WORST-CASE SPARSE PRINCIPAL SUBSPACE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse principal component analysis is normally posed as the maximization of an aggregate of captured variance—a single Rayleigh quotient, or the trace of the projected covariance—subject to a sparsity constraint. This choice matters when more than one component is required. An aggregate objective is dominated by its leading term, so a shared feature budget is spent first on the strongest components, and the weakest is returned as little more than an arbitrary direction. We make this precise by exhibiting covariances for which every maximizer of the trace criterion returns a degenerate subspace, one of whose directions carries no variance at all, while a worst-case objective yields a subspace in which every component keeps a fixed share of what it has in the population. Motivated by this, we study sparse principal subspace estimation under the worst-case componentwise retention ratio, defined as the smallest, across the population directions we wish to recover, of the fractions of variance that survive the restriction to the selected features. We show that this criterion is attained exactly by a choice of support, that it is largest precisely at the target eigenspace, and that taking the worst case over the individual columns of the estimated basis instead would give nothing