertible. Every remaining eigenvalue of $\mathbf{M}$ is $1 + \lambda_j(\mathbf{N})$ with $\lambda_j(\mathbf{N}) > 0$, and $\mathrm{rank}(\mathbf{N}) = \mathrm{rank}(\boldsymbol{\Sigma}_x) = r$ because $\mathbf{L}^{-1}$ is invertible. Hence exactly r eigenvalues exceed 1, and $\varepsilon = \mu_r - 1 = \lambda_r(\mathbf{N}) > 0$. □

Lemma 10 (The constraint and noise-free subspaces intersect trivially.)**.** *Under Assumption 2, $\mathcal{C} \cap \mathcal{N} = \{\mathbf{0}\}$. Consequently $\mathbf{P}_{\mathcal{N}^\perp}|_{\mathcal{C}}$ is injective, where $\mathbf{P}_{\mathcal{N}^\perp} := \mathbf{P}_{\mathcal{N}^\perp}$, and*

$$\cos\vartheta := \sigma_{\min}\left(\mathbf{P}_{\mathcal{N}^\perp}|_{\mathcal{C}}\right) = \min_{\mathbf{v}\in\mathcal{C},\,\|\mathbf{v}\|=1} \|\mathbf{P}_{\mathcal{N}^\perp}\mathbf{v}\| \;>\; 0. \tag{18}$$

Proof of Lemma 10. Suppose $\mathbf{v} \in \mathcal{C} \cap \mathcal{N}$ with $\mathbf{v} \neq \mathbf{0}$. Then $\mathbf{v}^\mathsf{T}\boldsymbol{\Sigma}_x\mathbf{v} = 0$ and $\mathbf{v}^\mathsf{T}\boldsymbol{\Sigma}_\epsilon\mathbf{v} = 0$, hence $\mathbf{v}^\mathsf{T}\boldsymbol{\Sigma}_y\mathbf{v} = 0$, contradicting Assumption 2. For the second claim, if $\mathbf{v} \in \mathcal{C}$ has $\mathbf{P}_{\mathcal{N}^\perp}\mathbf{v} = \mathbf{0}$ then $\mathbf{v} \in \mathcal{N}$, so $\mathbf{v} \in \mathcal{C} \cap \mathcal{N} = \{\mathbf{0}\}$. Thus $\mathbf{P}_{\mathcal{N}^\perp}|_{\mathcal{C}}$ is injective and its smallest singular value, which is the minimum in Eq. (18) over the compact unit sphere of $\mathcal{C}$, is attained and positive. □

D Proof of Lemma 12

We first record the net inequality for quadratic forms, with a proof, so that the argument does not depend on a numbered exercise.

Lemma 11 (A net suffices for the quadratic form of a symmetric matrix.)**.** *Let $\mathbf{E} \in \mathbb{S}^D$, let $\alpha \in (0, \frac{1}{2})$, and let $\mathcal{U}$ be a α-net of $\mathbb{S}^{D-1}$. Then $\|\mathbf{E}\|_{\mathrm{op}} \le (1-2\alpha)^{-1} \max_{\mathbf{w}\in\mathcal{U}} \left|\mathbf{w}^{\mathsf{T}}\mathbf{E}\mathbf{w}\right|$.*

Proof. For symmetric $\mathbf{E}$, $\|\mathbf{E}\|_{\mathrm{op}} = \max_{\mathbf{x}\in\mathbb{S}^{D-1}} \left|\mathbf{x}^{\mathsf{T}}\mathbf{E}\mathbf{x}\right|$, and the maximum is attained at some $\mathbf{x}^\star$ by compactness. Pick $\mathbf{w} \in \mathcal{U}$ with $\|\mathbf{w} - \mathbf{x}^\star\| \le \alpha$. Then

$$\begin{aligned}\left|\mathbf{w}^{\mathsf{T}}\mathbf{E}\mathbf{w} - \mathbf{x}^{\star\mathsf{T}}\mathbf{E}\mathbf{x}^\star\right| &= \left|(\mathbf{w}-\mathbf{x}^\star)^{\mathsf{T}}\mathbf{E}\mathbf{w} + \mathbf{x}^{\star\mathsf{T}}\mathbf{E}(\mathbf{w}-\mathbf{x}^\star)\right| \\ &\le \|\mathbf{E}\|_{\mathrm{op}} \|\mathbf{w}-\mathbf{x}^\star\| \left(\|\mathbf{w}\| + \|\mathbf{x}^\star\|\right) \;\le\; 2\alpha \|\mathbf{E}\|_{\mathrm{op}}, \end{aligned} \tag{19}$$

so $\left|\mathbf{w}^{\mathsf{T}}\mathbf{E}\mathbf{w}\right| \ge (1-2\alpha)\|\mathbf{E}\|_{\mathrm{op}}$, which rearranges to the claim. □

Lemma 12 (Sample covariance concentration.)**.** *Let $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)}$ be i.i.d. zero-mean random vectors in $\mathbb{R}^D$ with covariance $\mathbf{\Sigma}$ satisfying $\left\|\langle \mathbf{u}^{(t)}, \mathbf{w}\rangle\right\|_{\psi_2} \le K\sqrt{\mathbf{w}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{w}}$ for all $\mathbf{w}$, and let $\widehat{\mathbf{\Sigma}} := \frac{1}{n}\sum_t \mathbf{u}^{(t)}\mathbf{u}^{(t)\mathsf{T}}$. There are absolute constants $C_1, C_2 > 0$ such that for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\mathrm{op}} \;\le\; K^2 \|\mathbf{\Sigma}\|_{\mathrm{op}} \left[C_1\sqrt{\frac{D+\log(1/\delta)}{n}} + C_2\,\frac{D+\log(1/\delta)}{n}\right]. \tag{20}$$

Proof of Lemma 12. Notation. Write $\mathbf{E} := \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ and, for a unit vector $\mathbf{w} \in \mathbb{S}^{D-1}$, $W_t(\mathbf{w}) := \langle \mathbf{u}^{(t)}, \mathbf{w}\rangle$.

By hypothesis $\|W_t(\mathbf{w})\|_{\psi_2} \le K\sqrt{\mathbf{w}^{\mathsf{T}}\mathbf{\Sigma}\mathbf{w}} \le K\|\mathbf{\Sigma}\|_{\mathrm{op}}^{1/2}$. A random variable is sub-Gaussian if and only if its square is sub-exponential, with $\left\|W_t^2\right\|_{\psi_1} = \|W_t\|_{\psi_2}^2 \le K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$, and centering preserves this up to a factor 2 (Vershynin, 2018, §2.7). The quantity $\mathbf{w}^{\mathsf{T}}\mathbf{E}\mathbf{w} = \frac{1}{n}\sum_t (W_t^2 - \mathbb{E}W_t^2)$ is an average of n i.i.d. centered sub-exponential variables, so Bernstein's inequality for sub-exponential summands (Vershynin, 2018, §2.8) gives, in the form $\mathbb{P}[|\frac{1}{n}\sum_t X_t| \ge s] \le 2\exp\{-cn\min(s^2/L^2, s/L)\}$ valid for i.i.d. centered X_t with $\|X_t\|_{\psi_1} \le L$, and with $L = 2K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$,

$$\mathbb{P}\left[\left|\mathbf{w}^{\mathsf{T}}\mathbf{E}\mathbf{w}\right| \ge s\right] \;\le\; 2\exp\left\{-cn\min\left(\frac{s^2}{K^4\|\mathbf{\Sigma}\|_{\mathrm{op}}^2}, \frac{s}{K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}}\right)\right\} \tag{21}$$

after absorbing the factor 2 into the absolute constant c.

A α-net of $\mathbb{S}^{D-1}$ of cardinality at most $(1+2/\alpha)^D$ exists (Vershynin, 2018, §4.2). Taking $\alpha = \frac{1}{4}$ gives a net $\mathcal{U}$ with $|\mathcal{U}| \le 9^D$, and Lemma 11 gives

$$\|\mathbf{E}\|_{\mathrm{op}} \;\le\; 2\max_{\mathbf{w}\in\mathcal{U}} \left|\mathbf{w}^{\mathsf{T}}\mathbf{E}\mathbf{w}\right|. \tag{22}$$

Combining Eqs. (21) and (22) with a union bound over $\mathcal{U}$,

$$\mathbb{P}\left[\|\mathbf{E}\|_{\mathrm{op}} \ge 2s\right] \;\le\; 2\cdot 9^D \exp\left\{-cn\min\left(\frac{s^2}{K^4\|\mathbf{\Sigma}\|_{\mathrm{op}}^2}, \frac{s}{K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}}\right)\right\}. \tag{23}$$

Setting the right-hand side to δ and solving, the sub-Gaussian branch $s \le K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$ requires $s \ge K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}\sqrt{(D\log 9 + \log(2/\delta))/(cn)}$ and the sub-exponential branch $s > K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$ requires $s \ge K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}(D\log 9 + \log(2/\delta))/(cn)$. Both exponents carry $K^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$, not K. Summing the two branches and absorbing $\log 9$ and c into C_1, C_2 gives Eq. (20). □

Remark 1 (Three applications of Lemma 12.). With $\mathbf{u}^{(t)} = \mathbf{y}^{(t)}$ and $D = d$ we obtain Eq. (24). With $\mathbf{u}^{(t)} = \mathbf{z}^{(t)} = \mathbf{L}^{-1}\mathbf{y}^{(t)}$ and $D = d$, valid only when $\mathbf{\Sigma}_\epsilon \succ 0$, we obtain $\left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_{\mathrm{op}} \lesssim K^2 \left\|\mathbf{M}\right\|_{\mathrm{op}} \sqrt{(d + \log(1/\delta))/n}$, which is the form used in Theorem 18. With $\mathbf{u}^{(t)} = \mathbf{P}_{\mathcal{N}^\perp}\mathbf{y}^{(t)}$ and $D = d - s$ we obtain a bound on $\|\mathbf{P}_{\mathcal{N}^\perp}(\mathbf{S}_y - \mathbf{\Sigma}_y)\mathbf{P}_{\mathcal{N}^\perp}\|_{\mathrm{op}}$. This compressed bound controls only the restriction to $\mathcal{N}^\perp$; the full operator norm $\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}}$ still ranges over all directions in $\mathbb{R}^d$.

Applied to $\mathbf{y}^{(t)}$ with $D = d$, Lemma 12 controls the quantity that drives every bound in this paper,

$$\Delta_n := \|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}} \lesssim K^2 \|\mathbf{\Sigma}_y\|_{\mathrm{op}} \sqrt{\frac{d + \log(1/\delta)}{n}}, \tag{24}$$

valid in the regime $n \gtrsim K^4(d + \log(1/\delta))$ where the first term dominates. We write $\mathcal{E}_\delta$ for this event, which has probability at least $1 - \delta$, and C for the absolute constant in Eq. (24).

Several later arguments need Δ_n to sit below an explicit threshold. Since Δ_n is random we never assume this; instead we require n to be large enough that the right-hand side of Eq. (24) is already below the threshold, which is a condition on the model and not on the sample. Solving

$$CK^2 \|\mathbf{\Sigma}_y\|_{\mathrm{op}} \sqrt{\frac{d + \log(1/\delta)}{n}} \le \frac{\lambda_{\min}(\mathbf{\Sigma}_y)}{2} \qquad \text{for } n \tag{25}$$

gives the threshold used throughout,

$$n_\star(\delta) := 4C^2K^4 \kappa(\mathbf{\Sigma}_y)^2\big(d + \log(1/\delta)\big), \tag{26}$$

so that $n \ge n_\star(\delta)$ implies $\Delta_n \le \lambda_{\min}(\mathbf{\Sigma}_y)/2$ on $\mathcal{E}_\delta$.

E Perturbation Toolbox

Both the oracle sections and the end-to-end proof use the following two standard facts. We state them once here so that no later argument has to reprove them.

Lemma 13 (Perturbation of the bottom invariant subspace)**.** *Let $\mathbf{M}, \widehat{\mathbf{M}} \in \mathbb{S}^D$ and set $\mathbf{E} := \widehat{\mathbf{M}} - \mathbf{M}$. Suppose the m smallest eigenvalues of $\mathbf{M}$ are all equal to 0, with eigenspace $\mathcal{U}$, and that the remaining $D - m$ eigenvalues are at least $\gamma > 0$. Let $\widehat{\mathcal{U}}$ be the invariant subspace of $\widehat{\mathbf{M}}$ for its m smallest eigenvalues. If $\|\mathbf{E}\|_{\mathrm{op}} < \gamma/2$, then $\widehat{\mathcal{U}}$ is well defined, $\dim\widehat{\mathcal{U}} = m$, and*

$$\left\|\mathbf{P}_{\widehat{\mathcal{U}}} - \mathbf{P}_{\mathcal{U}}\right\|_{\mathrm{op}} \le \frac{\|\mathbf{E}\|_{\mathrm{op}}}{\gamma - \|\mathbf{E}\|_{\mathrm{op}}} \le \frac{2\|\mathbf{E}\|_{\mathrm{op}}}{\gamma}. \tag{27}$$

The same conclusion holds when the m smallest eigenvalues of $\mathbf{M}$ all equal a common value θ_0 and the rest are at least $\theta_0 + \gamma$, since replacing $\mathbf{M}$ by $\mathbf{M} - \theta_0\mathbf{I}$ and $\widehat{\mathbf{M}}$ by $\widehat{\mathbf{M}} - \theta_0\mathbf{I}$ changes neither $\mathbf{E}$ nor any eigenspace.

Proof. This is the Davis–Kahan $\sin\Theta$ theorem for invariant subspaces (Davis & Kahan, 1970); see also Stewart & Sun (1990, Chapter V). We give the argument in the form we use.

$\widehat{\mathcal{U}}$ is well defined. By Weyl's inequality the spectrum of $\widehat{\mathbf{M}}$ splits as

$$\underbrace{\mu \in [-\|\mathbf{E}\|_{\mathrm{op}},\ \|\mathbf{E}\|_{\mathrm{op}}]}_{m \text{ smallest}} \qquad \text{and} \qquad \underbrace{\mu \ge \gamma - \|\mathbf{E}\|_{\mathrm{op}}}_{\text{the rest}}, \tag{28}$$

and $\|\mathbf{E}\|_{\mathrm{op}} < \gamma/2$ makes $\gamma - \|\mathbf{E}\|_{\mathrm{op}} > \|\mathbf{E}\|_{\mathrm{op}}$, so the two groups are separated.

The bound. Write $\mathbf{P} := \mathbf{P}_{\mathcal{U}}$, $\widehat{\mathbf{P}} := \mathbf{P}_{\widehat{\mathcal{U}}}$, $\widehat{\mathbf{P}}^\perp := \mathbf{I} - \widehat{\mathbf{P}}$, and

$$\delta := \gamma - \|\mathbf{E}\|_{\mathrm{op}} > \frac{\gamma}{2} > 0. \tag{29}$$

Two structural facts drive the argument. First, $\mathcal{U}$ is the eigenspace of $\mathbf{M}$ at the eigenvalue 0, so

$$\mathbf{M}\mathbf{P} = \mathbf{0}. \tag{30}$$

Second, $\widehat{\mathcal{U}}^{\perp}$ is an invariant subspace of $\widehat{\mathbf{M}}$, so $\widehat{\mathbf{P}}^{\perp}\widehat{\mathbf{M}} = \widehat{\mathbf{P}}^{\perp}\widehat{\mathbf{M}}\widehat{\mathbf{P}}^{\perp}$, and by the spectral split above the symmetric operator

$$\mathbf{A} \;:=\; \widehat{\mathbf{P}}^{\perp}\widehat{\mathbf{M}}\widehat{\mathbf{P}}^{\perp} \qquad \text{restricted to } \widehat{\mathcal{U}}^{\perp} \tag{31}$$

has all $D-m$ of its eigenvalues at least δ; it is therefore invertible on $\widehat{\mathcal{U}}^{\perp}$, with inverse of operator norm at most $1/\delta$.

Now compress $\mathbf{E}$ between the two projectors and use Eq. (30):

$$\widehat{\mathbf{P}}^{\perp}\mathbf{E}\mathbf{P} \;=\; \widehat{\mathbf{P}}^{\perp}\left(\widehat{\mathbf{M}}-\mathbf{M}\right)\mathbf{P} \;=\; \widehat{\mathbf{P}}^{\perp}\widehat{\mathbf{M}}\mathbf{P} \;=\; \widehat{\mathbf{P}}^{\perp}\widehat{\mathbf{M}}\widehat{\mathbf{P}}^{\perp}\mathbf{P} \;=\; \mathbf{A}\left(\widehat{\mathbf{P}}^{\perp}\mathbf{P}\right). \tag{32}$$

The matrix $\widehat{\mathbf{P}}^{\perp}\mathbf{P}$ has its range inside $\widehat{\mathcal{U}}^{\perp}$, so Eq. (32) may be inverted there:

$$\left\|\widehat{\mathbf{P}}^{\perp}\mathbf{P}\right\|_{\mathrm{op}} \;\le\; \left\|\mathbf{A}^{-1}|_{\widehat{\mathcal{U}}^{\perp}}\right\|_{\mathrm{op}} \left\|\widehat{\mathbf{P}}^{\perp}\mathbf{E}\mathbf{P}\right\|_{\mathrm{op}} \;\le\; \frac{\|\mathbf{E}\|_{\mathrm{op}}}{\delta} \;=\; \frac{\|\mathbf{E}\|_{\mathrm{op}}}{\gamma-\|\mathbf{E}\|_{\mathrm{op}}} \;\le\; \frac{2\,\|\mathbf{E}\|_{\mathrm{op}}}{\gamma}, \tag{33}$$

the last step because $\delta > \gamma/2$. For two orthogonal projectors of equal rank, $\left\|\widehat{\mathbf{P}}-\mathbf{P}\right\|_{\mathrm{op}} = \left\|\widehat{\mathbf{P}}^{\perp}\mathbf{P}\right\|_{\mathrm{op}}$, both being the sine of the largest principal angle, which gives Eq. (27). See Yu et al. (2015) for the same bound in a form convenient for statistical applications. □

Lemma 14 (Subspace distance under an invertible map)**.** *Let $\mathbf{W} \in \mathbb{R}^{d\times d}$ be invertible and let $\mathcal{U}_1, \mathcal{U}_2$ be subspaces of equal dimension. Then $\|\mathbf{P}_{\mathbf{W}\mathcal{U}_1} - \mathbf{P}_{\mathbf{W}\mathcal{U}_2}\|_{\mathrm{op}} \le \kappa(\mathbf{W})\,\|\mathbf{P}_{\mathcal{U}_1} - \mathbf{P}_{\mathcal{U}_2}\|_{\mathrm{op}}$, where $\kappa(\mathbf{W}) = \|\mathbf{W}\|_{\mathrm{op}}\left\|\mathbf{W}^{-1}\right\|_{\mathrm{op}}$.*

Proof. For subspaces of equal dimension the projector distance is a one-sided directional distance,

$$\|\mathbf{P}_{\mathcal{U}_1} - \mathbf{P}_{\mathcal{U}_2}\|_{\mathrm{op}} \;=\; \max_{\mathbf{x}\in\mathcal{U}_2,\ \|\mathbf{x}\|=1} \mathrm{dist}(\mathbf{x},\mathcal{U}_1), \tag{34}$$

so it suffices to bound $\mathrm{dist}(\mathbf{y}, \mathbf{W}\mathcal{U}_1)$ for a unit $\mathbf{y} \in \mathbf{W}\mathcal{U}_2$. Write $\mathbf{y} = \mathbf{W}\mathbf{x}/\,\|\mathbf{W}\mathbf{x}\|$ with $\mathbf{x} \in \mathcal{U}_2$ a unit vector, and let $\mathbf{x}' \in \mathcal{U}_1$ attain $\mathrm{dist}(\mathbf{x},\mathcal{U}_1)$. The competitor $\mathbf{W}\mathbf{x}'/\,\|\mathbf{W}\mathbf{x}\|$ lies in $\mathbf{W}\mathcal{U}_1$, and $\|\mathbf{W}\mathbf{x}\| \ge \left\|\mathbf{W}^{-1}\right\|_{\mathrm{op}}^{-1}$, so

$$\mathrm{dist}(\mathbf{y}, \mathbf{W}\mathcal{U}_1) \;\le\; \frac{\|\mathbf{W}(\mathbf{x}-\mathbf{x}')\|}{\|\mathbf{W}\mathbf{x}\|} \;\le\; \frac{\|\mathbf{W}\|_{\mathrm{op}}}{\|\mathbf{W}\mathbf{x}\|}\,\|\mathbf{x}-\mathbf{x}'\| \;\le\; \|\mathbf{W}\|_{\mathrm{op}}\left\|\mathbf{W}^{-1}\right\|_{\mathrm{op}}\mathrm{dist}(\mathbf{x},\mathcal{U}_1). \tag{35}$$

Taking the maximum over $\mathbf{x}$ and applying Eq. (34) on both sides gives the claim. □

F Proofs for Section 3 (Singular Noise Covariance)

F.1 Subspace recovery

Theorem 3 (Subspace recovery when $\boldsymbol{\Sigma}_\epsilon$ is singular). *Let Assumptions 1 to 3 hold and let $\mathbf{P}_{\widehat{\mathbf{V}}}$ be the projector onto the bottom-m eigenspace of $\widehat{\boldsymbol{\Sigma}}_x := \mathbf{S}_y - \boldsymbol{\Sigma}_\epsilon$. If $\|\mathbf{S}_y - \boldsymbol{\Sigma}_y\|_{\mathrm{op}} < \sigma_{\mathcal{C}}^2/2$, then*

$$\left\|\mathbf{P}_{\widehat{\mathbf{V}}} - \mathbf{P}_{\mathbf{V}}\right\|_{\mathrm{op}} \;\le\; \frac{2}{\sigma_{\mathcal{C}}^2}\,\|\mathbf{S}_y - \boldsymbol{\Sigma}_y\|_{\mathrm{op}}. \tag{36}$$

Moreover, for $n \gtrsim K^4(d+\log(1/\delta))$, with probability at least $1-\delta$,

$$\left\|\mathbf{P}_{\widehat{\mathbf{V}}} - \mathbf{P}_{\mathbf{V}}\right\|_{\mathrm{op}} \;\lesssim\; \frac{K^2\,\|\boldsymbol{\Sigma}_y\|_{\mathrm{op}}}{\sigma_{\mathcal{C}}^2}\sqrt{\frac{d+\log(1/\delta)}{n}}. \tag{37}$$

Proof. Work with $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y - \boldsymbol{\Sigma}_\epsilon$ directly, which is what lets the Cholesky factor of $\boldsymbol{\Sigma}_\epsilon$ disappear. Because $\mathcal{C} = \mathrm{Null}(\boldsymbol{\Sigma}_x)$ and $\mathrm{rank}(\boldsymbol{\Sigma}_x) = r = d-m$, the eigenvalues of $\boldsymbol{\Sigma}_x \succeq 0$ satisfy

$$\lambda_1(\boldsymbol{\Sigma}_x) \ge \cdots \ge \lambda_r(\boldsymbol{\Sigma}_x) \ge \sigma_{\mathcal{C}}^2 > 0 = \lambda_{r+1}(\boldsymbol{\Sigma}_x) = \cdots = \lambda_d(\boldsymbol{\Sigma}_x), \tag{38}$$

so the m smallest eigenvalues of $\mathbf{\Sigma}_x$ are all zero with eigenspace $\mathcal{C}$, and the remaining $d-m$ are at least $\sigma_{\mathcal{C}}^2$. This is exactly the configuration of Lemma 13 with $\gamma = \sigma_{\mathcal{C}}^2$.

The perturbation is

$$\widehat{\mathbf{\Sigma}}_x - \mathbf{\Sigma}_x = (\mathbf{S}_y - \mathbf{\Sigma}_\epsilon) - (\mathbf{\Sigma}_y - \mathbf{\Sigma}_\epsilon) = \mathbf{S}_y - \mathbf{\Sigma}_y, \tag{39}$$

in which $\mathbf{\Sigma}_\epsilon$ cancels identically; neither its invertibility nor its Cholesky factor enters. Under the hypothesis $\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}} < \sigma_{\mathcal{C}}^2/2$, Lemma 13 applies and gives

$$\left\|\mathbf{P}_{\widehat{\mathbf{V}}} - \mathbf{P}_{\mathbf{V}}\right\|_{\mathrm{op}} \le \frac{2\,\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}}}{\sigma_{\mathcal{C}}^2}, \tag{40}$$

which is Eq. (36). Applying Lemma 12 to $\mathbf{y}^{(t)}$ with $D = d$, which is legitimate under Assumption 1 whether or not $\mathbf{\Sigma}_\epsilon$ is invertible, bounds $\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}} \lesssim K^2\,\|\mathbf{\Sigma}_y\|_{\mathrm{op}}\sqrt{(d+\log(1/\delta))/n}$ with probability at least $1-\delta$ in the regime $n \gtrsim K^4(d+\log(1/\delta))$, and substituting gives Eq. (37). In Frobenius norm the squared error is at most $2\min(m, d-m)$ times the square of the right-hand side, since the difference of two rank-m projectors has rank at most $2\min(m, d-m)$. □

Remark 2 (Which quantity diverges as the noise covariance becomes singular). The scaled analysis of Appendix I degrades as $\lambda_{\min}(\mathbf{\Sigma}_\epsilon) \to 0$, but not through the gap ε. Take $\mathbf{\Sigma}_x = \mathrm{diag}(1,1,0)$ and $\mathbf{\Sigma}_\epsilon = \mathrm{diag}(t,1,1)$ with $t \in (0,1]$. Then $\mathbf{M} = \mathrm{diag}(1+1/t,\,2,\,1)$, so $\varepsilon = \mu_r - 1 = 1$ for every t, while $\|\mathbf{M}\|_{\mathrm{op}} = 1 + 1/t$ blows up. Recovering $\mathbf{A}$ in scaled coordinates degrades the same way through $\left\|\mathbf{L}^{-1}\right\|_{\mathrm{op}} = \lambda_{\min}(\mathbf{\Sigma}_\epsilon)^{-1/2}$, as does a scaling-based covariance estimate through $\|\mathbf{M}\|_{\mathrm{op}}$. Theorem 3 avoids all three because it never forms $\mathbf{L}$, and $\sigma_{\mathcal{C}}^2$ is a property of $\mathbf{\Sigma}_x$ alone. We note that a diverging factor in an upper bound is not by itself a proof that the error diverges.

The claim is pointwise, not uniform. Singularity of the noise covariance alone does not degrade Eq. (3), but a sequence along which $\mathbf{\Sigma}_y$ becomes ill conditioned does, even at fixed $\sigma_{\mathcal{C}}^2$: for $\mathbf{\Sigma}_x = \mathrm{diag}(1,0)$ and $\mathbf{\Sigma}_\epsilon = t\mathbf{I}_2$ every assumption holds for each $t > 0$ with $\sigma_{\mathcal{C}}^2 = 1$, while $\kappa(\mathbf{\Sigma}_y) = (1+t)/t \to \infty$.

F.2 Noise covariance estimation

Theorem 5 (Noise covariance estimation). *Let Assumption 1 hold, let* $\mathbf{A}$ *have orthonormal rows, and let* $\mathcal{S}$ *satisfy* $q \le m(m+1)/2$ *with* $\Phi_{\mathbf{A},\mathcal{S}}$ *injective. Given the true* $\mathbf{A}$*, with probability at least* $1-\delta$*,*

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2 &\le \kappa_\Phi\sqrt{m}\,\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}} \\ &\lesssim \kappa_\Phi\sqrt{m}\,K^2\,\|\mathbf{\Sigma}_y\|_{\mathrm{op}}\sqrt{\frac{d+\log(1/\delta)}{n}},\end{aligned} \tag{41}$$

where the second line applies when $n \gtrsim K^4(d+\log(1/\delta))$.

Proof. The constraint $\mathbf{A}\mathbf{x}^{(t)} = \mathbf{0}$ gives $\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^\mathsf{T} = \mathbf{0}$, so the population moment equation is $\Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{\Sigma}_y\mathbf{A}^\mathsf{T}$. The estimator is the least-squares solution $\widehat{\boldsymbol{\theta}} := \Phi^+_{\mathbf{A},\mathcal{S}}(\mathbf{A}\mathbf{S}_y\mathbf{A}^\mathsf{T})$, which is what one computes whether or not the system is consistent; consistency can fail when $q < m(m+1)/2$, so we do not assume it. Injectivity gives $\Phi^+_{\mathbf{A},\mathcal{S}}\Phi_{\mathbf{A},\mathcal{S}} = \mathbf{I}$, so by linearity of the pseudoinverse

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \Phi^+_{\mathbf{A},\mathcal{S}}\big(\mathbf{A}(\mathbf{S}_y - \mathbf{\Sigma}_y)\mathbf{A}^\mathsf{T}\big), \qquad \text{hence} \qquad \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2 \le \kappa_\Phi\,\left\|\mathbf{A}(\mathbf{S}_y - \mathbf{\Sigma}_y)\mathbf{A}^\mathsf{T}\right\|_F. \tag{42}$$

Since $\mathbf{A}(\mathbf{S}_y - \mathbf{\Sigma}_y)\mathbf{A}^\mathsf{T}$ is $m \times m$, its Frobenius norm is at most $\sqrt{m}$ times its operator norm, and $\left\|\mathbf{A}(\mathbf{S}_y - \mathbf{\Sigma}_y)\mathbf{A}^\mathsf{T}\right\|_{\mathrm{op}} \le \|\mathbf{A}\|_{\mathrm{op}}^2\,\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}} = \|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}}$ because $\mathbf{A}$ has orthonormal rows. This gives the first inequality of Eq. (41), and Lemma 12 applied to $\mathbf{y}^{(t)}$ gives the second.

Nothing above used invertibility of $\mathbf{\Sigma}_\epsilon$. When $\mathbf{\Sigma}_\epsilon$ is singular with $\mathrm{Null}(\mathbf{\Sigma}_\epsilon) = \mathcal{N}$, the free parameters are those supported on the noisy coordinates: taking $\mathcal{N}$ to be spanned by standard basis vectors, which is the noise-free-coordinate setting of Section 4, and writing $\mathcal{J}$ for the noisy indices, the map acts on $\mathbb{R}^{q_s}$ with $q_s := |\{(i,j) \in \mathcal{S} : i, j \in \mathcal{J}\}|$ and the argument is unchanged with q replaced by q_s and κ_Φ by $\kappa_{\Phi,s}$. For a general $\mathcal{N}$ one parameterizes $\mathbf{\Sigma}_\epsilon$ in a basis adapted to $\mathcal{N}$ and q_s becomes the dimension of that parameterization. □

We state the normalization explicitly because κ_Φ is not scale free: the product $\kappa_\Phi \|\mathbf{A}\|_{\mathrm{op}}^2$ is invariant under rescaling and under orthogonal changes of row basis, but not under the general left multiplication that orthonormalizes an arbitrary representative, so fixing the representative once, as Section 2 does, removes the ambiguity.

Remark 3 (A sharper rate at a fixed constraint matrix). Passing through $\|\mathbf{S}_y - \mathbf{\Sigma}_y\|_{\mathrm{op}}$ is lossy, because the moment residual $\mathbf{A}(\mathbf{S}_y - \mathbf{\Sigma}_y)\mathbf{A}^\mathsf{T}$ is $m \times m$ and only $\dim(\mathbb{S}^m)$ directions of the perturbation are ever seen. Applying Lemma 12 instead to the compressed observation $\mathbf{u}^{(t)} := \mathbf{A}\mathbf{y}^{(t)} \in \mathbb{R}^m$, which is sub-Gaussian relative to its own covariance $\mathbf{A}\mathbf{\Sigma}_y\mathbf{A}^\mathsf{T}$ with the same constant K, bounds $\left\|\frac{1}{n}\sum_t \mathbf{u}^{(t)}\mathbf{u}^{(t)\mathsf{T}} - \mathbf{A}\mathbf{\Sigma}_y\mathbf{A}^\mathsf{T}\right\|_{\mathrm{op}}$ by $K^2\|\mathbf{\Sigma}_y\|_{\mathrm{op}}[c_1\sqrt{(m+\log(1/\delta))/n} + c_2(m+\log(1/\delta))/n]$, giving, in the regime $n \gtrsim K^4(m+\log(1/\delta))$ where the first term dominates,

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2 \;\lesssim\; \kappa_\Phi\sqrt{m}\,K^2\,\|\mathbf{\Sigma}_y\|_{\mathrm{op}}\sqrt{\frac{m+\log(1/\delta)}{n}}, \tag{43}$$

that is $\sqrt{m^2/n}$ with no explicit factor of d for a fixed oracle $\mathbf{A}$. The end-to-end analysis of Appendix J instead controls the residual uniformly over the data-dependent $\mathbf{A}^{(k)}$, which yields ambient-dimension dependence. Moreover, κ_Φ can itself grow with d: with all d diagonal entries free and $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}_m$ one has $\kappa_\Phi \geq d/m$.

G Proof of Theorem 7

Theorem 7 (Minimax lower bounds, both regimes). *(i) Invertible noise covariance. Let* $\mathcal{F}_{d,m,\kappa} := \{\mathcal{N}(\mathbf{0}, \kappa\sigma^2\mathbf{P}_{\mathcal{C}^\perp} + \sigma^2\mathbf{I}) : \mathcal{C} \in \mathcal{G}_{m,d}\}$*, one distribution per constraint subspace. Every member has* $\mathrm{rank}(\mathbf{\Sigma}_x) = r$ *and a flat signal spectrum,* $\lambda_1(\mathbf{\Sigma}_x) = \cdots = \lambda_r(\mathbf{\Sigma}_x) = \kappa\sigma^2$*, so* $\sigma_{\mathcal{C}}^2 = \kappa\sigma^2$ *and* $\|\mathbf{\Sigma}_y\|_{\mathrm{op}} = (1+\kappa)\sigma^2$ *throughout the class. There is* $c > 0$ *with*

$$\inf_{\widehat{\mathcal{C}}} \sup_{\mathbb{P}\in\mathcal{F}_{d,m,\kappa}} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathbf{P}_{\widehat{\mathcal{C}}} - \mathbf{P}_{\mathcal{C}}\right\|_F^2\right] \;\geq\; c\min\left\{1, \frac{m(d-m)(1+\kappa)}{n\kappa^2}\right\}. \tag{44}$$

(ii) Singular noise covariance. Fix a subspace $\mathcal{N}$ *with* $\dim\mathcal{N} = s$ *and suppose* $m \leq d - s - 1$*. Let*

$$\mathcal{F}_{d,m,\kappa}^s := \left\{\mathcal{N}\big(\mathbf{0},\ \kappa\sigma^2\mathbf{P}_{\mathcal{C}^\perp} + \sigma^2\mathbf{P}_{\mathcal{N}^\perp}\big)\ :\ \mathcal{C} \in \mathcal{G}_{m,d},\ \mathcal{C} \subseteq \mathcal{N}^\perp\right\}, \tag{45}$$

so that $\mathbf{\Sigma}_\epsilon = \sigma^2\mathbf{P}_{\mathcal{N}^\perp}$ *has rank* $d - s$ *and, as in part (i), the signal covariance is* $\mathbf{\Sigma}_x = \kappa\sigma^2\mathbf{P}_{\mathcal{C}^\perp}$ *with a flat spectrum, giving* $\sigma_{\mathcal{C}}^2 = \kappa\sigma^2$ *throughout the class. There is* $c_0 > 0$ *with*

$$\inf_{\widehat{\mathcal{C}}} \sup_{\mathbb{P}\in\mathcal{F}_{d,m,\kappa}^s} \mathbb{E}_{\mathbb{P}}\left[\left\|\mathbf{P}_{\widehat{\mathcal{C}}} - \mathbf{P}_{\mathcal{C}}\right\|_F^2\right] \;\geq\; c_0\min\left\{1, \frac{m(d-m-s)(1+\kappa)}{n\kappa^2}\right\}. \tag{46}$$

Pinning the signal spectrum from below as well as above is what makes the comparison with Theorem 17 meaningful: a class that only caps $\|\mathbf{\Sigma}_x\|_{\mathrm{op}}$ contains arbitrarily weak signals and has worst-case risk bounded away from zero at every n.

We first record the exact Kullback-Leibler identity used in both parts, since the constant matters for the comparison in Section 3.4.

Lemma 15 (Kullback-Leibler divergence between two constraint hypotheses.)**.** *Let* $\mathcal{C}_0, \mathcal{C}_1$ *be* m*-dimensional subspaces of* $\mathbb{R}^D$*, write* $\mathbf{P}_i := \mathbf{P}_{\mathcal{C}_i}$*, and let* $\mathbf{\Sigma}_i := \sigma^2[\mathbf{P}_i + (1+\kappa)(\mathbf{I}_D - \mathbf{P}_i)]$*. Then*

$$\mathrm{KL}\left(\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)\,\|\,\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)\right) \;=\; \frac{\kappa^2}{4(1+\kappa)}\,\|\mathbf{P}_0 - \mathbf{P}_1\|_F^2\,. \tag{47}$$

Proof. The two covariances have the same eigenvalue multiset, namely σ^2 with multiplicity m and $\sigma^2(1+\kappa)$ with multiplicity $D - m$, so the log-determinant term vanishes and

$$\mathrm{KL} = \tfrac{1}{2}\left[\mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0) - D\right]. \tag{48}$$

Write $t := \mathrm{tr}(\mathbf{P}_0\mathbf{P}_1)$ and use $\mathbf{\Sigma}_1^{-1} = \sigma^{-2}[\mathbf{P}_1 + (1+\kappa)^{-1}(\mathbf{I} - \mathbf{P}_1)]$. Expanding the four blocks and using $\mathrm{tr}\,\mathbf{P}_0 = \mathrm{tr}\,\mathbf{P}_1 = m$,

$$\mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0) = t + (1+\kappa)(m-t) + (1+\kappa)^{-1}(m-t) + (D - 2m + t) \tag{49}$$

$$= D - 2(m-t) + (m-t)\left[(1+\kappa) + (1+\kappa)^{-1}\right], \tag{50}$$

hence $\operatorname{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0) - D = (m-t)[(1+\kappa) + (1+\kappa)^{-1} - 2] = (m-t)\kappa^2/(1+\kappa)$. Finally $\|\mathbf{P}_0 - \mathbf{P}_1\|_F^2 = \operatorname{tr}\mathbf{P}_0 + \operatorname{tr}\mathbf{P}_1 - 2t = 2(m-t)$, so $m - t = \frac{1}{2}\|\mathbf{P}_0 - \mathbf{P}_1\|_F^2$, which gives Eq. (47). □

Lemma 16 (Exact projector distance between graph subspaces.)**.** *For $\mathbf{T} \in \mathbb{R}^{(D-m)\times m}$ let $\mathbf{U}_\mathbf{T} := \binom{\mathbf{I}_m}{\mathbf{T}}(\mathbf{I}_m + \mathbf{T}^\mathsf{T}\mathbf{T})^{-1/2}$ and $\mathbf{P}_\mathbf{T} := \mathbf{U}_\mathbf{T}\mathbf{U}_\mathbf{T}^\mathsf{T}$. Then*

$$\|\mathbf{P}_\mathbf{T} - \mathbf{P}_\mathbf{S}\|_F^2 = 2\left\|(\mathbf{I} + \mathbf{T}\mathbf{T}^\mathsf{T})^{-1/2}(\mathbf{S} - \mathbf{T})(\mathbf{I} + \mathbf{S}^\mathsf{T}\mathbf{S})^{-1/2}\right\|_F^2. \tag{51}$$

Consequently, if $\|\mathbf{T}\|_{\mathrm{op}} \le 1$ and $\|\mathbf{S}\|_{\mathrm{op}} \le 1$ then

$$\tfrac{1}{2}\|\mathbf{T} - \mathbf{S}\|_F^2 \;\le\; \|\mathbf{P}_\mathbf{T} - \mathbf{P}_\mathbf{S}\|_F^2 \;\le\; 2\|\mathbf{T} - \mathbf{S}\|_F^2, \tag{52}$$

with absolute constants that do not depend on D or m.

Proof. Define

$$\mathbf{U}_{\mathbf{T},\perp} \;:=\; \begin{pmatrix} -\mathbf{T}^\mathsf{T} \\ \mathbf{I}_{D-m} \end{pmatrix}\left(\mathbf{I}_{D-m} + \mathbf{T}\mathbf{T}^\mathsf{T}\right)^{-1/2}, \tag{53}$$

whose columns are orthonormal and span the orthogonal complement of the graph of $\mathbf{T}$. For two subspaces of equal dimension with orthonormal bases $\mathbf{U}_\mathbf{S}$ and $\mathbf{U}_\mathbf{T}$, the projector distance and the cross term are related by

$$\|\mathbf{P}_\mathbf{T} - \mathbf{P}_\mathbf{S}\|_F^2 \;=\; 2\left\|\mathbf{U}_{\mathbf{T},\perp}^\mathsf{T}\mathbf{U}_\mathbf{S}\right\|_F^2, \tag{54}$$

and a direct computation of that cross term gives the exact identity of Eq. (51),

$$\mathbf{U}_{\mathbf{T},\perp}^\mathsf{T}\mathbf{U}_\mathbf{S} \;=\; \left(\mathbf{I} + \mathbf{T}\mathbf{T}^\mathsf{T}\right)^{-1/2}(\mathbf{S} - \mathbf{T})\left(\mathbf{I} + \mathbf{S}^\mathsf{T}\mathbf{S}\right)^{-1/2}. \tag{55}$$

When $\|\mathbf{S}\|_{\mathrm{op}}, \|\mathbf{T}\|_{\mathrm{op}} \le 1$, each outer factor is symmetric positive definite with all singular values in $\left[1/\sqrt{2},\, 1\right]$, so the identity is a two-sided comparison and Eq. (52) follows. □

Proof of Theorem 7. Throughout, $D = d$ for part (i) and $D = d - s$ for part (ii), $p := m(D-m)$, and $\eta := \kappa^2/(1+\kappa)$ is the constant appearing in Lemma 15. We use graph coordinates: subspaces are parameterized by $\mathbf{T} \in \mathbb{R}^{(D-m)\times m}$ as in Lemma 16, after an orthogonal change of basis that puts a fixed reference subspace at $\mathbf{T} = \mathbf{0}$. The exact identity of Eq. (51) provides constants independent of D and m.

Hypotheses. Set $\mathbf{\Sigma}_x^{(\mathbf{T})} := \kappa\sigma^2\mathbf{P}_\mathbf{T}^\perp$ and $\mathbb{P}_\mathbf{T} := \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_x^{(\mathbf{T})} + \mathbf{\Sigma}_\epsilon)$, with $\mathbf{\Sigma}_\epsilon = \sigma^2\mathbf{I}_D$ in part (i) and $\mathbf{\Sigma}_\epsilon = \sigma^2\mathbf{P}_{\mathcal{N}^\perp}$ in part (ii), where in part (ii) the construction runs inside $\mathcal{N}^\perp$ and the block of $\mathbf{\Sigma}_y$ on $\mathcal{N}$ equals $\kappa\sigma^2\mathbf{I}_s$ for every hypothesis. Each $\mathbb{P}_\mathbf{T}$ lies in the stated class, by the verification given below when $\mathbf{\Sigma}_\epsilon$ is singular and immediately when $\mathbf{\Sigma}_\epsilon$ is invertible. By Lemma 15, $\mathrm{KL}(\mathbb{P}_\mathbf{T}^{\otimes n}\|\mathbb{P}_\mathbf{S}^{\otimes n}) = \frac{n\eta}{4}\|\mathbf{P}_\mathbf{T} - \mathbf{P}_\mathbf{S}\|_F^2$.

Large p. Suppose $p \ge p_0$ for the absolute constant p_0 fixed below. By the Varshamov-Gilbert bound applied to $\{0,1\}^p$ there is Ω with $\log|\Omega| \ge c_1 p$ and $d_H(\omega,\omega') \ge p/8$ for distinct elements. Put $\mathbf{T}_\omega := a\mathbf{B}_\omega$, where $\mathbf{B}_\omega \in \{0,1\}^{(D-m)\times m}$ reshapes ω and

$$a^2 := c\,\min\left\{\frac{1}{p},\; \frac{1}{n\eta}\right\}, \tag{56}$$

for a small absolute c. The cap $a^2 \le c/p$ gives $\|\mathbf{T}_\omega\|_{\mathrm{op}} \le \|\mathbf{T}_\omega\|_F \le a\sqrt{p} \le \sqrt{c} \le 1$, so Lemma 16 applies to every pair, uniformly in D and m. Since $\|\mathbf{T}_\omega - \mathbf{T}_{\omega'}\|_F^2 = a^2 d_H(\omega,\omega')$, Eq. (52) gives separation at least $a^2 p/16$ and maximum squared distance at most $2a^2 p$, so

$$\max_{\omega\neq\omega'} \mathrm{KL}(\mathbb{P}_{\mathbf{T}_\omega}^{\otimes n}\|\mathbb{P}_{\mathbf{T}_{\omega'}}^{\otimes n}) \le \tfrac{n\eta}{2}a^2 p \le \tfrac{c}{2}\,p, \tag{57}$$

using $a^2 \le c/(n\eta)$. Fano's inequality in the form of Wainwright (2019, Proposition 15.12), with ϱ the Frobenius distance between projectors and $\Psi(u) = u^2$, then bounds the minimax risk below by

$\frac{a^2 p}{64}\big(1 - (\frac{c}{2}p + \log 2)/(c_1 p)\big)$. Choosing $c \le c_1/2$ and $p_0 \ge 4\log 2/c_1$ makes the bracket at least $1/4$, so the risk is at least $a^2 p/256$, which by Eq. (56) is at least $c' \min\{1, p/(n\eta)\}$.

Bounded p. Suppose $p < p_0$. Take the two-point family obtained by rotating one constraint direction into one signal direction through an angle θ and leaving the remaining directions fixed, with

$$\sin^2\theta := c \min\left\{1, \frac{1}{n\eta}\right\}. \tag{58}$$

The two projectors differ on a two-dimensional block, so $\|\mathbf{P}_0 - \mathbf{P}_\theta\|_F^2 = 2\sin^2\theta$ exactly, and Lemma 15 gives $\mathrm{KL}(\mathbb{P}_0^{\otimes n} \| \mathbb{P}_\theta^{\otimes n}) = \frac{n\eta}{2}\sin^2\theta \le \frac{c}{2}$. By Pinsker's inequality the total variation distance is at most $\sqrt{c}/2$, which is at most $1/2$ for $c \le 1$. Converting an estimator into the test that reports the nearer projector, an error of that test forces distance at least $D/2$ from the truth, where $D^2 := \|\mathbf{P}_0 - \mathbf{P}_\theta\|_F^2 = 2\sin^2\theta$, so the standard two-point reduction (Wainwright, 2019, Eq. (15.14)) gives minimax risk at least

$$\frac{D^2}{8}\big(1 - \mathrm{TV}(\mathbb{P}_0^{\otimes n}, \mathbb{P}_\theta^{\otimes n})\big) \;\ge\; \frac{D^2}{16} \;=\; \frac{\sin^2\theta}{8} \;\ge\; c'' \min\{1, 1/(n\eta)\}. \tag{59}$$

Since $p < p_0$ with p_0 absolute, this is $\ge c'' \min\{1, p/(p_0 n\eta)\}$, which is the same bound up to an absolute constant. This regime includes $m = 1, d = 2$.

Conclusion. Combining the two regimes, and recalling $\eta = \kappa^2/(1+\kappa)$,

$$\inf_{\widehat{\mathcal{C}}} \sup \mathbb{E}\left[\left\|\mathbf{P}_{\widehat{\mathcal{C}}} - \mathbf{P}_{\mathcal{C}}\right\|_F^2\right] \;\ge\; c_0 \min\left\{1, \frac{p(1+\kappa)}{n\kappa^2}\right\} \;\ge\; c_0 \min\left\{1, \frac{p}{n\kappa^2}\right\}, \tag{60}$$

with $p = m(d-m)$ in part (i) and $p = m(d-m-s)$ in part (ii). This is Eqs. (44) and (46).

The singular hypotheses lie in the class. Fix $\mathcal{N}$ with $\dim\mathcal{N} = s$ and set $\mathbf{\Sigma}_\epsilon := \sigma^2 \mathbf{P}_{\mathcal{N}^\perp}$, so $\mathrm{Null}(\mathbf{\Sigma}_\epsilon) = \mathcal{N}$ and $\mathrm{rank}(\mathbf{\Sigma}_\epsilon) = d - s$. The graph construction runs inside $\mathcal{N}^\perp \cong \mathbb{R}^{d-s}$, giving $\mathcal{C}_{\mathbf{T}} \subseteq \mathcal{N}^\perp$ with $\dim\mathcal{C}_{\mathbf{T}} = m$, which needs $m \le d - s - 1$. With $\mathbf{\Sigma}_x^{(\mathbf{T})} = \kappa\sigma^2 \mathbf{P}_{\mathcal{C}_{\mathbf{T}}^\perp}$, the complement taken in $\mathbb{R}^d$, we have $\mathrm{Null}(\mathbf{\Sigma}_x^{(\mathbf{T})}) = \mathcal{C}_{\mathbf{T}}$ of dimension m and $\mathrm{rank}(\mathbf{\Sigma}_x^{(\mathbf{T})}) = d - m$, as the model requires. For $\mathbf{v} \ne \mathbf{0}$,

$$\mathbf{v}^\top \mathbf{\Sigma}_y^{(\mathbf{T})} \mathbf{v} = \kappa\sigma^2 \left\|\mathbf{P}_{\mathcal{C}_{\mathbf{T}}^\perp}\mathbf{v}\right\|^2 + \sigma^2 \left\|\mathbf{P}_{\mathcal{N}^\perp}\mathbf{v}\right\|^2, \tag{61}$$

which vanishes only if $\mathbf{v} \in \mathcal{C}_{\mathbf{T}} \cap \mathcal{N} = \{\mathbf{0}\}$, so $\mathbf{\Sigma}_y^{(\mathbf{T})} \succ 0$ and Assumption 2 holds. Since $\mathcal{C}_{\mathbf{T}} \subseteq \mathcal{N}^\perp$ gives $\mathcal{N} \subseteq \mathcal{C}_{\mathbf{T}}^\perp$, the noise-free directions carry signal and no noise. Finally $\mathbf{P}_{\mathcal{C}_{\mathbf{T}}}$ vanishes on $\mathcal{N}$ and maps $\mathcal{N}^\perp$ into itself, so $\mathbf{\Sigma}_y^{(\mathbf{T})}$ is block diagonal for $\mathbb{R}^d = \mathcal{N} \oplus \mathcal{N}^\perp$ with the block $\kappa\sigma^2\mathbf{I}_s$ on $\mathcal{N}$ independent of $\mathbf{T}$. The divergence is therefore exactly that of the $(d-s)$-dimensional blocks, and Lemma 15 applies with $D = d - s$. □

H Proof of Theorem 4

Theorem 4 (Invertibility of the moment map). *Fix a symmetric support $\mathcal{S}$ with $q = |\mathcal{S}|$ and write $\mathbf{a}_k$ for the k-th column of $\mathbf{A}$.*

(i) *If $q > m(m+1)/2$ then $\Phi_{\mathbf{A},\mathcal{S}}$ is not injective for any $\mathbf{A}$, so the solution set of the moment equation $\mathbf{A}\mathbf{\Sigma}_y\mathbf{A}^\top = \mathbf{A}\mathbf{\Sigma}_\epsilon\mathbf{A}^\top$ is an affine set of dimension at least $q - m(m+1)/2$ and $\boldsymbol{\theta}$ is not identified within the unrestricted space $\mathbb{R}^q$ of supported symmetric matrices. If moreover there are $\mathbf{v} \in \mathrm{Null}(\Phi_{\mathbf{A},\mathcal{S}}) \setminus \{\mathbf{0}\}$ and $\varepsilon > 0$ with $\mathrm{mat}_{\mathcal{S}}(\boldsymbol{\theta}^\star \pm \varepsilon\mathbf{v}) \in \mathcal{K}$, then the two distinct admissible covariances $\mathrm{mat}_{\mathcal{S}}(\boldsymbol{\theta}^\star \pm \varepsilon\mathbf{v})$ generate the same moment equation, so no estimator using that equation alone recovers $\mathbf{\Sigma}_\epsilon$ uniformly over the class at any sample size. Relative interior membership of $\boldsymbol{\theta}^\star$ is not on its own sufficient for this: it makes $\boldsymbol{\theta}^\star \pm \varepsilon\mathbf{v}$ admissible for directions $\mathbf{v}$ in $\mathrm{lin}(\mathcal{K} - \mathcal{K})$, the linear span of the difference set $\{\mathbf{M} - \mathbf{M}' : \mathbf{M}, \mathbf{M}' \in \mathcal{K}\}$, which is the smallest subspace containing every direction one can move in from a relative-interior point of $\mathcal{K}$ and stay in the affine hull of $\mathcal{K}$, so what is needed is $\mathrm{Null}(\Phi_{\mathbf{A},\mathcal{S}}) \cap \mathrm{lin}(\mathcal{K} - \mathcal{K}) \ne \{\mathbf{0}\}$ as well. The discussion after Theorem 4 exhibits a support for which $\mathcal{K}$ is a ray, $\mathrm{lin}(\mathcal{K} - \mathcal{K})$ is one-dimensional, and the kernel misses it entirely.*

(ii) *Injectivity of $\Phi_{\mathbf{A},\mathcal{S}}$ holds for Lebesgue-almost every $\mathbf{A} \in \mathbb{R}^{m\times d}$ if and only if the polynomial $p_{\mathcal{S}}(\mathbf{A}) := \det(\mathbf{D}_{\mathbf{A}}^{\mathsf{T}}\mathbf{D}_{\mathbf{A}})$ is not identically zero, where $\mathbf{D}_{\mathbf{A}}$ has columns $\operatorname{vech}(\mathbf{a}_i\mathbf{a}_j^{\mathsf{T}} + \mathbf{a}_j\mathbf{a}_i^{\mathsf{T}})$ for $(i,j) \in \mathcal{S}$ with $i < j$ and $\operatorname{vech}(\mathbf{a}_i\mathbf{a}_i^{\mathsf{T}})$ for $(i,i) \in \mathcal{S}$.*

(iii) *The condition $p_{\mathcal{S}} \not\equiv 0$ holds whenever $q \le m(m+1)/2$ and either $\mathcal{S}$ is contained in the diagonal, or the pairs of $\mathcal{S}$ use pairwise disjoint indices. For a general $\mathcal{S}$, evaluating $p_{\mathcal{S}}$ at a single $\mathbf{A}$ and obtaining a nonzero value certifies $p_{\mathcal{S}} \not\equiv 0$ deterministically.*

Proof. Notation. For $(i,j) \in \mathcal{S}$ define $\mathbf{J}_{ij} := \mathbf{e}_i\mathbf{e}_j^{\mathsf{T}} + \mathbf{e}_j\mathbf{e}_i^{\mathsf{T}}$ if $i < j$ and $\mathbf{J}_{ii} := \mathbf{e}_i\mathbf{e}_i^{\mathsf{T}}$, so that $\operatorname{mat}_{\mathcal{S}}(\boldsymbol{\theta}) = \sum_{(i,j)\in\mathcal{S}} \theta_{ij}\mathbf{J}_{ij}$. Write $\mathbf{a}_k$ for the k-th column of $\mathbf{A}$ and $\mathbf{a}_i \odot \mathbf{a}_j := \mathbf{A}\mathbf{J}_{ij}\mathbf{A}^{\mathsf{T}}$, which equals $\mathbf{a}_i\mathbf{a}_j^{\mathsf{T}} + \mathbf{a}_j\mathbf{a}_i^{\mathsf{T}}$ for $i < j$ and $\mathbf{a}_i\mathbf{a}_i^{\mathsf{T}}$ for $i = j$.

Part (i). By linearity, $\Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta}) = \sum_{(i,j)\in\mathcal{S}} \theta_{ij}\,\mathbf{a}_i \odot \mathbf{a}_j$, so $\Phi_{\mathbf{A},\mathcal{S}}$ is injective if and only if the q matrices $\{\mathbf{a}_i \odot \mathbf{a}_j\}_{(i,j)\in\mathcal{S}}$ are linearly independent in $\mathbb{S}^m$. Since $\dim \mathbb{S}^m = m(m+1)/2$, more than $m(m+1)/2$ of them are dependent for every $\mathbf{A}$. This proves (i) and shows the counting condition is necessary and not an artifact of any particular estimator.

Part (ii). Injectivity is equivalent to $\mathbf{D}_{\mathbf{A}}$ having full column rank, hence to $p_{\mathcal{S}}(\mathbf{A}) = \det(\mathbf{D}_{\mathbf{A}}^{\mathsf{T}}\mathbf{D}_{\mathbf{A}}) \neq 0$. Each entry of $\mathbf{D}_{\mathbf{A}}$ is a quadratic monomial in the entries of $\mathbf{A}$, so $p_{\mathcal{S}}$ is a polynomial in the md entries of $\mathbf{A}$. If $p_{\mathcal{S}} \equiv 0$ then injectivity fails everywhere. If $p_{\mathcal{S}} \not\equiv 0$ then its zero set has Lebesgue measure zero, since the zero set of a nonzero real polynomial in N variables is Lebesgue null, which follows by induction on N together with Fubini's theorem. This proves (ii).

Part (iii), diagonal supports. Let $\mathcal{S} \subseteq \{(i,i)\}$ with $q \le m(m+1)/2$, and put $\mathcal{V} := \{\mathbf{a}\mathbf{a}^{\mathsf{T}} : \mathbf{a} \in \mathbb{R}^m\}$, the set of rank-one squares.

First, $\mathcal{V}$ spans $\mathbb{S}^m$: it contains every $\mathbf{e}_i\mathbf{e}_i^{\mathsf{T}}$, and

$$\mathbf{J}_{ij} \;=\; (\mathbf{e}_i + \mathbf{e}_j)\,(\mathbf{e}_i + \mathbf{e}_j)^{\mathsf{T}} - \mathbf{e}_i\mathbf{e}_i^{\mathsf{T}} - \mathbf{e}_j\mathbf{e}_j^{\mathsf{T}} \;\in\; \operatorname{span}(\mathcal{V}), \tag{62}$$

so $\operatorname{span}(\mathcal{V})$ contains a basis of $\mathbb{S}^m$; equivalently, $\mathcal{V}$ lies in no proper linear subspace of $\mathbb{S}^m$.

Second, q generic points of $\mathcal{V}$ are linearly independent whenever $q \le \dim \mathbb{S}^m$. We induct on q. For $q = 1$ a generic point is nonzero. Assume $\mathbf{a}_1\mathbf{a}_1^{\mathsf{T}}, \ldots, \mathbf{a}_{q-1}\mathbf{a}_{q-1}^{\mathsf{T}}$ are independent and write V for their span, so $\dim V = q - 1 < \dim \mathbb{S}^m$ and V is a proper subspace. By the previous paragraph $\mathcal{V} \not\subseteq V$, so

$$\{\mathbf{a} \in \mathbb{R}^m : \mathbf{a}\mathbf{a}^{\mathsf{T}} \notin V\} \;=\; \mathbb{R}^m \setminus \{\mathbf{a} : \mathbf{a}\mathbf{a}^{\mathsf{T}} \in V\} \tag{63}$$

is the complement of a proper algebraic subset, hence nonempty and generic; any such $\mathbf{a}$ extends the independent family. The resulting $\mathbf{A}$ is a witness at which $p_{\mathcal{S}} \neq 0$, so $p_{\mathcal{S}} \not\equiv 0$.

Part (iii), supports with disjoint pairs. Suppose the pairs of $\mathcal{S}$ use pairwise disjoint indices. Then each $\mathbf{a}_i \odot \mathbf{a}_j$ depends on a group of columns of $\mathbf{A}$ disjoint from every other, so the columns entering different elements can be perturbed independently. The same induction applies, using at each step that the image of the map $(\mathbf{a}, \mathbf{b}) \mapsto \mathbf{a} \odot \mathbf{b}$ spans $\mathbb{S}^m$ and is therefore not contained in the span V of the previously chosen elements.

Part (iii), the certificate. For a general $\mathcal{S}$ the two arguments above do not combine, because an index may appear in several pairs and the corresponding elements cannot be perturbed independently. In that case $p_{\mathcal{S}} \not\equiv 0$ remains a property of $\mathcal{S}$ alone and is certified by a single evaluation: if $p_{\mathcal{S}}(\mathbf{A}_0) \neq 0$ for one matrix $\mathbf{A}_0$, then $p_{\mathcal{S}}$ is not the zero polynomial, and part (ii) applies. This certificate is exact rather than probabilistic, and we verified it for every support used in Section 4. □

I Invertible Noise Covariance

Throughout this appendix $\boldsymbol{\Sigma}_\epsilon \succ 0$ with Cholesky factor $\boldsymbol{\Sigma}_\epsilon = \mathbf{L}\mathbf{L}^{\mathsf{T}}$, $\mathbf{M} = \mathbf{L}^{-1}\boldsymbol{\Sigma}_y\mathbf{L}^{-\mathsf{T}}$, $\mathbf{z}^{(t)} = \mathbf{L}^{-1}\mathbf{y}^{(t)}$ and $\widehat{\mathbf{M}} = \frac{1}{n}\sum_t \mathbf{z}^{(t)}\mathbf{z}^{(t)\mathsf{T}}$. Assumption 1 applied to $\mathbf{u} = \mathbf{L}^{-\mathsf{T}}\mathbf{w}$ gives $\big\|\langle \mathbf{z}^{(t)}, \mathbf{w}\rangle\big\|_{\psi_2} \le K\sqrt{\mathbf{w}^{\mathsf{T}}\mathbf{M}\mathbf{w}}$, so $\mathbf{z}^{(t)}$ is sub-Gaussian relative to its own covariance with the same constant and Lemma 12 applies to it with $D = d$.

These statements are the $s = 0$ specializations of Section 3, written in scaled coordinates. We give them in full because this is the standard formulation for invertible $\mathbf{\Sigma}_\epsilon$ and because the scaled gap ε has no singular counterpart.

I.1 Subspace consistency

Theorem 17 (Subspace consistency when $\mathbf{\Sigma}_\epsilon$ is invertible)**.** *Let $\mathbf{P}_{\widehat{\mathbf{U}}}$ and $\mathbf{P}_{\mathbf{U}}$ denote the projectors onto the bottom-m eigenspaces of $\widehat{\mathbf{M}}$ and $\mathbf{M}$. Under Assumptions 1 to 3 with $\mathbf{\Sigma}_\epsilon \succ 0$, there are absolute constants $C_1, C_2, C_3 > 0$ such that whenever $n \geq C_3 K^4 \|\mathbf{M}\|_{\mathrm{op}}^2 (d + \log(1/\delta))/\varepsilon^2$, with probability at least $1 - \delta$,*

$$\left\|\mathbf{P}_{\widehat{\mathbf{U}}} - \mathbf{P}_{\mathbf{U}}\right\|_{\mathrm{op}} \leq \frac{K^2 \|\mathbf{M}\|_{\mathrm{op}}}{\varepsilon} \left[C_1 \sqrt{\frac{d + \log(1/\delta)}{n}} + C_2 \frac{d + \log(1/\delta)}{n}\right]. \tag{64}$$

Proof. By Lemma 9 the eigenspace of $\mathbf{M}$ at eigenvalue 1 is exactly $\mathbf{L}^\mathsf{T}\mathcal{C}$, of dimension m, and every other eigenvalue is at least $1 + \varepsilon$. The population spectrum is therefore

$$\underbrace{\mu_{r+1} = \cdots = \mu_d = 1}_{m \text{ constraint directions}} \qquad \text{and} \qquad \underbrace{\mu_j \geq 1 + \varepsilon \ \ (j \leq r)}_{r \text{ signal directions}}, \tag{65}$$

which is the shifted configuration covered by Lemma 13 with $\theta_0 = 1$ and $\gamma = \varepsilon$. Lemma 12 applied to $\mathbf{z}^{(t)}$ bounds $\left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_{\mathrm{op}}$ by the bracket in Eq. (64), and the sample-size condition makes that bracket at most $\varepsilon/2$, so Lemma 13 applies and yields Eq. (64). □

I.2 Rank consistency

Theorem 18 (Rank consistency)**.** *Let $\mathbf{\Sigma}_\epsilon \succ 0$, let $\hat{\mu}_1 \geq \cdots \geq \hat{\mu}_d$ be the eigenvalues of $\widehat{\mathbf{M}}$, and set $\hat{m}(\tau) := |\{j : |\hat{\mu}_j - 1| \leq \tau\}|$. Fix $\tau \in (0, \varepsilon/2)$ and $\delta \in (0, 1)$. There is $C_3 > 0$ such that if $n \geq C_3 K^4 \|\mathbf{M}\|_{\mathrm{op}}^2 (d + \log(1/\delta))/\tau^2$ then $\hat{m}(\tau) = m$ with probability at least $1 - \delta$.*

Proof. By Lemma 9 the population spectrum has $\mu_{r+1} = \cdots = \mu_d = 1$ and $\mu_j \geq 1 + \varepsilon$ for $j \leq r$. By Weyl's inequality $|\hat{\mu}_j - \mu_j| \leq \left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_{\mathrm{op}}$ for every j, and the sample-size condition together with Lemma 12 gives, for a suitable absolute C_3,

$$\left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_{\mathrm{op}} \leq \frac{\tau}{2} \qquad \text{with probability at least } 1 - \delta. \tag{66}$$

On that event the two groups separate across the threshold τ:

$$\underbrace{|\hat{\mu}_j - 1| \leq \tfrac{\tau}{2} < \tau}_{m \text{ constraints: counted}} \quad \text{and} \quad \underbrace{\hat{\mu}_j \geq 1 + \varepsilon - \tfrac{\tau}{2} > 1 + \tau}_{r \text{ signal directions: not counted}}, \tag{67}$$

the second because $\tau < \varepsilon/2$ gives $\varepsilon > 2\tau$ and hence $\varepsilon - \tau/2 > 3\tau/2 > \tau$. Hence $\hat{m}(\tau) = m$. □

A threshold of the form $\tau_n = CK^2 \|\mathbf{M}\|_{\mathrm{op}} \sqrt{(d + \log n)/n}$ keeps ε out of the formula, in the spirit of information-criterion approaches to factor-number selection (Bai & Ng, 2002), but not out of the conclusion, which still requires $\tau_n < \varepsilon/2$. Nor is the rule data-driven, since K and $\|\mathbf{M}\|_{\mathrm{op}}$ are not estimated here.

I.3 Noise covariance estimation, scaled form

Theorem 19 (Noise covariance estimation when $\mathbf{\Sigma}_\epsilon$ is invertible)**.** *Let $\mathbf{\Sigma}_\epsilon \succ 0$ with support $\mathcal{S}$, $q = |\mathcal{S}| \leq m(m+1)/2$ and $\Phi_{\mathbf{A},\mathcal{S}}$ injective, and let $\mathbf{A}$ have orthonormal rows. Then with probability at least $1 - \delta$,*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2 \leq \kappa_\Phi \sqrt{m}\, \lambda_{\max}(\mathbf{\Sigma}_\epsilon)\, K^2 \|\mathbf{M}\|_{\mathrm{op}} \left[C_1 \sqrt{\frac{d + \log(1/\delta)}{n}} + C_2 \frac{d + \log(1/\delta)}{n}\right]. \tag{68}$$

Proof. As in the proof of Theorem 5, $\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|_2 \le \kappa_\Phi\sqrt{m}\,\|\mathbf{S}_y-\boldsymbol{\Sigma}_y\|_{\mathrm{op}}$. Undoing the scaling,

$$\begin{aligned}\mathbf{S}_y-\boldsymbol{\Sigma}_y \;&=\; \mathbf{L}\left(\widehat{\mathbf{M}}-\mathbf{M}\right)\mathbf{L}^{\mathsf{T}},\\ \|\mathbf{S}_y-\boldsymbol{\Sigma}_y\|_{\mathrm{op}} \;&\le\; \|\mathbf{L}\|_{\mathrm{op}}^2\left\|\widehat{\mathbf{M}}-\mathbf{M}\right\|_{\mathrm{op}} \;=\; \lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\left\|\widehat{\mathbf{M}}-\mathbf{M}\right\|_{\mathrm{op}}. \end{aligned} \tag{69}$$

Applying Lemma 12 to $\mathbf{z}^{(t)}$ to bound $\left\|\widehat{\mathbf{M}}-\mathbf{M}\right\|_{\mathrm{op}}$ then gives Eq. (68). □

The product $\lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\,\|\mathbf{M}\|_{\mathrm{op}}$ is the price of measuring the perturbation after scaling: the concentration bound for the scaled covariance is scale-free, so the original scale must be restored by $\|\mathbf{L}\|_{\mathrm{op}}^2$, and $\|\mathbf{M}\|_{\mathrm{op}}$ itself diverges as $\boldsymbol{\Sigma}_\epsilon$ degenerates. Theorem 5 carries $\|\boldsymbol{\Sigma}_y\|_{\mathrm{op}}$ in place of both factors and holds for every $\boldsymbol{\Sigma}_\epsilon \succeq 0$, which is why it is the statement we prefer; Eq. (68) is recorded only because it is the form usually quoted when $\boldsymbol{\Sigma}_\epsilon$ is invertible.

J Proof of Theorem 8

J.1 The selection rule of line 3

Lemma 20 (The selection rule is a symmetric-definite eigenproblem)**.** *Let $\mathbf{S}_y \succ 0$ and $\widehat{\boldsymbol{\Sigma}}_\epsilon \succeq 0$, set $\widehat{\boldsymbol{\Sigma}}_x := \mathbf{S}_y - \widehat{\boldsymbol{\Sigma}}_\epsilon$, and write $\mathbf{G} := \mathbf{S}_y^{-1/2}\widehat{\boldsymbol{\Sigma}}_x\mathbf{S}_y^{-1/2}$ and $\mathbf{G}_\star := \mathbf{S}_y^{-1/2}\boldsymbol{\Sigma}_x\mathbf{S}_y^{-1/2}$.*

(i) $\mathbf{S}_y\mathbf{v} = \lambda\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v}$ with λ finite and nonzero if and only if

$$\widehat{\boldsymbol{\Sigma}}_x\,\mathbf{v} = \nu\,\mathbf{S}_y\,\mathbf{v}, \qquad \nu = 1-1/\lambda, \tag{70}$$

and the infinite eigenvalues of $(\mathbf{S}_y, \widehat{\boldsymbol{\Sigma}}_\epsilon)$ are exactly the eigenvalues $\nu = 1$ of $(\widehat{\boldsymbol{\Sigma}}_x, \mathbf{S}_y)$. The pencil $(\widehat{\boldsymbol{\Sigma}}_x, \mathbf{S}_y)$ is symmetric definite, so it has d real eigenvalues $\nu_1 \le \cdots \le \nu_d$ and a basis of $\mathbf{S}_y$-orthogonal eigenvectors, namely the eigendecomposition of $\mathbf{G}$.

(ii) Every $\nu_j \le 1$. Consequently every generalized eigenvalue of $(\mathbf{S}_y, \widehat{\boldsymbol{\Sigma}}_\epsilon)$ lies in $(0, +\infty]$, and

$$|\lambda_j - 1| \;=\; \mathrm{h}(\nu_j), \qquad \mathrm{h}(\nu) := \frac{|\nu|}{1-\nu}\;\;(\nu<1), \qquad \mathrm{h}(1) := +\infty. \tag{71}$$

(iii) (Selection.) *Suppose $\boldsymbol{\Sigma}_x \succeq 0$ has $\mathrm{Null}(\boldsymbol{\Sigma}_x) = \mathcal{C}$ with $\dim\mathcal{C} = m < d$, let $g > 0$ be the smallest nonzero eigenvalue of $\mathbf{G}_\star$, and let $\eta \ge \|\mathbf{G}-\mathbf{G}_\star\|_{\mathrm{op}}$. If $\eta < g/2$, then exactly m eigenvalues of $\mathbf{G}$ satisfy $|\nu| \le \eta$ and the remaining $d-m$ satisfy $\nu \ge g-\eta > \eta$; the former are both the m eigenvalues of $\mathbf{G}$ smallest in absolute value and the m finite generalized eigenvalues of $(\mathbf{S}_y, \widehat{\boldsymbol{\Sigma}}_\epsilon)$ closest to 1. In particular line 3 of Algorithm 1 returns $\mathbf{S}_y^{-1/2}\widehat{\mathcal{U}}$, where $\widehat{\mathcal{U}}$ is the invariant subspace of $\mathbf{G}$ for those m eigenvalues.*

Proof. (i) Suppose $\mathbf{S}_y\mathbf{v} = \lambda\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v}$ with $\lambda \notin \{0,\infty\}$. Then $\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v} = \lambda^{-1}\mathbf{S}_y\mathbf{v}$, so $\widehat{\boldsymbol{\Sigma}}_x\mathbf{v} = \mathbf{S}_y\mathbf{v} - \widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v} = (1-\lambda^{-1})\mathbf{S}_y\mathbf{v}$, which is Eq. (70). Conversely, if $\widehat{\boldsymbol{\Sigma}}_x\mathbf{v} = \nu\mathbf{S}_y\mathbf{v}$ with $\nu \ne 1$, then $\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v} = (1-\nu)\mathbf{S}_y\mathbf{v}$ with $1-\nu \ne 0$, so $\mathbf{S}_y\mathbf{v} = (1-\nu)^{-1}\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v}$ and $\lambda = (1-\nu)^{-1}$ is finite. The case $\nu = 1$ is $\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v} = \mathbf{0}$, which is the infinite eigenvalue of $(\mathbf{S}_y, \widehat{\boldsymbol{\Sigma}}_\epsilon)$. Since $\mathbf{S}_y \succ 0$, the pencil $(\widehat{\boldsymbol{\Sigma}}_x, \mathbf{S}_y)$ is symmetric definite, and $\widehat{\boldsymbol{\Sigma}}_x\mathbf{v} = \nu\mathbf{S}_y\mathbf{v}$ is equivalent to $\mathbf{G}\mathbf{u} = \nu\mathbf{u}$ with $\mathbf{u} = \mathbf{S}_y^{1/2}\mathbf{v}$, so the eigenvalues are real and the eigenvectors are $\mathbf{S}_y$-orthogonal.

(ii) If $\widehat{\boldsymbol{\Sigma}}_x\mathbf{v} = \nu\mathbf{S}_y\mathbf{v}$ with $\mathbf{v} \ne \mathbf{0}$ then $\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v} = (1-\nu)\mathbf{S}_y\mathbf{v}$, so $\mathbf{v}^{\mathsf{T}}\widehat{\boldsymbol{\Sigma}}_\epsilon\mathbf{v} = (1-\nu)\,\mathbf{v}^{\mathsf{T}}\mathbf{S}_y\mathbf{v}$. The left side is nonnegative because $\widehat{\boldsymbol{\Sigma}}_\epsilon \succeq 0$ and the factor $\mathbf{v}^{\mathsf{T}}\mathbf{S}_y\mathbf{v}$ is strictly positive because $\mathbf{S}_y \succ 0$, so $\nu \le 1$. Hence $\lambda = 1/(1-\nu) \in (0,+\infty]$, and for $\nu < 1$, $\lambda - 1 = 1/(1-\nu) - 1 = \nu/(1-\nu)$, which gives Eq. (71).

(iii) Since $\mathbf{S}_y \succ 0$, $\mathbf{G}_\star \succeq 0$ with $\mathrm{Null}(\mathbf{G}_\star) = \mathbf{S}_y^{1/2}\mathcal{C}$ of dimension m, so its eigenvalues are m zeros followed by $d-m$ values at least g. Weyl's inequality applied to $\mathbf{G} = \mathbf{G}_\star + (\mathbf{G} - \mathbf{G}_\star)$ moves each eigenvalue by at most η, so the spectrum of $\mathbf{G}$ splits into

$$\underbrace{\nu \in [-\eta,\ \eta]}_{m \text{ "target" eigenvalues}} \qquad \text{and} \qquad \underbrace{\nu \in [g-\eta,\ 1]}_{d-m \text{ "non-target" eigenvalues}}, \tag{72}$$

the upper limit 1 coming from part (ii). The hypothesis $\eta < g/2$ gives $g - \eta > \eta$, so the two intervals are disjoint and the first group is exactly the set of m eigenvalues smallest in absolute value.

It remains to check that the *same* m indices minimize $|\lambda - 1| = \mathrm{h}(\nu)$. Recall from Eq. (71) the three properties of h we use:

$$\mathrm{h}(0) = 0; \qquad \mathrm{h} \text{ is strictly increasing on } [0,1] \text{ with } \mathrm{h}(1) = +\infty;$$
$$\mathrm{h}(-t) = \frac{t}{1+t} \ < \ \frac{t}{1-t} = \mathrm{h}(t) \qquad \text{for } t \in (0,1). \tag{73}$$

First, $\eta < 1$. Indeed $m < d$, so a non-target eigenvalue exists; by Eq. (72) it satisfies $g - \eta \le \nu \le 1$, whence

$$\eta \ < \ g - \eta \ \le \ 1. \tag{74}$$

Now compare the two groups under h. Every target has $|\nu| \le \eta < 1$, and the third property in Eq. (73) says h is larger on the positive side, so the maximum of h over $[-\eta, \eta]$ is attained at $\nu = \eta$:

$$\mathrm{h}(\nu) \ \le \ \mathrm{h}(\eta) \qquad \text{for every target } \nu. \tag{75}$$

Every non-target has $\nu \in [g-\eta, 1]$, where h is increasing, so

$$\mathrm{h}(\nu) \ \ge \ \mathrm{h}(g-\eta) \qquad \text{for every non-target } \nu, \tag{76}$$

with the value $+\infty$ for a non-target sitting at $\nu = 1$. Finally, Eq. (74) puts both η and $g - \eta$ in $[0,1]$ with $\eta < g - \eta$, and h is strictly increasing there, so

$$\max_{\text{targets}} \mathrm{h}(\nu) \ \le \ \mathrm{h}(\eta) \ < \ \mathrm{h}(g-\eta) \ \le \ \min_{\text{non-targets}} \mathrm{h}(\nu). \tag{77}$$

The m targets therefore have strictly the m smallest values of $|\lambda - 1|$, and their λ are finite because $|\nu| \le \eta < 1$. □

Remark 4 (The hypothesis of part (iii) cannot be dropped.). Globally the two orderings differ, because h weights the two sides of 0 unequally. Two eigenvalues $\nu = 0.5$ and $\nu = -0.8$ have $|\nu|$ equal to 0.5 and 0.8, but $|\lambda - 1|$ equal to 1 and $4/9$, so the two rules disagree about which is closer to the constraint cluster. Part (iii) rules this out by confining the targets to $[-\eta, \eta]$ and the rest to $[g-\eta, 1]$ with $\eta < g - \eta$, which is the only regime in which the lemma is invoked. Both proofs that use it verify this condition explicitly.

Lemma 21 (Monotonicity of pencil eigenvalues in the weight.). *Let $\mathbf{A} \in \mathbb{S}^d$ with $\mathbf{A} \succeq 0$, let $\mathbf{B}, \mathbf{B}' \succ 0$, and suppose $\mathbf{B} \preceq c\,\mathbf{B}'$ for some $c > 0$. Write $\nu_1(\mathbf{A},\mathbf{B}) \le \cdots \le \nu_d(\mathbf{A},\mathbf{B})$ for the eigenvalues of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$. Then $\nu_j(\mathbf{A},\mathbf{B}) \ge \nu_j(\mathbf{A},\mathbf{B}')/c$ for every j.*

Proof. Substituting $\mathbf{u} = \mathbf{B}^{1/2}\mathbf{v}$ in the Courant-Fischer characterization of the eigenvalues of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, and using that $\mathbf{v} \mapsto \mathbf{B}^{1/2}\mathbf{v}$ is a bijection on j-dimensional subspaces,

$$\nu_j(\mathbf{A},\mathbf{B}) = \min_{\dim \mathcal{V} = j} \ \max_{\mathbf{v} \in \mathcal{V}\setminus\{\mathbf{0}\}} f_{\mathbf{B}}(\mathbf{v}), \qquad f_{\mathbf{B}}(\mathbf{v}) := \frac{\mathbf{v}^\mathsf{T}\mathbf{A}\mathbf{v}}{\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v}}. \tag{78}$$

The hypothesis gives $\mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v} \le c\,\mathbf{v}^\mathsf{T}\mathbf{B}'\mathbf{v}$ for every $\mathbf{v}$, hence $f_{\mathbf{B}} \ge f_{\mathbf{B}'}/c$ pointwise. Taking the maximum over each $\mathcal{V}$ and then the minimum over $\mathcal{V}$ preserves the inequality. □

Lemma 22 (Rule-P subspace perturbation). *Let $\mathbf{S}_y \succ 0$, let $\widehat{\mathbf{\Sigma}_\epsilon} \succeq 0$, and write $\mathbf{G} := \mathbf{S}_y^{-1/2}(\mathbf{S}_y - \widehat{\mathbf{\Sigma}_\epsilon})\mathbf{S}_y^{-1/2}$ and $\mathbf{G}_\star := \mathbf{S}_y^{-1/2}\mathbf{\Sigma}_x\mathbf{S}_y^{-1/2}$, and let g be the smallest nonzero eigenvalue of $\mathbf{G}_\star$. If $\eta := \|\mathbf{G} - \mathbf{G}_\star\|_{\mathrm{op}} < g/2$, then the subspace $\widehat{\mathcal{C}}$ returned by line 3 of Algorithm 1 satisfies*

$$\left\|\mathbf{P}_{\widehat{\mathcal{C}}} - \mathbf{P}_{\mathcal{C}}\right\|_{\mathrm{op}} \ \le \ \frac{2\kappa(\mathbf{S}_y)^{1/2}}{g}\,\eta \ \le \ \beta\,\left\|(\mathbf{S}_y - \widehat{\mathbf{\Sigma}_\epsilon}) - \mathbf{\Sigma}_x\right\|_{\mathrm{op}}, \qquad \beta := \frac{2\kappa(\mathbf{S}_y)^{3/2}}{\sigma_{\mathcal{C}}^2}. \tag{79}$$

Moreover, on the event $\Delta_n \le \lambda_{\min}(\mathbf{\Sigma}_y)/2$ *of Eq.* (24)*,*

$$\beta \;\le\; \beta_\star := \frac{6\sqrt{3}\,\kappa(\mathbf{\Sigma}_y)^{3/2}}{\sigma_{\mathcal{C}}^2}, \tag{80}$$

a deterministic function of the population covariances, and

$$\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,\beta_\star \;\ge\; 6\sqrt{3}. \tag{81}$$

Proof. The population matrix. $\mathbf{G}_\star$ is positive semidefinite, and

$$\mathbf{G}_\star\mathbf{u} = \mathbf{0} \iff \mathbf{\Sigma}_x\mathbf{S}_y^{-1/2}\mathbf{u} = \mathbf{0} \iff \mathbf{S}_y^{-1/2}\mathbf{u} \in \mathcal{C}, \qquad \text{so} \qquad \mathrm{Null}(\mathbf{G}_\star) = \mathbf{S}_y^{1/2}\mathcal{C}. \tag{82}$$

Its eigenvalues are therefore m zeros followed by $d-m$ values at least g.

The first inequality. By Lemma 20(iii), line 3 of Algorithm 1 returns $\widehat{\mathcal{C}} = \mathbf{S}_y^{-1/2}\widehat{\mathcal{U}}$, where $\widehat{\mathcal{U}}$ is the invariant subspace of $\mathbf{G}$ for its m eigenvalues smallest in absolute value; under $\eta < g/2$ these are also its m algebraically smallest. Applying Lemma 13 with $\gamma = g$ and then Lemma 14 with $\mathbf{W} = \mathbf{S}_y^{-1/2}$, for which $\kappa(\mathbf{W}) = \kappa(\mathbf{S}_y)^{1/2}$, gives

$$\begin{aligned}\left\|\mathbf{P}_{\widehat{\mathcal{U}}} - \mathbf{P}_{\mathbf{S}_y^{1/2}\mathcal{C}}\right\|_{\mathrm{op}} \;&\le\; \frac{2\eta}{g},\\ \left\|\mathbf{P}_{\widehat{\mathcal{C}}} - \mathbf{P}_{\mathcal{C}}\right\|_{\mathrm{op}} \;&\le\; \kappa(\mathbf{S}_y)^{1/2}\left\|\mathbf{P}_{\widehat{\mathcal{U}}} - \mathbf{P}_{\mathbf{S}_y^{1/2}\mathcal{C}}\right\|_{\mathrm{op}} \;\le\; \frac{2\kappa(\mathbf{S}_y)^{1/2}}{g}\,\eta.\end{aligned} \tag{83}$$

The second inequality. Two bounds are needed. Conjugation by $\mathbf{S}_y^{-1/2}$ inflates the operator norm by at most $\lambda_{\min}(\mathbf{S}_y)^{-1}$, so

$$\eta \;=\; \left\|\mathbf{S}_y^{-1/2}\big[(\mathbf{S}_y - \widehat{\mathbf{\Sigma}_\epsilon}) - \mathbf{\Sigma}_x\big]\mathbf{S}_y^{-1/2}\right\|_{\mathrm{op}} \;\le\; \frac{\left\|(\mathbf{S}_y - \widehat{\mathbf{\Sigma}_\epsilon}) - \mathbf{\Sigma}_x\right\|_{\mathrm{op}}}{\lambda_{\min}(\mathbf{S}_y)}. \tag{84}$$

For the gap, apply Lemma 21 with $\mathbf{A} = \mathbf{\Sigma}_x$, $\mathbf{B} = \mathbf{S}_y$, $\mathbf{B}' = \mathbf{I}$ and $c = \|\mathbf{S}_y\|_{\mathrm{op}}$, admissible because $\mathbf{S}_y \preceq \|\mathbf{S}_y\|_{\mathrm{op}}\mathbf{I}$:

$$g \;=\; \nu_{m+1}(\mathbf{\Sigma}_x, \mathbf{S}_y) \;\ge\; \frac{\nu_{m+1}(\mathbf{\Sigma}_x, \mathbf{I})}{\|\mathbf{S}_y\|_{\mathrm{op}}} \;=\; \frac{\sigma_{\mathcal{C}}^2}{\|\mathbf{S}_y\|_{\mathrm{op}}}, \tag{85}$$

where the last equality holds because $\nu_j(\mathbf{\Sigma}_x, \mathbf{I})$ is the jth smallest eigenvalue of $\mathbf{\Sigma}_x$, and $\mathbf{\Sigma}_x$ has exactly m zero eigenvalues followed by $\lambda_{m+1}(\mathbf{\Sigma}_x) = \sigma_{\mathcal{C}}^2$. Substituting Eqs. (84) and (85) into Eq. (83),

$$\left\|\mathbf{P}_{\widehat{\mathcal{C}}} - \mathbf{P}_{\mathcal{C}}\right\|_{\mathrm{op}} \;\le\; \underbrace{\frac{2\kappa(\mathbf{S}_y)^{1/2}\,\|\mathbf{S}_y\|_{\mathrm{op}}}{\sigma_{\mathcal{C}}^2\,\lambda_{\min}(\mathbf{S}_y)}}_{=\;2\kappa(\mathbf{S}_y)^{3/2}/\sigma_{\mathcal{C}}^2\;=\;\beta} \left\|(\mathbf{S}_y - \widehat{\mathbf{\Sigma}_\epsilon}) - \mathbf{\Sigma}_x\right\|_{\mathrm{op}}. \tag{86}$$

The two bounds on β*.* On the event $\Delta_n \le \lambda_{\min}(\mathbf{\Sigma}_y)/2$ of Eq. (24) one has $\kappa(\mathbf{S}_y) \le 3\kappa(\mathbf{\Sigma}_y)$, hence Eq. (80):

$$\beta \;=\; \frac{2\kappa(\mathbf{S}_y)^{3/2}}{\sigma_{\mathcal{C}}^2} \;\le\; \frac{2\cdot 3^{3/2}\,\kappa(\mathbf{\Sigma}_y)^{3/2}}{\sigma_{\mathcal{C}}^2} \;=\; \frac{6\sqrt{3}\,\kappa(\mathbf{\Sigma}_y)^{3/2}}{\sigma_{\mathcal{C}}^2} \;=\; \beta_\star. \tag{87}$$

Finally, $\sigma_{\mathcal{C}}^2 = \lambda_{\min}(\mathbf{\Sigma}_x|_{\mathcal{C}^\perp}) \le \|\mathbf{\Sigma}_x\|_{\mathrm{op}}$ and $\kappa(\mathbf{\Sigma}_y) \ge 1$, so

$$\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,\beta_\star \;=\; 6\sqrt{3}\,\kappa(\mathbf{\Sigma}_y)^{3/2}\,\frac{\|\mathbf{\Sigma}_x\|_{\mathrm{op}}}{\sigma_{\mathcal{C}}^2} \;\ge\; 6\sqrt{3}, \tag{88}$$

which is Eq. (81). □

The proof uses two facts about how the moment map responds to a perturbation of its constraint matrix. We isolate them because the basin condition in Eq. (94) depends on their constants.

Lemma 23 (Size and stability of the moment map.)**.** *Let $\mathbf{A}, \mathbf{A}' \in \mathbb{R}^{m\times d}$ have orthonormal rows, let $\mathbf{R} \in \mathrm{O}(m)$, and set $e := \|\mathbf{A}' - \mathbf{R}\mathbf{A}\|_{\mathrm{op}}$. For every symmetric support $\mathcal{S}$:*

(i) $\sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}}) \le \sqrt{2}$*, hence* $\kappa_\Phi \ge 1/\sqrt{2}$ *and* $\sqrt{m}\,\kappa_\Phi \ge 1/\sqrt{2}$*;*

(ii) $\sigma_{\min}(\Phi_{\mathbf{A}',\mathcal{S}}) \ge \sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}}) - \sqrt{2}\,e\,(2+e)$*.*

In particular, if $e \le 1/(6\sqrt{2}\,\kappa_\Phi)$ then $e \le 1/6$ and $\sigma_{\min}(\Phi_{\mathbf{A}',\mathcal{S}}) \ge \sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}})/2$, so $\Phi_{\mathbf{A}',\mathcal{S}}$ is injective with $\left\|\Phi^+_{\mathbf{A}',\mathcal{S}}\right\|_{\mathrm{op}} \le 2\kappa_\Phi$.

Proof. Here $\sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}}) = \min_{\|\boldsymbol{\theta}\|_2=1} \|\Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta})\|_F$, and for a unit $\boldsymbol{\theta}$ we write $\mathbf{M} := \mathrm{mat}_{\mathcal{S}}(\boldsymbol{\theta})$, so that

$$\|\mathbf{M}\|_F^2 = \sum_{(i,i)\in\mathcal{S}} \theta_{ii}^2 + 2 \sum_{(i,j)\in\mathcal{S},\, i<j} \theta_{ij}^2 \;\le\; 2\,\|\boldsymbol{\theta}\|_2^2 = 2. \tag{89}$$

(i) For any unit $\boldsymbol{\theta}$, $\|\Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta})\|_F = \left\|\mathbf{A}\mathbf{M}\mathbf{A}^\mathsf{T}\right\|_F \le \|\mathbf{A}\|_{\mathrm{op}}^2\,\|\mathbf{M}\|_F \le \sqrt{2}$, using $\|\mathbf{A}\|_{\mathrm{op}} = 1$. A minimum over the unit sphere is at most the value at any single point, so $\sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}}) \le \sqrt{2}$. Since $m \ge 1$, the remaining claims follow.

(ii) First, $\Phi_{\mathbf{R}\mathbf{A},\mathcal{S}}(\boldsymbol{\theta}) = \mathbf{R}\,\Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta})\,\mathbf{R}^\mathsf{T}$, and the Frobenius norm is invariant under conjugation by an orthogonal matrix, so $\sigma_{\min}(\Phi_{\mathbf{R}\mathbf{A},\mathcal{S}}) = \sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}})$. Next, writing $\mathbf{F} := \mathbf{A}' - \mathbf{R}\mathbf{A}$ and expanding,

$$\Phi_{\mathbf{A}',\mathcal{S}}(\boldsymbol{\theta}) - \Phi_{\mathbf{R}\mathbf{A},\mathcal{S}}(\boldsymbol{\theta}) = \mathbf{F}\mathbf{M}\mathbf{F}^\mathsf{T} + \mathbf{F}\mathbf{M}\mathbf{A}^\mathsf{T}\mathbf{R}^\mathsf{T} + \mathbf{R}\mathbf{A}\mathbf{M}\mathbf{F}^\mathsf{T}. \tag{90}$$

Bound each term by $\|\mathbf{X}\mathbf{Y}\mathbf{Z}\|_F \le \|\mathbf{X}\|_{\mathrm{op}}\,\|\mathbf{Y}\|_F\,\|\mathbf{Z}\|_{\mathrm{op}}$, using $\|\mathbf{A}\|_{\mathrm{op}} = \|\mathbf{R}\|_{\mathrm{op}} = 1$ and $\|\mathbf{F}\|_{\mathrm{op}} = e$:

$$\|\Phi_{\mathbf{A}',\mathcal{S}}(\boldsymbol{\theta}) - \Phi_{\mathbf{R}\mathbf{A},\mathcal{S}}(\boldsymbol{\theta})\|_F \;\le\; \|\mathbf{M}\|_F\,\left(e^2 + 2e\right) \overset{(89)}{\le} \sqrt{2}\,e\,(2+e). \tag{91}$$

This holds for every unit $\boldsymbol{\theta}$, so taking minima over the unit sphere gives (ii).

For the final claim, suppose $e \le 1/(6\sqrt{2}\,\kappa_\Phi)$. Part (i) gives $\kappa_\Phi \ge 1/\sqrt{2}$, so

$$e \;\le\; \frac{1}{6\sqrt{2}\,\kappa_\Phi} \;\le\; \frac{1}{6\sqrt{2}\cdot(1/\sqrt{2})} \;=\; \frac{1}{6}, \tag{92}$$

and therefore $2 + e \le 3$, which lets us drop the quadratic term:

$$\sqrt{2}\,e\,(2+e) \;\le\; 3\sqrt{2}\,e \;\le\; 3\sqrt{2}\cdot\frac{1}{6\sqrt{2}\,\kappa_\Phi} \;=\; \frac{1}{2\kappa_\Phi} \;=\; \frac{\sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}})}{2}. \tag{93}$$

Substituting into (ii) gives $\sigma_{\min}(\Phi_{\mathbf{A}',\mathcal{S}}) \ge \sigma_{\min}(\Phi_{\mathbf{A},\mathcal{S}})/2$, hence $\left\|\Phi^+_{\mathbf{A}',\mathcal{S}}\right\|_{\mathrm{op}} \le 2\kappa_\Phi$. □

Theorem 8 (End-to-end guarantee for Algorithm 1). *Let Assumptions 1 to 3 hold with $q \le m(m+1)/2$ and $\Phi_{\mathbf{A},\mathcal{S}}$ injective. With $\beta_\star$ the deterministic constant of Eq.* (79)*, define the basin radius*

$$\rho := \frac{c_0}{\sqrt{m}\,\kappa_\Phi\,\|\boldsymbol{\Sigma}_x\|_{\mathrm{op}}\,\beta_\star^2} = \frac{c_0\,\sigma_{\mathcal{C}}^4}{108\,\sqrt{m}\,\kappa_\Phi\,\|\boldsymbol{\Sigma}_x\|_{\mathrm{op}}\,\kappa(\boldsymbol{\Sigma}_y)^3}, \qquad c_0 := \frac{1}{48}. \tag{94}$$

Suppose $\left\|\boldsymbol{\Sigma}_\epsilon^{(1)} - \boldsymbol{\Sigma}_\epsilon\right\|_{\mathrm{op}} \le \rho$ and

$$n \;\ge\; n_{\star\star}(\delta) := \max\left\{ n_\star(\delta),\; \frac{C^2K^4\,\|\boldsymbol{\Sigma}_y\|_{\mathrm{op}}^2\,\big(d + \log(1/\delta)\big)\,(1+\sqrt{m}\,\kappa_\Phi)^2}{c_2^2\,\rho^2} \right\}, \qquad c_2 := \tfrac{1}{8}, \tag{95}$$

with C the absolute constant of Eq. (24)*, so that on $\mathcal{E}_\delta$ both $\Delta_n \le \lambda_{\min}(\mathbf{\Sigma}_y)/2$ and $\Delta_n \le c_2\rho/(1+\sqrt{m}\kappa_\Phi)$ hold. Then, on $\mathcal{E}_\delta$ and hence with probability at least $1-\delta$, every iterate of Algorithm 1 run with the exact projection of line 4 satisfies*

$$\left\|\mathbf{\Sigma}_\epsilon^{(k+1)} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}} \;\le\; \tfrac{1}{2}\left\|\mathbf{\Sigma}_\epsilon^{(k)} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}} \;+\; c_1\sqrt{m}\,\kappa_\Phi\,\Delta_n, \qquad c_1 := 4, \tag{96}$$

so that, writing

$$U_n \;:=\; K^2\,\|\mathbf{\Sigma}_y\|_{\mathrm{op}}\sqrt{\frac{d+\log(1/\delta)}{n}}, \qquad T_\star(\delta) \;:=\; \log_2\left(\frac{\rho}{U_n}\right), \tag{97}$$

one has, for every $k \ge T_\star(\delta)$,

$$\begin{aligned}\left\|\mathbf{\Sigma}_\epsilon^{(k)} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}} &\lesssim \sqrt{m}\,\kappa_\Phi\,U_n,\\ \left\|\mathbf{P}_{\mathcal{C}^{(k)}} - \mathbf{P}_{\mathcal{C}}\right\|_{\mathrm{op}} &\lesssim \frac{\sqrt{m}\,\kappa_\Phi\,\kappa(\mathbf{\Sigma}_y)^{3/2}}{\sigma_{\mathcal{C}}^2}\,U_n.\end{aligned} \tag{98}$$

In particular, if the iteration budget of Algorithm 1 satisfies $T \ge T_\star(\delta)$, both bounds apply to the returned pair $(\mathbf{A}^{(T)}, \mathbf{\Sigma}_\epsilon^{(T+1)})$, since $T+1 > T \ge T_\star(\delta)$. No step uses the true $\mathbf{\Sigma}_\epsilon$ or the true $\mathbf{A}$.

Proof of Theorem 8. Notation. Write $b_k := \left\|\mathbf{\Sigma}_\epsilon^{(k)} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}}$ and $\mathbf{F}_k := \mathbf{A}^{(k)} - \mathbf{R}_k\mathbf{A}$, where $\mathbf{R}_k \in \mathrm{O}(m)$ is the Procrustes-optimal rotation aligning the two bases, and put $e_k := \|\mathbf{F}_k\|_{\mathrm{op}}$. Both $\mathbf{A}$ and $\mathbf{A}^{(k)}$ have orthonormal rows, the latter by construction in line 3 of Algorithm 1.

The only randomness is the single draw of $\mathbf{S}_y$, and the iteration is a deterministic map of $\mathbf{S}_y$, so every statement below holds on the one event of Eq. (24). No union bound over iterations is needed. Write $\beta_\star$ for the deterministic constant of Eq. (79) and set

$$\rho := \frac{c_0}{\sqrt{m}\,\kappa_\Phi\,\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,\beta_\star^2}, \qquad c_0 := \tfrac{1}{48}. \tag{99}$$

Step 0: the standing inequalities. We argue throughout on the event $\mathcal{E}_\delta$. Two facts hold there. The sample-size condition Eq. (95) was chosen precisely so that

$$\Delta_n \;\le\; \frac{c_2\,\rho}{1+\sqrt{m}\,\kappa_\Phi} \;\le\; c_2\,\rho \;\le\; \rho, \tag{100}$$

the middle step because $1+\sqrt{m}\kappa_\Phi \ge 1$ and the last because $c_2 = 1/8 \le 1$. And if the current iterate satisfies $b_k \le \rho$, then adding Eq. (100) gives

$$\Delta_n + b_k \;\le\; 2\rho. \tag{101}$$

Everything else in the proof follows from one estimate on the product $\rho\beta_\star$:

$$\rho\,\beta_\star \overset{\text{(a)}}{=} \frac{c_0}{\sqrt{m}\,\kappa_\Phi\,\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,\beta_\star} \overset{\text{(b)}}{\le} \frac{c_0}{6\sqrt{3}\,\kappa_\Phi} \overset{\text{(c)}}{=} \frac{1}{288\sqrt{3}\,\kappa_\Phi} \overset{\text{(d)}}{\le} \min\left\{\frac{1}{24\,\kappa_\Phi},\,\frac{1}{4}\right\}, \tag{102}$$

where (a) substitutes the definition of ρ, (b) uses $\sqrt{m} \ge 1$ together with $\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,\beta_\star \ge 6\sqrt{3}$ from Eq. (81), (c) is $c_0 = 1/48$, and (d) uses $288\sqrt{3} \ge 24$ for the first entry and $\kappa_\Phi \ge 1/\sqrt{2}$ from Lemma 23(i) for the second.

Two consequences will be used repeatedly. Combining Eqs. (101) and (102),

$$\beta_\star\,(\Delta_n + b_k) \;\le\; 2\rho\,\beta_\star \;\le\; \frac{1}{2}, \tag{103}$$

which is the hypothesis the selection step of Step 1 needs, and, feeding Eq. (103) through the bound on e_k in Eq. (109) below,

$$e_k \;\le\; \sqrt{2}\,\beta_\star\,(\Delta_n + b_k) \;\le\; 2\sqrt{2}\,\rho\,\beta_\star \;\le\; \frac{2\sqrt{2}}{24\,\kappa_\Phi} \;=\; \frac{1}{6\sqrt{2}\,\kappa_\Phi}, \tag{104}$$

which is exactly the range in which Lemma 23 applies. Both hypotheses used later are therefore consequences of the basin condition, not extra assumptions.

Step 1: the subspace update. Line 3 is run with $\widehat{\mathbf{\Sigma}}_\epsilon = \mathbf{\Sigma}_\epsilon^{(k)}$, not with the true $\mathbf{\Sigma}_\epsilon$, so we must check that Lemma 20(iii) applies at *every* iterate and not only at $k = 1$. Write

$$\widehat{\mathbf{\Sigma}}_x^{(k)} := \mathbf{S}_y - \mathbf{\Sigma}_\epsilon^{(k)}, \qquad \mathbf{G}_k := \mathbf{S}_y^{-1/2}\widehat{\mathbf{\Sigma}}_x^{(k)}\mathbf{S}_y^{-1/2}, \qquad \mathbf{G}_\star := \mathbf{S}_y^{-1/2}\mathbf{\Sigma}_x\mathbf{S}_y^{-1/2}. \tag{105}$$

The plug-in error splits into a sampling part and a covariance part,

$$\widehat{\mathbf{\Sigma}}_x^{(k)} - \mathbf{\Sigma}_x \;=\; \underbrace{(\mathbf{S}_y - \mathbf{\Sigma}_y)}_{\|\cdot\|_{\mathrm{op}} \le \Delta_n} \;-\; \underbrace{(\mathbf{\Sigma}_\epsilon^{(k)} - \mathbf{\Sigma}_\epsilon)}_{\|\cdot\|_{\mathrm{op}} \le b_k}, \tag{106}$$

so, conjugating by $\mathbf{S}_y^{-1/2}$ and using $\left\|\mathbf{S}_y^{-1/2}\mathbf{M}\mathbf{S}_y^{-1/2}\right\|_{\mathrm{op}} \le \|\mathbf{M}\|_{\mathrm{op}} / \lambda_{\min}(\mathbf{S}_y)$,

$$\eta_k \;:=\; \|\mathbf{G}_k - \mathbf{G}_\star\|_{\mathrm{op}} \;\le\; \frac{\Delta_n + b_k}{\lambda_{\min}(\mathbf{S}_y)}. \tag{107}$$

The eigengap bound Eq. (85) involves only $\mathbf{\Sigma}_x$ and $\mathbf{S}_y$, never $\mathbf{\Sigma}_\epsilon^{(k)}$, so $g \ge \sigma_{\mathcal{C}}^2 / \|\mathbf{S}_y\|_{\mathrm{op}}$ holds at every iterate. Chaining this with Eqs. (107) and (103), and using $\beta_\star \ge \beta = 2\kappa(\mathbf{S}_y)^{3/2}/\sigma_{\mathcal{C}}^2 \ge 2\kappa(\mathbf{S}_y)/\sigma_{\mathcal{C}}^2$ so that Eq. (103) reads $\Delta_n + b_k \le \sigma_{\mathcal{C}}^2/(4\kappa(\mathbf{S}_y))$,

$$\eta_k \;\le\; \frac{\Delta_n + b_k}{\lambda_{\min}(\mathbf{S}_y)} \;\le\; \frac{\sigma_{\mathcal{C}}^2}{4\,\kappa(\mathbf{S}_y)\,\lambda_{\min}(\mathbf{S}_y)} \;=\; \frac{\sigma_{\mathcal{C}}^2}{4\,\|\mathbf{S}_y\|_{\mathrm{op}}} \;\le\; \frac{g}{4} \;<\; \frac{g}{2}, \tag{108}$$

the equality because $\kappa(\mathbf{S}_y)\lambda_{\min}(\mathbf{S}_y) = \|\mathbf{S}_y\|_{\mathrm{op}}$. Hence Lemma 20(iii) applies for every k: line 3 returns $\mathcal{C}^{(k)} = \mathbf{S}_y^{-1/2}\widehat{\mathcal{U}}_k$, where $\widehat{\mathcal{U}}_k$ is the invariant subspace of $\mathbf{G}_k$ for its m eigenvalues smallest in absolute value, and those are also its m algebraically smallest. Lemma 22, applied with perturbation $\Delta_n + b_k$, then gives

$$\|\mathbf{P}_{\mathcal{C}^{(k)}} - \mathbf{P}_{\mathcal{C}}\|_{\mathrm{op}} \;\le\; \beta\,(\Delta_n + b_k) \;\le\; \beta_\star\,(\Delta_n + b_k)\,, \qquad e_k \;\le\; \sqrt{2}\,\beta_\star\,(\Delta_n + b_k)\,. \tag{109}$$

The second bound in Eq. (109) converts a projector distance into an aligned-basis distance, and that conversion needs the principal angles to be acute. Let $\vartheta_1, \dots, \vartheta_m$ be the principal angles between $\mathcal{C}^{(k)}$ and $\mathcal{C}$. Then

$$\sigma_i\left(\mathbf{A}^{(k)} - \mathbf{R}_k\mathbf{A}\right) = 2\sin(\vartheta_i/2), \qquad \|\mathbf{P}_{\mathcal{C}^{(k)}} - \mathbf{P}_{\mathcal{C}}\|_{\mathrm{op}} = \sin\vartheta_{\max},$$
$$2\sin\tfrac{\vartheta}{2} \;\le\; \sqrt{2}\,\sin\vartheta \qquad \text{on } \left[0, \tfrac{\pi}{2}\right], \tag{110}$$

the first at the Procrustes-optimal $\mathbf{R}_k$. The required range is available because Eqs. (109) and (102) give

$$\sin\vartheta_{\max} \;\le\; \beta_\star\,(\Delta_n + b_k) \;\le\; 2\rho\,\beta_\star \;\le\; \tfrac{1}{2}, \qquad \text{so} \qquad \vartheta_{\max} \le \tfrac{\pi}{2}. \tag{111}$$

Step 2: the covariance update. By Lemma 23 together with Eq. (104), the map $\Phi_{\mathbf{A}^{(k)},\mathcal{S}}$ is injective and

$$\left\|\Phi^{+}_{\mathbf{A}^{(k)},\mathcal{S}}\right\|_{\mathrm{op}} \;\le\; 2\kappa_\Phi. \tag{112}$$

Line 4 sets $\boldsymbol{\theta}^{(k+1)} := \Phi^{+}_{\mathbf{A}^{(k)},\mathcal{S}}(\mathbf{A}^{(k)}\mathbf{S}_y(\mathbf{A}^{(k)})^{\mathsf{T}})$, the least-squares solution, which is what the algorithm computes whether or not the moment system is consistent; consistency can fail when $q < m(m+1)/2$, so we do not assume it. Injectivity gives $\Phi^{+}_{\mathbf{A}^{(k)},\mathcal{S}}\Phi_{\mathbf{A}^{(k)},\mathcal{S}} = \mathbf{I}$, and the truth satisfies $\Phi_{\mathbf{A}^{(k)},\mathcal{S}}(\boldsymbol{\theta}^\star) = \mathbf{A}^{(k)}\mathbf{\Sigma}_\epsilon(\mathbf{A}^{(k)})^{\mathsf{T}}$. Subtracting, and splitting $\mathbf{S}_y - \mathbf{\Sigma}_\epsilon = (\mathbf{S}_y - \mathbf{\Sigma}_y) + \mathbf{\Sigma}_x$,

$$\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star = \Phi^{+}_{\mathbf{A}^{(k)},\mathcal{S}}\Big(\mathbf{A}^{(k)}(\mathbf{S}_y - \mathbf{\Sigma}_y)(\mathbf{A}^{(k)})^{\mathsf{T}} + \mathbf{A}^{(k)}\mathbf{\Sigma}_x(\mathbf{A}^{(k)})^{\mathsf{T}}\Big). \tag{113}$$

The first term in the bracket is the sampling error and the second is the subspace error. The second is where the argument turns. Because $\mathcal{C} = \mathrm{row}(\mathbf{A}) = \mathrm{Null}(\mathbf{\Sigma}_x)$ we have $\mathbf{\Sigma}_x\mathbf{A}^{\mathsf{T}} = \mathbf{0}$, so for any rotation $\mathbf{R}_k$,

$$\mathbf{A}^{(k)}\mathbf{\Sigma}_x(\mathbf{A}^{(k)})^{\mathsf{T}} = (\mathbf{F}_k + \mathbf{R}_k\mathbf{A})\mathbf{\Sigma}_x(\mathbf{F}_k + \mathbf{R}_k\mathbf{A})^{\mathsf{T}} = \mathbf{F}_k\mathbf{\Sigma}_x\mathbf{F}_k^{\mathsf{T}}, \tag{114}$$

since each of the other three terms carries a factor $\mathbf{\Sigma}_x\mathbf{A}^\mathsf{T}$ or its transpose $\mathbf{A}\mathbf{\Sigma}_x$. Because $\mathbf{F}_k\mathbf{\Sigma}_x\mathbf{F}_k^\mathsf{T}$ is an $m \times m$ matrix, its Frobenius norm is at most $\sqrt{m}$ times its operator norm, and therefore

$$\left\|\mathbf{A}^{(k)}\mathbf{\Sigma}_x(\mathbf{A}^{(k)})^\mathsf{T}\right\|_F \;\le\; \sqrt{m}\,\left\|\mathbf{F}_k\mathbf{\Sigma}_x\mathbf{F}_k^\mathsf{T}\right\|_{\mathrm{op}} \;\le\; \sqrt{m}\,\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,e_k^2. \tag{115}$$

This is the whole point of the argument: the subspace error enters at order e_k^2 and not e_k, which is what lets the iteration contract instead of merely propagating error.

For the sampling error we need a bound that is valid uniformly over the data-dependent $\mathbf{A}^{(k)}$, so we use only that $\mathbf{A}^{(k)}$ has orthonormal rows:

$$\left\|\mathbf{A}^{(k)}(\mathbf{S}_y - \mathbf{\Sigma}_y)(\mathbf{A}^{(k)})^\mathsf{T}\right\|_F \;\le\; \sqrt{m}\,\left\|\mathbf{A}^{(k)}\right\|_{\mathrm{op}}^2\,\Delta_n \;=\; \sqrt{m}\,\Delta_n. \tag{116}$$

Applying Eq. (112) to Eq. (113) and inserting Eqs. (115) and (116),

$$\left\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\right\|_2 \;\le\; 2\kappa_\Phi\sqrt{m}\left(\Delta_n + \|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,e_k^2\right). \tag{117}$$

Step 3: from the parameter error to b_{k+1}. This passage is carried out in Frobenius norm, because Frobenius is the norm in which the projection is non-expansive. The set $\mathcal{K}$ of Eq. (2) is closed, convex and contains $\mathbf{\Sigma}_\epsilon$, so $\Pi_\mathcal{K}$ fixes $\mathbf{\Sigma}_\epsilon$ and is 1-Lipschitz there. Writing $\widehat{\mathbf{\Sigma}}_\epsilon^{(k+1)} := \Pi_\mathcal{K}(\mathrm{mat}_\mathcal{S}(\boldsymbol{\theta}^{(k+1)}))$,

$$\begin{aligned} b_{k+1} = \left\|\widehat{\mathbf{\Sigma}}_\epsilon^{(k+1)} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}} \;&\le\; \left\|\Pi_\mathcal{K}\big(\mathrm{mat}_\mathcal{S}(\boldsymbol{\theta}^{(k+1)})\big) - \Pi_\mathcal{K}(\mathbf{\Sigma}_\epsilon)\right\|_F \\ &\le\; \left\|\mathrm{mat}_\mathcal{S}(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star)\right\|_F \;\le\; \sqrt{2}\left\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\right\|_2, \end{aligned} \tag{118}$$

where the first step is $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$, the second is non-expansiveness, and the third is Eq. (89).

Step 4: the quadratic recurrence. Squaring the bound on e_k in Eq. (109) gives $e_k^2 \le 2\beta_\star^2(\Delta_n + b_k)^2$. Substituting that into Eq. (117) and then into Eq. (118) yields

$$b_{k+1} \;\le\; 2\sqrt{2}\,\sqrt{m}\,\kappa_\Phi\,\Delta_n \;+\; Q\,(\Delta_n + b_k)^2, \qquad Q \;:=\; 4\sqrt{2}\,\sqrt{m}\,\kappa_\Phi\,\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,\beta_\star^2. \tag{119}$$

On the region $\Delta_n + b_k \le 1/(2Q)$ the quadratic term obeys $Q(\Delta_n + b_k)^2 \le (\Delta_n + b_k)/2$, so Eq. (119) becomes

$$b_{k+1} \;\le\; \left(2\sqrt{2}\,\sqrt{m}\,\kappa_\Phi + \tfrac{1}{2}\right)\Delta_n \;+\; \tfrac{1}{2}b_k \;\le\; c_1\sqrt{m}\,\kappa_\Phi\,\Delta_n \;+\; \tfrac{1}{2}b_k, \qquad c_1 := 4, \tag{120}$$

which is Eq. (96). The stray $\Delta_n/2$ is absorbed, not discarded: Lemma 23(i) gives $\sqrt{m}\,\kappa_\Phi \ge 1/\sqrt{2}$, hence

$$\begin{aligned} \tfrac{1}{2} \;&\le\; \tfrac{1}{\sqrt{2}}\sqrt{m}\,\kappa_\Phi \qquad\Longrightarrow \\ 2\sqrt{2}\,\sqrt{m}\,\kappa_\Phi + \tfrac{1}{2} \;&\le\; \left(2\sqrt{2} + \tfrac{1}{\sqrt{2}}\right)\sqrt{m}\,\kappa_\Phi \;=\; \tfrac{5}{\sqrt{2}}\sqrt{m}\,\kappa_\Phi \;\le\; 4\sqrt{m}\,\kappa_\Phi. \end{aligned} \tag{121}$$

Step 5: the basin is reached and forward invariant. The region $\Delta_n + b_k \le 1/(2Q)$ is reached, because the definitions of ρ and Q give

$$\rho\,Q \;=\; 4\sqrt{2}\,c_0 \;=\; \frac{\sqrt{2}}{12} \;\le\; \frac{1}{4} \qquad\Longrightarrow\qquad \Delta_n + b_k \overset{(101)}{\le} 2\rho \;\le\; \frac{1}{2Q}. \tag{122}$$

It is also forward invariant. Assume $b_k \le \rho$. Applying Eq. (120) and then Eq. (100),

$$b_{k+1} \;\le\; c_1\sqrt{m}\,\kappa_\Phi \cdot \frac{c_2\,\rho}{1 + \sqrt{m}\,\kappa_\Phi} \;+\; \frac{\rho}{2} \;\le\; c_1c_2\,\rho \;+\; \frac{\rho}{2} \;\le\; \frac{\rho}{2} + \frac{\rho}{2} \;=\; \rho, \tag{123}$$

the middle step because $x/(1+x) \le 1$ for $x \ge 0$ and the last because $c_1c_2 = 4 \cdot \frac{1}{8} = \frac{1}{2}$. Induction from $b_1 \le \rho$ therefore keeps every iterate in the basin, which is what licenses Step 0 at every k.

Step 6: unrolling. Iterating Eq. (120) and summing the geometric series $\sum_{j\ge0} 2^{-j} = 2$,

$$b_k \;\le\; \underbrace{2^{-(k-1)}b_1}_{\text{transient}} \;+\; \underbrace{2c_1\sqrt{m}\,\kappa_\Phi\,\Delta_n}_{\text{statistical floor}}. \tag{124}$$

The two terms are handled separately, and neither uses the realized value of Δ_n from below. For the transient, $k \geq T_\star(\delta)$ means $2^{-k} \leq U_n/\rho$, so with $b_1 \leq \rho$,

$$2^{-(k-1)} b_1 \;\leq\; 2\,\rho\, 2^{-k} \;\leq\; 2\,U_n, \tag{125}$$

the first step because $b_1 \leq \rho$. For the floor, Eq. (24) gives $\Delta_n \leq C\,U_n$ on $\mathcal{E}_\delta$. Substituting both into Eq. (124) and using $\sqrt{m}\,\kappa_\Phi \geq 1/\sqrt{2}$ from Lemma 23(i) to absorb the transient,

$$b_k \;\leq\; 2U_n \;+\; 2c_1 C\sqrt{m}\,\kappa_\Phi\,U_n \;\leq\; \left(2\sqrt{2} + 2c_1 C\right)\sqrt{m}\,\kappa_\Phi\,U_n \;\lesssim\; \sqrt{m}\,\kappa_\Phi\,U_n. \tag{126}$$

Substituting Eq. (126) into Eq. (109) and applying Eq. (80),

$$\|\mathbf{P}_{\mathcal{C}^{(k)}} - \mathbf{P}_{\mathcal{C}}\|_{\mathrm{op}} \;\lesssim\; \beta_\star\,\sqrt{m}\,\kappa_\Phi\,U_n \;\lesssim\; \frac{\sqrt{m}\,\kappa_\Phi\,\kappa(\mathbf{\Sigma}_y)^{3/2}}{\sigma_{\mathcal{C}}^2}\,U_n, \tag{127}$$

which is Eq. (98). No step used the true $\mathbf{\Sigma}_\epsilon$ or the true $\mathbf{A}$, only the support $\mathcal{S}$ and the number of linear relations m. □

Remark 5 (Iteration budget.). Theorem 8 controls every iterate with index $k \geq T_\star(\delta)$. Therefore a fixed budget $T \geq T_\star(\delta)$ returns an iterate at the statistical floor, where $T_\star(\delta) = \mathcal{O}(\log\sqrt{n/d})$ by Eq. (97).

Remark 6 (Implementation of line 4.). The experiments implement the projected pseudoinverse update stated in Algorithm 1 and analyzed in Theorem 8, up to numerical solution tolerance. They use at most 100 iterations and a numerical convergence tolerance on successive constraint projectors.

Remark 7 (Scaling of the sufficient basin.). The radius ρ of Eq. (94) scales as $\sigma_{\mathcal{C}}^4/(\sqrt{m}\,\kappa_\Phi\,\|\mathbf{\Sigma}_x\|_{\mathrm{op}}\,\kappa(\mathbf{\Sigma}_y)^3)$. Thus the theorem gives an explicit sufficient initialization region whose radius reflects the conditioning of both the moment map and the observation covariance. The constants in this sufficient condition are not optimized.

Remark 8 (Uniform rate for the data-dependent iterate.). Because $\mathbf{A}^{(k)}$ depends on the data, the end-to-end proof controls $\left\|\mathbf{A}^{(k)}(\mathbf{S}_y - \mathbf{\Sigma}_y)(\mathbf{A}^{(k)})^{\mathsf{T}}\right\|_F$ uniformly over matrices with orthonormal rows and therefore has the $\sqrt{d/n}$ dependence of Theorem 8. The compressed $\sqrt{m^2/n}$ rate in Remark 3 applies to the oracle setting with a fixed $\mathbf{A}$.

K Estimating the Number of Linear Relations

This section gives the formal version of Theorem 6. For the known support $\mathcal{S}$, let $q := |\mathcal{S}|$ and define the candidate set

$$\mathcal{I}_q \;:=\; \{k \in \{1,\ldots,d-1\} : q \leq k(k+1)/2\}\,. \tag{128}$$

Algorithm 2 skips every $k \notin \mathcal{I}_q$. The standing identifiability condition $q \leq m(m+1)/2$ guarantees $m \in \mathcal{I}_q$, and every $k > m$ then also belongs to $\mathcal{I}_q$. Thus this check removes only candidate values for which the unrestricted supported moment map cannot be injective; it does not remove the true value m or alter the proof below.

Under Assumption 4, every candidate run initializes in and projects each covariance update onto the closed convex set

$$\mathcal{K}_{\mathrm{cap}} \;:=\; \left\{\mathbf{M} \in \mathcal{K} : \mathbf{M} \preceq \bar{\sigma}^2\mathbf{I}\right\}, \qquad \Pi_{\mathcal{K}_{\mathrm{cap}}}(\widetilde{\mathbf{M}}) \;:=\; \operatorname*{arg\,min}_{\mathbf{M}\in\mathcal{K}_{\mathrm{cap}}} \left\|\mathbf{M} - \widetilde{\mathbf{M}}\right\|_F. \tag{129}$$

After the candidate run with k linear relations, recompute the generalized eigenvalues of the final pencil $(\mathbf{S}_y, \widehat{\mathbf{\Sigma}}_\epsilon^{[k]})$ and define

$$c_k \;:=\; \#\left\{j : \lambda_j \text{ is finite and } |\lambda_j - 1| \leq \tau\right\}, \qquad \mathrm{PASS}_k \;:\Longleftrightarrow\; c_k \geq k. \tag{130}$$

An invalid candidate output is declared to fail. The recomputation in Eq. (130) is needed because the last covariance update follows the last generalized-eigenvalue solve in Algorithm 1.

Let $\mathcal{E}_\delta$ be the covariance-concentration event of Eq. (24), and put

$$c_\dagger := 2\sqrt{2} + 2c_1 C, \tag{131}$$

where $c_1 = 4$ is the contraction constant and C is the concentration constant. Choose

$$0 < \tau \;\le\; \frac{\sigma_{\mathcal{C}}^2 - \bar{\sigma}^2}{3\,\|\mathbf{\Sigma}_y\|_{\mathrm{op}}} \tag{132}$$

and suppose

$$n \;\ge\; n_{\mathrm{ord}}(\delta,\tau) \;:=\; \max\left\{ n_{\star\star}(\delta), \frac{16(C + c_\dagger\sqrt{m}\,\kappa_\Phi)^2 K^4 \kappa(\mathbf{\Sigma}_y)^2\big(d + \log(1/\delta)\big)}{\tau^2} \right\}. \tag{133}$$

Lemma 24 (Overspecified values fail)**.** *Under Assumption 4 and Eqs.* (132)–(133)*, on $\mathcal{E}_\delta$ every candidate output satisfies $c_k \le m$. Consequently every tested value $k > m$ fails, irrespective of its initialization or convergence behavior.*

Proof. Fix any candidate output $\widehat{\mathbf{\Sigma}}_\epsilon \in \mathcal{K}_{\mathrm{cap}}$ and set

$$\mathbf{G} \;:=\; \mathbf{S}_y^{-1/2}(\mathbf{S}_y - \widehat{\mathbf{\Sigma}}_\epsilon)\mathbf{S}_y^{-1/2}, \tag{134}$$

with eigenvalues $\nu_1 \le \cdots \le \nu_d$. By Lemma 20, a finite generalized eigenvalue satisfies $\lambda = 1/(1-\nu)$ and

$$|\lambda - 1| = \frac{|\nu|}{1-\nu}. \tag{135}$$

In particular, $|\lambda - 1| \le \tau$ implies $\nu \le \tau$: this is immediate for $\nu < 0$, while for $\nu \ge 0$ the right-hand side of Eq. (135) is at least ν.

On $\mathcal{E}_\delta$, the sample-size conditions imply

$$\lambda_{\min}(\mathbf{S}_y) \ge \tfrac{1}{2}\lambda_{\min}(\mathbf{\Sigma}_y), \qquad \|\mathbf{S}_y\|_{\mathrm{op}} \le \tfrac{3}{2}\,\|\mathbf{\Sigma}_y\|_{\mathrm{op}}, \qquad \Delta_n \le \tfrac{1}{4}(\sigma_{\mathcal{C}}^2 - \bar{\sigma}^2). \tag{136}$$

The first two inequalities follow from $n \ge n_{\star\star}(\delta) \ge n_\star(\delta)$. For the third, Eq. (133) gives $CU_n \le \tau\lambda_{\min}(\mathbf{\Sigma}_y)/4$, and Eq. (132) makes this no larger than $(\sigma_{\mathcal{C}}^2 - \bar{\sigma}^2)/12$.

By Courant–Fischer,

$$\nu_{m+1}(\mathbf{G}) \;=\; \min_{\dim W = m+1}\; \max_{\mathbf{v}\in W\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^\mathsf{T}(\mathbf{S}_y - \widehat{\mathbf{\Sigma}}_\epsilon)\mathbf{v}}{\mathbf{v}^\mathsf{T}\mathbf{S}_y\mathbf{v}}. \tag{137}$$

Every $(m+1)$-dimensional W contains a unit vector $\mathbf{v} \in \mathcal{C}^\perp$. For that vector,

$$\begin{aligned}\mathbf{v}^\mathsf{T}(\mathbf{S}_y - \widehat{\mathbf{\Sigma}}_\epsilon)\mathbf{v} &= \mathbf{v}^\mathsf{T}\mathbf{\Sigma}_x\mathbf{v} + \mathbf{v}^\mathsf{T}\mathbf{\Sigma}_\epsilon\mathbf{v} - \mathbf{v}^\mathsf{T}\widehat{\mathbf{\Sigma}}_\epsilon\mathbf{v} + \mathbf{v}^\mathsf{T}(\mathbf{S}_y - \mathbf{\Sigma}_y)\mathbf{v} \\ &\ge \sigma_{\mathcal{C}}^2 - \bar{\sigma}^2 - \Delta_n \ge \tfrac{3}{4}(\sigma_{\mathcal{C}}^2 - \bar{\sigma}^2),\end{aligned} \tag{138}$$

where the cap controls $\mathbf{v}^\mathsf{T}\widehat{\mathbf{\Sigma}}_\epsilon\mathbf{v}$. Combining this numerator with Eq. (136),

$$\nu_{m+1}(\mathbf{G}) \;\ge\; \frac{\sigma_{\mathcal{C}}^2 - \bar{\sigma}^2}{2\,\|\mathbf{\Sigma}_y\|_{\mathrm{op}}} \;>\; \tau. \tag{139}$$

Hence at most the first m eigenvalues can produce finite generalized eigenvalues within τ of one, so $c_k \le m$. If $k > m$, then $c_k < k$ and the candidate fails. □

Lemma 25 (The true value passes)**.** *Under Assumption 4 and Eqs.* (132)–(133)*, suppose the run with m linear relations starts in the basin of Theorem 8 and uses $T \ge T_\star(\delta)$ iterations. Then, on $\mathcal{E}_\delta$, $c_m = m$ and the true value passes.*

Proof. The set $\mathcal{K}_{\mathrm{cap}}$ is closed and convex, contains the true $\mathbf{\Sigma}_\epsilon$, and is a subset of $\mathcal{K}$. Replacing $\Pi_{\mathcal{K}}$ by $\Pi_{\mathcal{K}_{\mathrm{cap}}}$ therefore preserves every projection inequality in the proof of Theorem 8. Equation (126) gives for the returned estimate

$$b := \left\|\widehat{\mathbf{\Sigma}}_\epsilon^{[m]} - \mathbf{\Sigma}_\epsilon\right\|_{\mathrm{op}} \;\le\; c_\dagger\sqrt{m}\,\kappa_\Phi U_n. \tag{140}$$

Table 5: Exact configuration of Case Study 1. The support $\mathcal{S}$ consists of five diagonal pairs and one off-diagonal pair $(1,3)$, giving $q = 6 = m(m+1)/2$ free symmetric parameters.

Quantity	Value
d	5
m	3
Support $\mathcal{S}$	diagonal entries and $(1,3)$
$\mathrm{diag}(\Sigma_e)$	$[0.10, 0.26, 0.375, 0.425, 0.095]$
$\Sigma_e(1,3) = \Sigma_e(3,1)$	0.0871
Variance SNR floor	10
Sample sizes	$[50, 100, 200, 500, 1000, 2000, 5000]$
Monte Carlo trials	100

Let $\mathbf{G}_\star := \mathbf{S}_y^{-1/2}\mathbf{\Sigma}_x\mathbf{S}_y^{-1/2}$. It is positive semidefinite and has exactly m zero eigenvalues. Using Eq. (136),

$$\eta := \|\mathbf{G} - \mathbf{G}_\star\|_{\mathrm{op}} \le \frac{\Delta_n + b}{\lambda_{\min}(\mathbf{S}_y)} \le \frac{2(C + c_\dagger\sqrt{m}\,\kappa_\Phi)U_n}{\lambda_{\min}(\mathbf{\Sigma}_y)} \le \frac{\tau}{2}, \tag{141}$$

where the last step is Eq. (133). Weyl's inequality places the m eigenvalues corresponding to the null space of $\mathbf{G}_\star$ in $[-\eta, \eta]$. Since $\tau < 1/3$, Eq. (135) gives

$$|\lambda_j - 1| \le \frac{\eta}{1-\eta} \le 2\eta \le \tau, \qquad j = 1, \dots, m. \tag{142}$$

Thus $c_m \ge m$. Lemma 24 gives the reverse inequality, so $c_m = m$. □

Theorem 6 (formal recovery of the number of linear relations). *Under the conditions of Lemma 25, Algorithm 2 returns $\widehat{m} = m$ with probability at least $1 - \delta$. On this event it invokes Algorithm 1 at exactly $d - m$ candidate values.*

Proof. Work on $\mathcal{E}_\delta$, whose probability is at least $1-\delta$. The identifiability check retains m because $q \le m(m+1)/2$. It also retains every $k > m$. Lemma 24 shows that each candidate $d-1, d-2, \dots, m+1$ fails, while Lemma 25 shows that m passes. The first passing value in the descending sweep is therefore exactly m. Both conclusions hold on the same event, so no union bound over candidate values is required. The tested candidates are precisely $d-1, d-2, \dots, m$, a total of $d-m$ runs. □

L Case Study 1: Five-Variable Identifiable System

This appendix records the exact configuration of the first case study from Section 4.

Generalized eigenvalue spectrum. At $n = 5000$ with $\mathbf{\Sigma}_\epsilon \succ 0$, the m fitted constraint eigenvalues deviate from one by about 1.7×10^{-15} on average, while the smallest non-constraint eigenvalue is about 22.2, rising to about 39.3 with $s = 1$. Because $q = 6 = m(m+1)/2$, the exactly determined moment update enforces $\mathbf{A}\mathbf{S}_y\mathbf{A}^\mathsf{T} = \Phi_{\mathbf{A},\mathcal{S}}(\boldsymbol{\theta})$, so the fitted constraint directions form a unity cluster in the sample pencil. The separation from the remaining generalized eigenvalues provides the spectral diagnostic used by the order-selection procedure.

When $s = 2$ and $\mathbf{\Sigma}_\epsilon$ is singular, the pencil has only $m = 3$ finite near-one eigenvalues after accounting for the two infinite directions, so the next-finite-eigenvalue diagnostic is undefined and is reported as NaN in the saved summaries.

L.1 Misspecifying the Number of Linear Relations with Fixed Support

We first isolate misspecification of m from misspecification of the nuisance model. Data are generated from the $d = 5$, true-$m = 3$, $s = 1$ configuration, with coordinate 5 noise-free and support

Table 6: Supported-noise parameter error for Case Study 1 with $\mathbf{\Sigma}_\epsilon \succ 0$ and $n \geq 500$. Each entry reports $\left\|\widehat{\theta} - \theta^\star\right\|_2$ as mean $\pm$ standard deviation over 100 Monte Carlo replicates. "d.n.e." means *does not estimate*: the method returns no estimate of $\mathbf{\Sigma}_\epsilon$.

n	OLS	Std PCA	RPCA	GoDec	IPCA	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	d.n.e.	d.n.e.	d.n.e.	d.n.e.	$\mathbf{0.1 \pm 0.0}$	5.9 ± 1.1	4.0 ± 0.4	$\mathbf{0.1 \pm 0.0}$
1000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	$\mathbf{0.0 \pm 0.0}$	5.7 ± 1.6	3.9 ± 0.3	$\mathbf{0.0 \pm 0.0}$
2000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	$\mathbf{0.0 \pm 0.0}$	5.7 ± 1.5	3.8 ± 0.2	$\mathbf{0.0 \pm 0.0}$
5000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	$\mathbf{0.0 \pm 0.0}$	5.8 ± 1.3	3.8 ± 0.1	$\mathbf{0.0 \pm 0.0}$

Table 7: Supported-noise parameter error for Case Study 1 with singular $\mathbf{\Sigma}_\epsilon$ and $n \geq 500$. Each entry reports $\left\|\widehat{\theta} - \theta^\star\right\|_2$ as mean $\pm$ standard deviation over 100 Monte Carlo replicates. "d.n.e." means *does not estimate*: the method returns no estimate of $\mathbf{\Sigma}_\epsilon$.

s	n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
1	500	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.8 ± 1.1	3.9 ± 0.4	$\mathbf{0.1 \pm 0.0}$
1	1000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.8 ± 1.3	3.8 ± 0.2	$\mathbf{0.0 \pm 0.0}$
1	2000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.7 ± 1.3	3.7 ± 0.2	$\mathbf{0.0 \pm 0.0}$
1	5000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.8 ± 1.2	3.7 ± 0.1	$\mathbf{0.0 \pm 0.0}$
2	500	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.7 ± 1.4	2.9 ± 0.6	$\mathbf{0.0 \pm 0.0}$
2	1000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.6 ± 1.4	3.0 ± 0.5	$\mathbf{0.0 \pm 0.0}$
2	2000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.7 ± 1.2	2.9 ± 0.3	$\mathbf{0.0 \pm 0.0}$
2	5000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	5.5 ± 1.7	3.0 ± 0.2	$\mathbf{0.0 \pm 0.0}$

$\mathcal{S} = \{(1,1),(2,2),(3,3),(4,4),(1,3)\}$. Both candidate values of $\widetilde{m}$ receive this same support and the same samples; only the supplied number of linear relations changes. For a candidate $\widetilde{m}$, let $\widehat{\lambda}_{(\widetilde{m})}$ denote the $\widetilde{m}$th finite generalized eigenvalue after sorting by $|\widehat{\lambda} - 1|$.

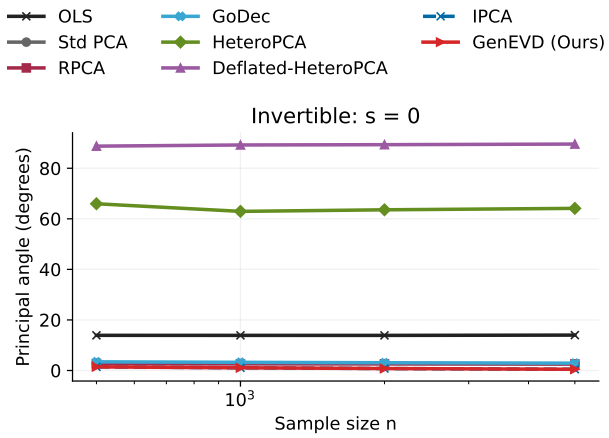
Table 8: Fixed-support relation-number misspecification experiment over 100 replicates. Entries are the candidate unity eigenvalues, reported as mean $\pm$ standard deviation. "d.n.e." means that the estimator does not return a finite fourth generalized eigenvalue.

n	$\widehat{\lambda}_{(3)}, \widetilde{m} = 3$	$\widehat{\lambda}_{(4)}, \widetilde{m} = 4$
500	1.007 ± 0.094	d.n.e.
1000	0.998 ± 0.064	d.n.e.
2000	1.005 ± 0.051	d.n.e.
5000	1.005 ± 0.031	d.n.e.

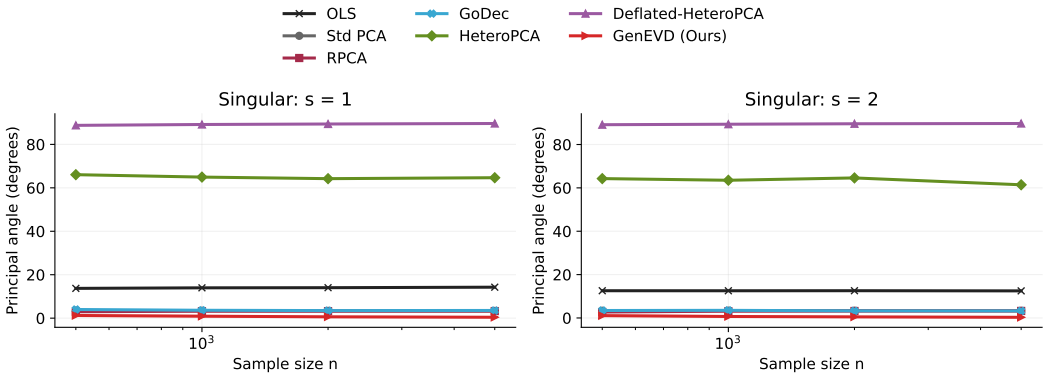
The over-specified fit leaves only three finite generalized eigenvalues, while four are requested. In contrast, if the support is expanded to all ten entries of the noisy 4×4 block, the ten covariance parameters can interpolate the ten moment equations, absorb a signal direction, and force four sample eigenvalues to one. Holding $\mathcal{S}$ fixed is therefore essential when the experiment is intended to change only m.

L.2 Incorrect Off-Diagonal Support

We next retain the same data-generating process but give both candidate values the incorrect support $\widetilde{\mathcal{S}} = \{(1,1),(2,2),(3,3),(4,4),(1,2)\}$, replacing the true pair $(1,3)$ by $(1,2)$. The four noisy variances, SNR floor of 10, sample-size grid, and 100 common-seed replicates are unchanged. This keeps the number of nuisance parameters fixed at five and tests whether the pencil itself reveals support misspecification.



(a) Principal-angle curve with invertible noise covariance.



(b) Singular principal-angle curves for $s = 1$ and $s = 2$.

Figure 1: Principal-angle curves for Case Study 1, including Deflated-HeteroPCA. Lines show means over 100 replicates; red bands show one standard deviation for GenEVD. The tables give standard deviations for all methods.

Table 9: Covariance-support misspecification experiment over 100 replicates. Covariance error is $\left\|\widehat{\boldsymbol{\Sigma}_\epsilon} - \boldsymbol{\Sigma}_\epsilon\right\|_F$. The fourth-eigenvalue statistics use the finite returned values, whose counts across increasing n are 13, 9, 13, and 11 out of 100.

n	$\widehat{\lambda}_{(3)}$	$\lvert\widehat{\lambda}_{(3)} - 1\rvert$	covariance error, $\widetilde{m} = 3$	$\widehat{\lambda}_{(4)}$, $\widetilde{m} = 4$
500	0.912 ± 0.324	0.331 ± 0.047	0.164 ± 0.016	$(-0.089 \pm 24.691) \times 10^{14}$
1000	0.899 ± 0.303	0.317 ± 0.031	0.159 ± 0.013	$(2.422 \pm 25.132) \times 10^{14}$
2000	0.867 ± 0.288	0.315 ± 0.022	0.154 ± 0.007	$(-2.428 \pm 24.029) \times 10^{14}$
5000	0.871 ± 0.287	0.313 ± 0.014	0.153 ± 0.005	$(3.511 \pm 23.362) \times 10^{14}$

At $n = 5000$, GenEVD places -0.0447 ± 0.0031 at the supplied pair $(1, 2)$ while the true nonzero pair $(1, 3)$ is excluded. Its covariance error is consequently much larger than the 0.0197 ± 0.0086 obtained under the correct support. The numerically unbounded fourth eigenvalue under $\widetilde{m} = 4$ supplies a second diagnostic of the incorrect support. OLS, Standard PCA, HeteroPCA, Deflated-HeteroPCA, RPCA, and GoDec do not read $\mathcal{S}$; rerunning them on the identical samples produces exactly the same projectors (and, for the two HeteroPCA variants, exactly the same diagonal covariance estimates). They therefore supply no diagnostic for choosing between $(1, 2)$ and $(1, 3)$.

L.3 Unknown Diagonal Support

Finally, we set every off-diagonal noise entry to zero and generate $\mathrm{diag}(\mathbf{\Sigma}_\epsilon) = (0.10, 0.26, 0.375, 0.425, 0)$. Coordinate 5 is noise-free in the data, but this fact is withheld from GenEVD: the supplied candidate support contains all five diagonal pairs and no off-diagonal pair. Thus the method must estimate $[\mathbf{\Sigma}_\epsilon]_{55}$ along with the four positive variances. We use the same $d = 5$, true $m = 3$, SNR floor, sample-size grid, and 100 replicates.

Table 10: Unknown diagonal-support experiment. Entries are mean $\pm$ standard deviation over 100 replicates.

	$\widetilde{m} = 3$			$\widetilde{m} = 4$		
n	$\widehat{\lambda}_{(3)}$	$[\widehat{\mathbf{\Sigma}_\epsilon}]_{55}$	covariance error	$\widehat{\lambda}_{(4)}$	$[\widehat{\mathbf{\Sigma}_\epsilon}]_{55}$	covariance error
500	0.965 ± 0.066	0.004 ± 0.005	0.060 ± 0.023	2.440 ± 0.025	0.868 ± 0.059	1.655 ± 0.117
1000	0.960 ± 0.048	0.003 ± 0.004	0.042 ± 0.019	2.438 ± 0.015	0.866 ± 0.038	1.645 ± 0.082
2000	0.979 ± 0.033	0.002 ± 0.003	0.030 ± 0.012	2.436 ± 0.011	0.868 ± 0.027	1.650 ± 0.058
5000	0.982 ± 0.025	0.001 ± 0.002	0.019 ± 0.007	2.434 ± 0.006	0.868 ± 0.018	1.650 ± 0.038

For the correct value $m = 3$, the median fifth variance is exactly zero at $n = 5000$, 96% of estimates are at most 0.005, and every estimate is at most 0.01. Thresholding therefore recovers the withheld noise-free coordinate. The incorrect value $\widetilde{m} = 4$ instead treats substantial signal variation as measurement noise, which is visible both in $\widehat{\lambda}_{(4)}$ and in the inflated fifth variance.

Table 11: Noise-covariance diagnostics at $n = 5000$ and $\widetilde{m} = 3$ when the zero fifth variance is withheld. Methods marked "d.n.e." do not return a noise-covariance estimate.

Method	$\left\|\widehat{\mathbf{\Sigma}_\epsilon} - \mathbf{\Sigma}_\epsilon\right\|_F$	$[\widehat{\mathbf{\Sigma}_\epsilon}]_{55}$
OLS	d.n.e.	d.n.e.
Standard PCA	d.n.e.	d.n.e.
RPCA	d.n.e.	d.n.e.
GoDec	d.n.e.	d.n.e.
HeteroPCA	5.533 ± 1.960	0.788 ± 0.279
Deflated-HeteroPCA	3.838 ± 0.104	0.870 ± 0.018
GenEVD	$\mathbf{0.019 \pm 0.007}$	$\mathbf{0.001 \pm 0.002}$

M Additional Experimental Results

Case Study 2. We evaluate GenEVD on a higher-dimensional singular instance with $d = 10$, $m = 5$, and $s = 3$ noise-free coordinates. The support has $q = 15 = m(m+1)/2$ free symmetric parameters, comprising seven diagonal pairs and eight off-diagonal pairs, and the moment system is exactly identifiable under the chosen constraint matrix. Tables 12 and 13 report mean $\pm$ standard deviation over 40 replicates at $n = 20000$. GenEVD attains the lowest error in both subspace recovery and error covariance estimation.

Table 12: Subspace recovery error for Case Study 2 with $d = 10$, $m = 5$, $s = 3$, and $q = 15$. Each entry reports $\theta_{\max}$ in degrees as mean $\pm$ standard deviation over 40 replicates at $n = 20000$.

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
20000	3.8 ± 0.1	2.0 ± 0.1	1.9 ± 0.1	2.1 ± 0.2	87.1 ± 0.4	86.9 ± 0.4	$\mathbf{0.3 \pm 0.1}$

Table 13: Supported-noise parameter error for Case Study 2 at $n = 20000$ over 40 replicates. "d.n.e." means *does not estimate*: the method returns no estimate of $\mathbf{\Sigma}_\epsilon$.

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
20000	d.n.e.	d.n.e.	d.n.e.	d.n.e.	6.3 ± 0.1	4.9 ± 0.0	$\mathbf{0.0 \pm 0.0}$

Table 14: Runtime in milliseconds, mean $\pm$ standard deviation over the same synthetic replicates used in the accuracy tables. Case Study 1 uses $n = 5000$ and Case Study 2 uses $n = 20000$. "n/a" marks a configuration the method does not apply to: IPCA requires $\mathbf{\Sigma}_\epsilon \succ 0$.

Method	Case 1: $s = 0$	Case 1: $s = 1$	Case 1: $s = 2$	Case 2
OLS	57.414 ± 7.518	56.425 ± 4.293	61.835 ± 10.830	1819.693 ± 461.411
Std PCA	0.066 ± 0.010	0.070 ± 0.021	0.074 ± 0.024	0.213 ± 0.280
RPCA	3.375 ± 0.233	3.446 ± 0.218	3.400 ± 0.343	34.079 ± 3.437
GoDec	48.215 ± 1.877	52.046 ± 1.544	51.943 ± 4.309	410.219 ± 20.865
HeteroPCA	0.792 ± 0.166	0.778 ± 0.076	0.853 ± 0.242	2.200 ± 1.807
Deflated-HeteroPCA	0.145 ± 0.039	0.146 ± 0.021	0.159 ± 0.084	1.141 ± 1.233
IPCA	0.781 ± 0.170	n/a	n/a	n/a
GenEVD (Ours)	0.391 ± 0.167	0.648 ± 0.123	0.525 ± 0.167	30.128 ± 42.406

M.1 Relation-Number Misspecification in Case Study 2

We retain the Case Study 2 constraint matrix with $d = 10$, true $m = 5$, and noise-free coordinates $\{8, 9, 10\}$, but use a common ten-parameter noise support for all candidate values $\widetilde{m} \in \{4, 5, 6\}$. The support contains the seven noisy diagonal pairs and the three off-diagonal pairs $(3, 5)$, $(3, 6)$, and $(4, 5)$, each with correlation 0.35. It therefore uses exactly the ten moment equations available under $\widetilde{m} = 4$. The true-support moment map has rank 10, its condition number is 12.46, and the minimum finite coordinate-wise variance SNR is 12.24. At each sample size, all three candidates receive the same samples in each of 40 replicates. As before, $\widehat{\lambda}_{(\widetilde{m})}$ is the $\widetilde{m}$th finite generalized eigenvalue after sorting by distance to one.

Table 15: Candidate unity eigenvalues in the Case Study 2 relation-number misspecification experiment, reported as mean $\pm$ standard deviation over the finite returned values. The $\widetilde{m} = 4$ and 5 columns use all 40 replicates. The $\widetilde{m} = 6$ column uses respectively 38, 37, and 40 finite returned values.

n	$\widehat{\lambda}_{(4)}, \widetilde{m} = 4$	$\widehat{\lambda}_{(5)}, \widetilde{m} = 5$	$\widehat{\lambda}_{(6)}, \widetilde{m} = 6$
20000	0.887 ± 0.285	0.995 ± 0.021	$(-4.112 \pm 1.254) \times 10^{11}$
50000	0.894 ± 0.279	0.997 ± 0.013	$(-4.122 \pm 1.379) \times 10^{11}$
100000	0.899 ± 0.307	0.999 ± 0.010	$(-4.499 \pm 1.426) \times 10^{11}$

The correctly specified fifth eigenvalue concentrates around one as n increases, whereas the over-specified sixth eigenvalue is numerically divergent with magnitude about 10^{11}. Hence the descending selector rejects $\widetilde{m} = 6$ and identifies $\widetilde{m} = 5$.

M.2 Semi-Synthetic Experiments on Real-World Datasets

Table 16: Semi-synthetic experiments constructed from real-world datasets at the largest sample size, reproduced from Table 3 so that this appendix is self-contained. Entries are principal angles in degrees, mean ± standard deviation. IPCA does not apply when the noise covariance is singular and is omitted, as in Table 2. Bold denotes the lowest mean in each row. The per-dataset sample-size grids are in Appendix M.2.

Dataset	n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
Energy	768	17.0 ± 0.6	13.7 ± 0.4	13.4 ± 0.4	13.8 ± 0.9	80.6 ± 3.6	12.6 ± 0.5	$\mathbf{1.5 \pm 0.6}$
Air Quality	8991	55.8 ± 3.5	41.3 ± 0.8	40.6 ± 0.8	33.3 ± 4.8	89.8 ± 0.2	51.6 ± 0.7	$\mathbf{0.8 \pm 0.3}$
Appliances	19735	13.1 ± 0.3	19.9 ± 0.5	20.2 ± 0.5	32.7 ± 9.0	86.1 ± 0.2	50.7 ± 0.4	$\mathbf{0.5 \pm 0.2}$
Gas Drift	13910	9.9 ± 0.1	9.0 ± 0.2	9.1 ± 0.2	11.7 ± 1.8	76.5 ± 0.6	13.3 ± 0.2	$\mathbf{0.4 \pm 0.1}$
MHEALTH	10000	13.2 ± 0.2	9.5 ± 0.2	9.4 ± 0.2	9.2 ± 1.2	63.2 ± 0.2	78.3 ± 0.3	$\mathbf{0.4 \pm 0.2}$
Power Plant	9568	30.2 ± 0.3	21.8 ± 0.4	21.7 ± 0.4	22.5 ± 10.6	87.8 ± 0.6	70.9 ± 0.6	$\mathbf{0.5 \pm 0.2}$
Household	20000	18.1 ± 0.2	13.2 ± 0.4	13.5 ± 0.4	50.9 ± 25.5	75.8 ± 0.7	85.1 ± 1.6	$\mathbf{0.3 \pm 0.1}$

M.2.1 Data Sources and Reference Construction

All source datasets are obtained from the UCI Machine Learning Repository (Kelly et al., 2023). Let n_{data} denote the number of observations retained after preprocessing and before trial-level subsampling. Energy Efficiency (UCI 242) uses all $n_{\mathrm{data}} = 768$ observations and coordinates $(Y_1, Y_2, X_1, X_2, X_5)$. Air Quality (UCI 360) treats -200 as missing and retains $n_{\mathrm{data}} = 8991$ complete observations of PT08.S1(CO)–PT08.S5(O3). Appliances Energy (UCI 374) retains $n_{\mathrm{data}} = 19735$ complete observations of $(T_1, \ldots, T_5)$ in temporal order.

Gas Sensor Drift (UCI 270) concatenates batches 1:10, selects columns $(1, 9, 17, 25, 33)$, applies $z \mapsto \mathrm{sign}(z)\log(1 + |z|)$ coordinatewise, and retains $n_{\mathrm{data}} = 13910$ observations. MHEALTH (UCI 319) uses subject 1, discards label 0 and incomplete observations, and retains $n_{\mathrm{data}} = 35174$ observations of columns $(9, 16, 18, 19, 20)$. Power Plant (UCI 294) uses the first worksheet and the $n_{\mathrm{data}} = 9568$ complete observations of $(\mathrm{AT}, \mathrm{V}, \mathrm{AP}, \mathrm{RH}, \mathrm{PE})$. Household Power (UCI 235) treats "?" as missing and retains the $n_{\mathrm{data}} = 2049280$ complete observations of global active power, voltage, global intensity, and sub-metering channels 1 and 3.

Let $R \in \mathbb{R}^{n_{\mathrm{data}} \times 5}$ denote a processed dataset matrix. We standardize each coordinate using all retained observations:

$$\mu_j := \frac{1}{n_{\mathrm{data}}} \sum_{i=1}^{n_{\mathrm{data}}} R_{ij}, \quad s_j^2 := \frac{1}{n_{\mathrm{data}}} \sum_{i=1}^{n_{\mathrm{data}}} (R_{ij} - \mu_j)^2, \quad Z_{ij} := \frac{R_{ij} - \mu_j}{s_j}, \qquad j \in [5]. \quad (143)$$

After a numerical recentering, let $Z = UDV^{\mathsf{T}}$ be its singular value decomposition and write $U_2 := U_{[:,1:2]}$, $D_2 := D_{[1:2,1:2]}$, and $V_2 := V_{[:,1:2]}$. The rank-two reference and its constraint projector are

$$X_{\mathrm{ref}} := U_2 D_2 V_2^{\mathsf{T}} = Z V_2 V_2^{\mathsf{T}}, \qquad P_\star := I_5 - V_2 V_2^{\mathsf{T}}. \quad (144)$$

Thus the processed dataset defines a fixed semi-synthetic population. The evaluation concerns the constructed rank-two constraint space and injected measurement noise, not unknown natural constraints or supervised prediction on the original datasets.

M.2.2 Covariance, Sampling, and Metrics

Let v_j be the sample variance of column j of X_{ref} (denominator $n_{\mathrm{data}} - 1$), $w \in \mathbb{R}^5$ the positive noise-scale vector listed below, F the two supplied noise-free indices, and $J = [5] \setminus F$. Define

$$a_j = \max\{v_j w_j / (2 \max_{\ell \in [5]} w_\ell), 10^{-8}\}, \quad j \in J, \qquad a_j = 0, \quad j \in F,$$

$$[\mathbf{\Sigma}_\epsilon]_{jj} = a_j, \qquad [\mathbf{\Sigma}_\epsilon]_{ij} = \rho\sqrt{a_i a_j}, \quad i \neq j.$$

The maximum in the normalization includes all five weights, even those of noise-free variables. Each noisy 3×3 block is positive definite, and $\mathbf{\Sigma}_\epsilon$ has rank 3. The free support consists of its three diagonal and three unordered off-diagonal pairs, so $q = 6$; (i, j) and (j, i) are one parameter. The saved

moment-design matrices have rank 6, not just six columns. All configurations satisfy $v_j/[\mathbf{\Sigma}_\epsilon]_{jj} \geq 2$ for $j \in J$ (noise-free coordinates have infinite SNR).

At grid index g and trial t, both one-based, set the MATLAB Mersenne Twister seed to $b + 10000g + t + 100000$, where b is the dataset's base seed. Draw n reference rows without replacement using $\texttt{randperm}(n_{\text{data}}, n)$. For an eigendecomposition $\mathbf{\Sigma}_\epsilon = QD_eQ^\mathsf{T}$, generate $E = QD_e^{1/2}G$, with $G \in \mathbb{R}^{5\times n}$ having independent standard normal entries, and form $Y = X_{\text{ref}}[\text{selected rows}, :]^\mathsf{T} + E$. Each method sees this same Y. Use $S_y = YY^\mathsf{T}/n$ without further trial-level centering. Grid points use different seeds and are not nested samples. Run methods in the order OLS, Std PCA, HeteroPCA, Deflated-HeteroPCA, RPCA, GoDec, GenEVD; GoDec uses the continuing trial random stream.

For each trial report $\theta_{\max} = (180/\pi)\arcsin(\min\{1, \|\widehat{P} - P_\star\|_{\text{op}}\})$ and the relative full-matrix covariance error $\|\widehat{\mathbf{\Sigma}_\epsilon} - \mathbf{\Sigma}_\epsilon\|_F/\|\mathbf{\Sigma}_\epsilon\|_F$. The latter is different from the unnormalized supported-parameter L_2 error used in the synthetic case studies. Means and sample standard deviations (denominator trials minus one) are computed across trials, without discarding failures; all reported saved runs completed successfully. The plotted bands show GenEVD variability, while tables give variability for every method. Tables bold the lowest mean, not a statistical-significance claim.

Table 17: Saved real-world-derived configurations. Noise-free indices use the feature ordering above.

Dataset	Dataset observations	Trials	Base seed	Noise-free	ρ
Energy	768	20	1310000	$3, 5$	0.5
Air Quality	8991	20	5710000	$4, 5$	0.5
Appliances	19735	30	6120000	$3, 4$	0.65
Gas Drift	13910	30	6210000	$3, 5$	0.65
MHEALTH	35174	30	6320000	$4, 5$	0.65
Power Plant	9568	30	6410000	$3, 5$	0.65
Household	2049280	30	6510000	$3, 4$	0.65

M.2.3 Method Implementations and Settings

All methods receive the target signal rank $r = 2$ (equivalently $m = 3$); the true constraint basis and the covariance values are used only for evaluation. Only GenEVD consumes the covariance support and noise-free indices. The following settings also describe the implementations used in the synthetic comparisons, with $r = d - m$ adjusted to the case study.

OLS. Pivoted QR applied to Y^T selects r predictor coordinates. The remaining m coordinates are regressed on them without an intercept, and the m fitted equations are orthonormalized to obtain the constraint projector. This is ordinary least squares: it does not correct measurement error in the predictors or estimate $\mathbf{\Sigma}_\epsilon$.

PCA and HeteroPCA. Standard PCA uses the bottom m eigenvectors of S_y. HeteroPCA follows Algorithm 1 of Zhang et al. (2022), and its constraint estimate is the orthogonal complement of the returned signal space. Its reported noise estimate is the diagonal residual surrogate produced by that model.

Deflated-HeteroPCA. Deflated-HeteroPCA follows Algorithm 2 of Zhou & Chen (2025) with the settings proposed there. Both HeteroPCA variants model diagonal noise, so their residual-noise estimates cannot represent the off-diagonal entries used in our experiments; their theory likewise does not cover these correlated-noise comparisons.

RPCA and GoDec. RPCA uses principal-component pursuit (Candès et al., 2011), and GoDec follows Zhou & Tao (2011); in both cases we extract the rank-r signal space returned by the published procedure. Neither method nor Standard PCA estimates a measurement-noise covariance, so no covariance error is reported for them.

GenEVD and IPCA. GenEVD implements Algorithm 1, including its generalized-eigenvalue selection and projected supported-moment update. It is initialized from the diagonal of S_y with

supplied noise-free entries set to zero and is run for at most $T = 100$ iterations with a numerical convergence tolerance on successive constraint projectors. IPCA uses scaling by the inverse Cholesky factor and the corresponding supported-moment update in the invertible synthetic comparisons; it is not run on the singular real-world-derived cases.

Scope of the comparisons. These experiments are conditional case studies under the fixed configurations reported in the appendix. GenEVD receives structural support information that the generic baselines do not use. Observations in the source datasets may be temporally dependent, whereas the theoretical guarantees are stated for independent samples.

Table 18: Relative full-matrix Frobenius noise-covariance error at the largest sample size, mean $\pm$ sample standard deviation. HeteroPCA variants provide diagonal surrogates; GenEVD estimates the supplied full support.

Dataset	n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
Energy	768	2.255 ± 0.250	0.696 ± 0.012	$\mathbf{0.060 \pm 0.026}$
Air Quality	8991	2.117 ± 0.020	0.984 ± 0.006	$\mathbf{0.016 \pm 0.006}$
Appliances	19735	1.270 ± 0.009	0.800 ± 0.002	$\mathbf{0.011 \pm 0.006}$
Gas Drift	13910	2.197 ± 0.032	0.791 ± 0.001	$\mathbf{0.013 \pm 0.007}$
MHEALTH	10000	5.988 ± 0.100	2.713 ± 0.048	$\mathbf{0.013 \pm 0.006}$
Power Plant	9568	1.399 ± 0.020	1.673 ± 0.018	$\mathbf{0.015 \pm 0.011}$
Household	20000	2.339 ± 0.033	1.952 ± 0.130	$\mathbf{0.009 \pm 0.004}$

M.2.4 ENERGY

The reference contains 768 rows; each grid point uses 20 trials. The sample-size grid is $100, 200, 400, 768$. The noise-scale vector is $(1.9, 2, 0.85, 0.95, 0.35)$, the base seed is 1310000, and $\rho = 0.5$. For the ordered noisy coordinates $J = (1, 2, 4)$, the true covariance has the following block (all other entries are zero; displayed values are rounded):

$$[\mathbf{\Sigma}_\epsilon]_{JJ} = \begin{bmatrix} 0.47 & 0.24 & 0.17 \\ 0.24 & 0.49 & 0.17 \\ 0.17 & 0.17 & 0.24 \end{bmatrix}.$$

Table 19: Energy: principal angle in degrees (mean ± sample standard deviation).

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
100	17.5 ± 2.0	14.6 ± 1.6	14.0 ± 1.8	15.3 ± 2.7	74.0 ± 5.3	13.2 ± 1.6	$\mathbf{5.1 \pm 2.4}$
200	17.1 ± 1.2	14.0 ± 0.9	13.3 ± 1.1	14.2 ± 1.5	78.8 ± 5.2	12.9 ± 1.0	$\mathbf{3.5 \pm 1.4}$
400	16.9 ± 0.8	13.6 ± 0.7	13.2 ± 0.7	13.8 ± 0.9	78.3 ± 3.0	12.6 ± 0.8	$\mathbf{2.2 \pm 0.7}$
768	17.0 ± 0.6	13.7 ± 0.4	13.4 ± 0.4	13.8 ± 0.9	80.6 ± 3.6	12.6 ± 0.5	$\mathbf{1.5 \pm 0.6}$

Table 20: Energy: relative full-matrix Frobenius covariance error (mean ± sample standard deviation).

n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
100	3.096 ± 0.757	0.714 ± 0.031	$\mathbf{0.177 \pm 0.070}$
200	2.436 ± 0.543	0.693 ± 0.024	$\mathbf{0.109 \pm 0.042}$
400	2.380 ± 0.288	0.703 ± 0.017	$\mathbf{0.079 \pm 0.028}$
768	2.255 ± 0.250	0.696 ± 0.012	$\mathbf{0.060 \pm 0.026}$

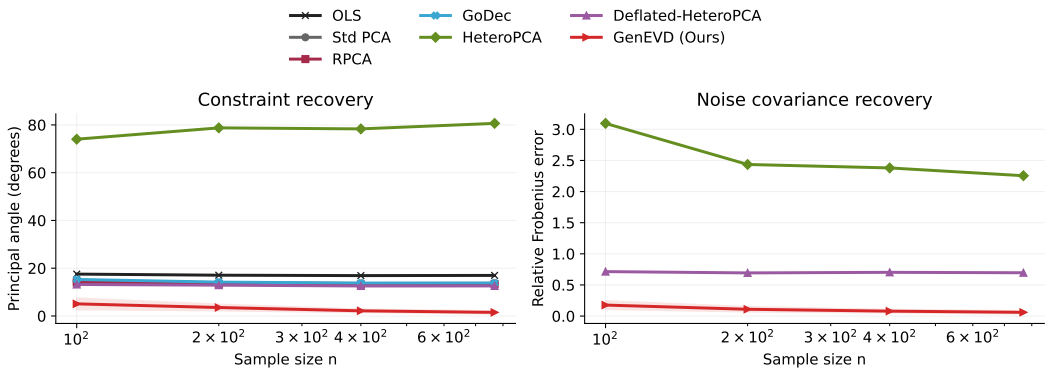


Figure 2: Energy: means over the saved trials. The red band denotes one standard deviation for GenEVD; tables report standard deviations for every method.

M.2.5 AIR QUALITY

The reference contains 8991 rows; each grid point uses 20 trials. The sample-size grid is $500, 1000, 2000, 5000, 8991$. The noise-scale vector is $(1.3, 1.15, 0.8, 0.95, 1.45)$, the base seed is 5710000, and $\rho = 0.5$. For the ordered noisy coordinates $J = (1, 2, 3)$, the true covariance has the following block (all other entries are zero; displayed values are rounded):

$$[\mathbf{\Sigma}_\epsilon]_{JJ} = \begin{bmatrix} 0.40 & 0.19 & 0.15 \\ 0.19 & 0.37 & 0.15 \\ 0.15 & 0.15 & 0.23 \end{bmatrix}.$$

Table 21: Air Quality: principal angle in degrees (mean $\pm$ sample standard deviation).

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	56.3 ± 4.9	42.3 ± 6.8	39.7 ± 7.0	40.3 ± 11.1	89.2 ± 0.7	51.9 ± 3.6	$\mathbf{2.9 \pm 1.4}$
1000	58.7 ± 4.6	41.1 ± 3.1	39.1 ± 3.1	34.4 ± 7.0	89.5 ± 0.4	51.1 ± 2.5	$\mathbf{2.5 \pm 1.0}$
2000	56.7 ± 4.4	41.6 ± 2.4	40.1 ± 2.3	36.7 ± 8.9	89.6 ± 0.3	51.7 ± 1.3	$\mathbf{1.3 \pm 0.7}$
5000	57.7 ± 4.2	41.7 ± 1.7	40.7 ± 1.7	40.1 ± 11.5	89.7 ± 0.2	52.0 ± 0.9	$\mathbf{1.0 \pm 0.4}$
8991	55.8 ± 3.5	41.3 ± 0.8	40.6 ± 0.8	33.3 ± 4.8	89.8 ± 0.2	51.6 ± 0.7	$\mathbf{0.8 \pm 0.3}$

Table 22: Air Quality: relative full-matrix Frobenius covariance error (mean $\pm$ sample standard deviation).

n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	2.140 ± 0.166	0.990 ± 0.031	$\mathbf{0.085 \pm 0.033}$
1000	2.133 ± 0.133	0.980 ± 0.022	$\mathbf{0.046 \pm 0.017}$
2000	2.130 ± 0.092	0.985 ± 0.015	$\mathbf{0.034 \pm 0.017}$
5000	2.090 ± 0.052	0.985 ± 0.009	$\mathbf{0.025 \pm 0.011}$
8991	2.117 ± 0.020	0.984 ± 0.006	$\mathbf{0.016 \pm 0.006}$

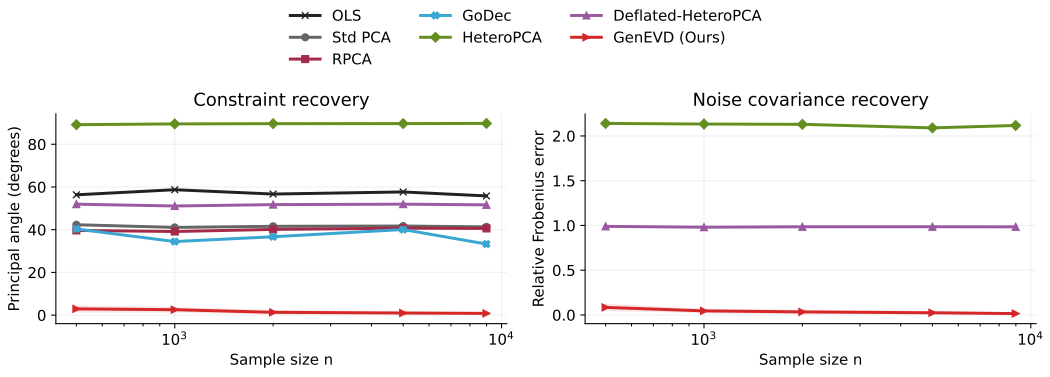


Figure 3: Air Quality: means over the saved trials. The red band denotes one standard deviation for GenEVD; tables report standard deviations for every method.

M.2.6 APPLIANCES

The reference contains 19735 rows; each grid point uses 30 trials. The sample-size grid is $500, 1000, 2000, 5000, 10000, 19735$. The noise-scale vector is $(1.1, 1, 0.9, 0.85, 1.2)$, the base seed is 6120000, and $\rho = 0.65$. For the ordered noisy coordinates $J = (1, 2, 5)$, the true covariance has the following block (all other entries are zero; displayed values are rounded):

$$[\mathbf{\Sigma}_\epsilon]_{JJ} = \begin{bmatrix} 0.43 & 0.27 & 0.29 \\ 0.27 & 0.41 & 0.29 \\ 0.29 & 0.29 & 0.47 \end{bmatrix}.$$

Table 23: Appliances: principal angle in degrees (mean $\pm$ sample standard deviation).

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	14.2 ± 3.5	21.8 ± 4.0	26.4 ± 3.5	36.0 ± 9.4	85.4 ± 4.2	50.4 ± 2.5	$\mathbf{3.1 \pm 1.9}$
1000	13.9 ± 3.2	20.5 ± 3.4	23.8 ± 3.5	30.2 ± 8.0	86.1 ± 0.8	50.5 ± 1.9	$\mathbf{2.1 \pm 0.9}$
2000	13.4 ± 0.8	20.1 ± 1.9	22.1 ± 1.8	30.7 ± 8.8	86.1 ± 0.5	51.0 ± 1.2	$\mathbf{1.5 \pm 0.7}$
5000	13.2 ± 0.7	20.0 ± 1.3	21.0 ± 1.3	33.8 ± 17.8	86.1 ± 0.3	50.8 ± 0.9	$\mathbf{0.9 \pm 0.5}$
10000	13.0 ± 0.4	20.1 ± 0.8	20.5 ± 0.8	30.4 ± 9.9	86.1 ± 0.2	50.5 ± 0.6	$\mathbf{0.6 \pm 0.3}$
19735	13.1 ± 0.3	19.9 ± 0.5	20.2 ± 0.5	32.7 ± 9.0	86.1 ± 0.2	50.7 ± 0.4	$\mathbf{0.5 \pm 0.2}$

Table 24: Appliances: relative full-matrix Frobenius covariance error (mean $\pm$ sample standard deviation).

n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	1.234 ± 0.103	0.796 ± 0.011	$\mathbf{0.068 \pm 0.039}$
1000	1.275 ± 0.044	0.799 ± 0.008	$\mathbf{0.042 \pm 0.024}$
2000	1.275 ± 0.040	0.801 ± 0.005	$\mathbf{0.031 \pm 0.017}$
5000	1.272 ± 0.022	0.800 ± 0.004	$\mathbf{0.024 \pm 0.011}$
10000	1.271 ± 0.018	0.800 ± 0.002	$\mathbf{0.017 \pm 0.008}$
19735	1.270 ± 0.009	0.800 ± 0.002	$\mathbf{0.011 \pm 0.006}$

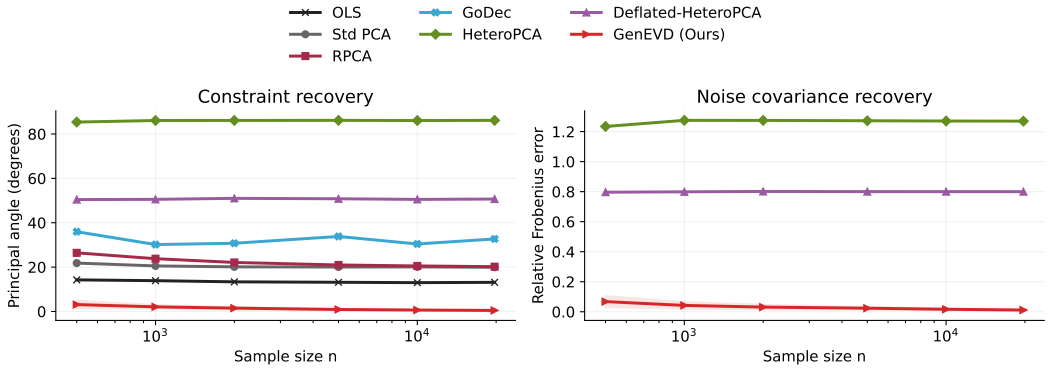


Figure 4: Appliances: means over the saved trials. The red band denotes one standard deviation for GenEVD; tables report standard deviations for every method.

M.2.7 GAS DRIFT

The reference contains 13910 rows; each grid point uses 30 trials. The sample-size grid is $500, 1000, 2000, 5000, 10000, 13910$. The noise-scale vector is $(1.25, 1.15, 0.9, 1.05, 0.85)$, the base seed is 6210000, and $\rho = 0.65$. For the ordered noisy coordinates $J = (1, 2, 4)$, the true covariance has the following block (all other entries are zero; displayed values are rounded):

$$[\mathbf{\Sigma}_\epsilon]_{JJ} = \begin{bmatrix} 0.45 & 0.28 & 0.28 \\ 0.28 & 0.41 & 0.26 \\ 0.28 & 0.26 & 0.41 \end{bmatrix}.$$

Table 25: Gas Drift: principal angle in degrees (mean $\pm$ sample standard deviation).

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	10.3 ± 1.3	9.6 ± 1.4	9.9 ± 1.2	12.4 ± 2.8	74.9 ± 5.0	13.3 ± 1.4	$\mathbf{2.0 \pm 0.7}$
1000	10.3 ± 1.7	9.2 ± 1.0	9.6 ± 0.9	11.3 ± 1.5	76.2 ± 3.1	13.7 ± 1.1	$\mathbf{1.2 \pm 0.7}$
2000	10.3 ± 1.2	9.0 ± 0.7	9.2 ± 0.6	10.8 ± 1.8	75.5 ± 3.2	13.4 ± 0.9	$\mathbf{0.8 \pm 0.4}$
5000	10.0 ± 0.6	9.0 ± 0.5	9.2 ± 0.5	11.5 ± 1.7	76.4 ± 2.0	13.3 ± 0.5	$\mathbf{0.5 \pm 0.2}$
10000	9.9 ± 0.3	9.0 ± 0.3	9.1 ± 0.3	10.8 ± 1.6	76.6 ± 0.9	13.4 ± 0.3	$\mathbf{0.4 \pm 0.2}$
13910	9.9 ± 0.1	9.0 ± 0.2	9.1 ± 0.2	11.7 ± 1.8	76.5 ± 0.6	13.3 ± 0.2	$\mathbf{0.4 \pm 0.1}$

Table 26: Gas Drift: relative full-matrix Frobenius covariance error (mean $\pm$ sample standard deviation).

n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	3.031 ± 1.032	0.789 ± 0.008	$\mathbf{0.066 \pm 0.022}$
1000	2.408 ± 0.665	0.790 ± 0.005	$\mathbf{0.051 \pm 0.028}$
2000	2.441 ± 0.716	0.792 ± 0.003	$\mathbf{0.029 \pm 0.011}$
5000	2.276 ± 0.362	0.791 ± 0.003	$\mathbf{0.023 \pm 0.013}$
10000	2.193 ± 0.062	0.791 ± 0.001	$\mathbf{0.014 \pm 0.006}$
13910	2.197 ± 0.032	0.791 ± 0.001	$\mathbf{0.013 \pm 0.007}$

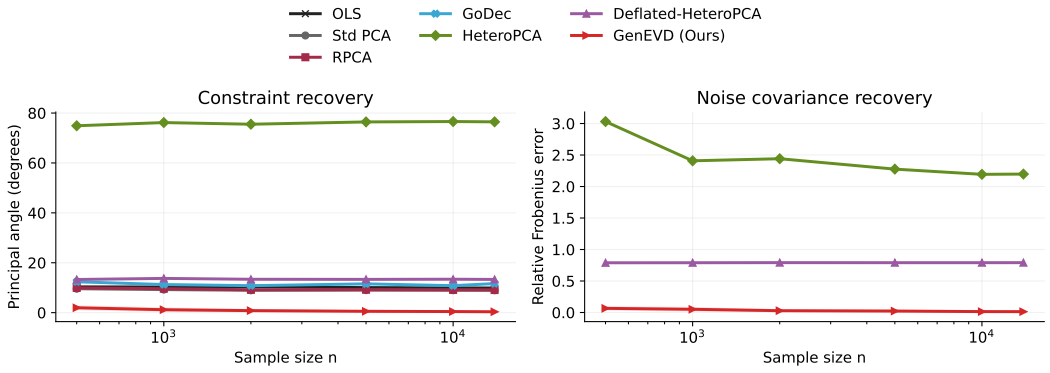


Figure 5: Gas Drift: means over the saved trials. The red band denotes one standard deviation for GenEVD; tables report standard deviations for every method.

M.2.8 MHEALTH

The reference contains 35174 rows; each grid point uses 30 trials. The sample-size grid is $500, 1000, 2000, 5000, 10000$. The noise-scale vector is $(1.2, 0.95, 1.05, 1.1, 1)$, the base seed is 6320000, and $\rho = 0.65$. For the ordered noisy coordinates $J = (1, 2, 3)$, the true covariance has the following block (all other entries are zero; displayed values are rounded):

$$[\mathbf{\Sigma}_\epsilon]_{JJ} = \begin{bmatrix} 0.42 & 0.23 & 0.22 \\ 0.23 & 0.29 & 0.18 \\ 0.22 & 0.18 & 0.27 \end{bmatrix}.$$

Table 27: MHEALTH: principal angle in degrees (mean $\pm$ sample standard deviation).

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	13.3 ± 0.8	9.6 ± 0.8	8.9 ± 0.9	9.2 ± 1.2	63.6 ± 0.9	78.7 ± 1.2	$\mathbf{1.7 \pm 0.7}$
1000	13.3 ± 0.7	9.8 ± 0.5	9.3 ± 0.6	9.2 ± 1.0	63.3 ± 0.5	78.7 ± 0.8	$\mathbf{1.1 \pm 0.3}$
2000	13.1 ± 0.5	9.5 ± 0.5	9.2 ± 0.5	9.2 ± 0.7	63.3 ± 0.4	78.4 ± 0.8	$\mathbf{0.7 \pm 0.3}$
5000	13.1 ± 0.2	9.5 ± 0.3	9.2 ± 0.3	8.8 ± 1.0	63.2 ± 0.2	78.3 ± 0.4	$\mathbf{0.5 \pm 0.2}$
10000	13.2 ± 0.2	9.5 ± 0.2	9.4 ± 0.2	9.2 ± 1.2	63.2 ± 0.2	78.3 ± 0.3	$\mathbf{0.4 \pm 0.2}$

Table 28: MHEALTH: relative full-matrix Frobenius covariance error (mean $\pm$ sample standard deviation).

n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	5.874 ± 0.402	2.690 ± 0.201	$\mathbf{0.065 \pm 0.034}$
1000	5.961 ± 0.232	2.702 ± 0.116	$\mathbf{0.045 \pm 0.019}$
2000	5.956 ± 0.224	2.705 ± 0.110	$\mathbf{0.037 \pm 0.019}$
5000	5.998 ± 0.115	2.714 ± 0.061	$\mathbf{0.022 \pm 0.012}$
10000	5.988 ± 0.100	2.713 ± 0.048	$\mathbf{0.013 \pm 0.006}$

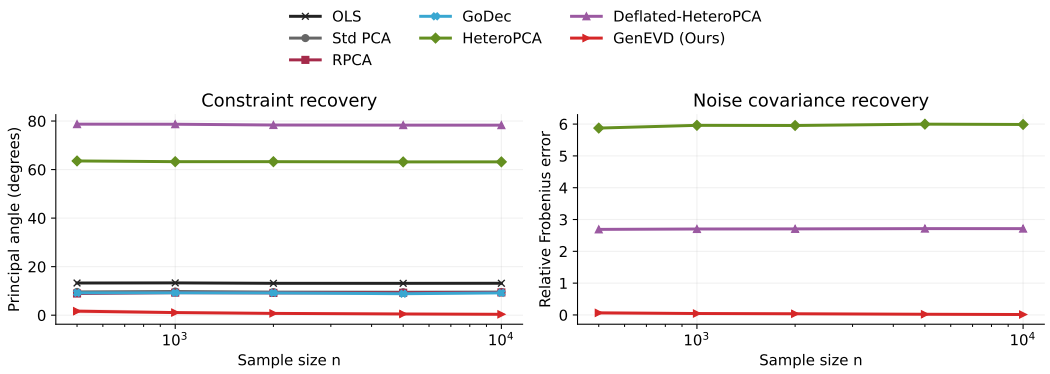


Figure 6: MHEALTH: means over the saved trials. The red band denotes one standard deviation for GenEVD; tables report standard deviations for every method.

M.2.9 POWER PLANT

The reference contains 9568 rows; each grid point uses 30 trials. The sample-size grid is $500, 1000, 2000, 5000, 9568$. The noise-scale vector is $(1.2, 1, 0.85, 0.95, 1.1)$, the base seed is 6410000, and $\rho = 0.65$. For the ordered noisy coordinates $J = (1, 2, 4)$, the true covariance has the following block (all other entries are zero; displayed values are rounded):

$$[\mathbf{\Sigma}_\epsilon]_{JJ} = \begin{bmatrix} 0.48 & 0.26 & 0.26 \\ 0.26 & 0.33 & 0.22 \\ 0.26 & 0.22 & 0.34 \end{bmatrix}.$$

Table 29: Power Plant: principal angle in degrees (mean ± sample standard deviation).

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	29.8 ± 1.6	21.2 ± 2.3	20.4 ± 2.5	21.3 ± 9.4	87.6 ± 1.3	71.3 ± 2.9	$\mathbf{2.1 \pm 1.2}$
1000	30.1 ± 1.2	21.8 ± 1.6	21.4 ± 1.6	18.3 ± 6.7	87.8 ± 1.1	70.9 ± 1.8	$\mathbf{1.5 \pm 0.8}$
2000	30.1 ± 0.7	21.6 ± 0.8	21.4 ± 0.8	20.7 ± 9.1	87.7 ± 1.1	71.4 ± 1.2	$\mathbf{1.1 \pm 0.5}$
5000	30.2 ± 0.4	21.9 ± 0.7	21.7 ± 0.7	24.8 ± 13.6	87.8 ± 0.8	71.0 ± 0.9	$\mathbf{0.7 \pm 0.3}$
9568	30.2 ± 0.3	21.8 ± 0.4	21.7 ± 0.4	22.5 ± 10.6	87.8 ± 0.6	70.9 ± 0.6	$\mathbf{0.5 \pm 0.2}$

Table 30: Power Plant: relative full-matrix Frobenius covariance error (mean ± sample standard deviation).

n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	1.481 ± 0.106	1.668 ± 0.088	$\mathbf{0.072 \pm 0.026}$
1000	1.425 ± 0.088	1.664 ± 0.071	$\mathbf{0.047 \pm 0.030}$
2000	1.424 ± 0.059	1.663 ± 0.036	$\mathbf{0.030 \pm 0.015}$
5000	1.397 ± 0.026	1.667 ± 0.030	$\mathbf{0.021 \pm 0.012}$
9568	1.399 ± 0.020	1.673 ± 0.018	$\mathbf{0.015 \pm 0.011}$

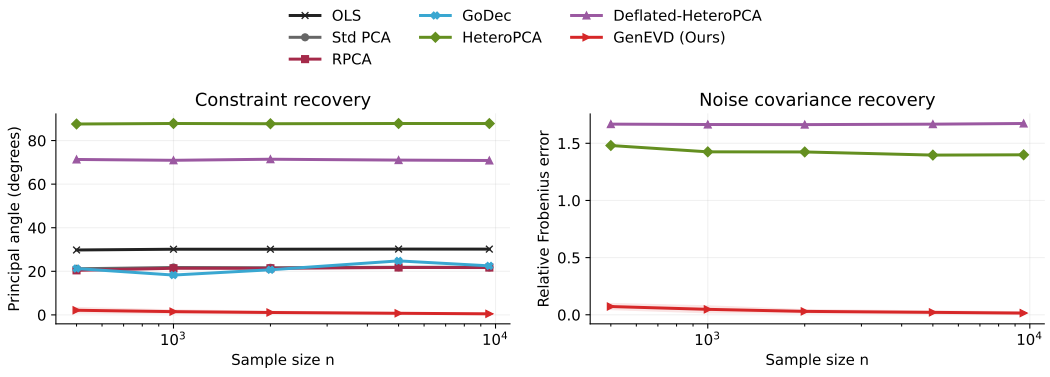


Figure 7: Power Plant: means over the saved trials. The red band denotes one standard deviation for GenEVD; tables report standard deviations for every method.

M.2.10 HOUSEHOLD

The reference contains 2049280 rows; each grid point uses 30 trials. The sample-size grid is $500, 1000, 2000, 5000, 10000, 20000$. The noise-scale vector is $(1.25, 0.85, 1.2, 1, 0.95)$, the base seed is 6510000, and $\rho = 0.65$. For the ordered noisy coordinates $J = (1, 2, 5)$, the true covariance has the following block (all other entries are zero; displayed values are rounded):

$$[\mathbf{\Sigma}_\epsilon]_{JJ} = \begin{bmatrix} 0.46 & 0.14 & 0.24 \\ 0.14 & 0.11 & 0.12 \\ 0.24 & 0.12 & 0.30 \end{bmatrix}.$$

Table 31: Household: principal angle in degrees (mean $\pm$ sample standard deviation).

n	OLS	Std PCA	RPCA	GoDec	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	18.3 ± 1.5	13.7 ± 3.0	16.8 ± 4.9	36.0 ± 26.0	76.1 ± 4.0	80.4 ± 16.0	$\mathbf{2.3 \pm 1.2}$
1000	18.3 ± 0.8	13.8 ± 1.6	15.6 ± 2.2	39.3 ± 26.8	74.9 ± 2.3	84.3 ± 2.9	$\mathbf{1.5 \pm 0.8}$
2000	18.2 ± 1.4	13.8 ± 1.3	15.0 ± 1.8	38.4 ± 27.2	75.6 ± 1.6	85.5 ± 2.8	$\mathbf{0.8 \pm 0.2}$
5000	18.1 ± 0.4	13.2 ± 0.7	13.8 ± 0.8	40.3 ± 27.8	75.9 ± 1.1	84.9 ± 1.9	$\mathbf{0.6 \pm 0.2}$
10000	18.1 ± 0.3	13.3 ± 0.6	13.7 ± 0.7	40.2 ± 24.8	75.5 ± 0.8	84.8 ± 1.7	$\mathbf{0.4 \pm 0.2}$
20000	18.1 ± 0.2	13.2 ± 0.4	13.5 ± 0.4	50.9 ± 25.5	75.8 ± 0.7	85.1 ± 1.6	$\mathbf{0.3 \pm 0.1}$

Table 32: Household: relative full-matrix Frobenius covariance error (mean $\pm$ sample standard deviation).

n	HeteroPCA	Deflated-HeteroPCA	GenEVD (Ours)
500	2.374 ± 0.304	1.903 ± 0.346	$\mathbf{0.066 \pm 0.044}$
1000	2.435 ± 0.244	1.991 ± 0.276	$\mathbf{0.044 \pm 0.015}$
2000	2.370 ± 0.131	1.959 ± 0.203	$\mathbf{0.032 \pm 0.019}$
5000	2.352 ± 0.108	1.999 ± 0.163	$\mathbf{0.021 \pm 0.010}$
10000	2.348 ± 0.077	1.960 ± 0.157	$\mathbf{0.014 \pm 0.007}$
20000	2.339 ± 0.033	1.952 ± 0.130	$\mathbf{0.009 \pm 0.004}$

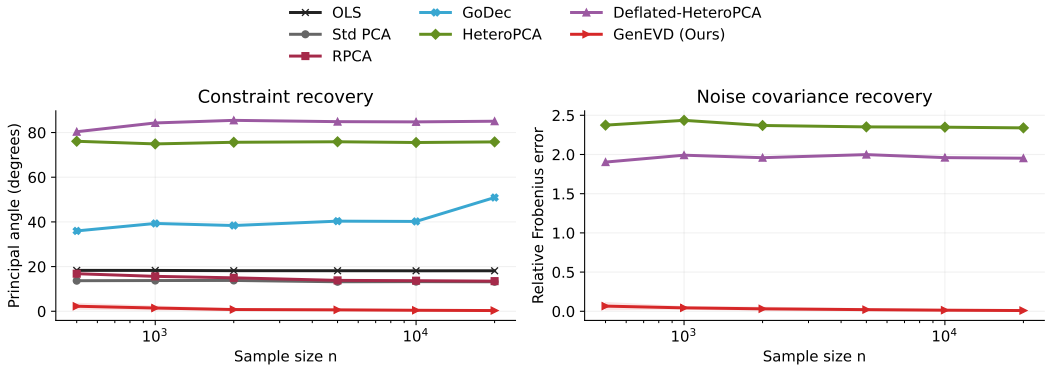


Figure 8: Household: means over the saved trials. The red band denotes one standard deviation for GenEVD; tables report standard deviations for every method.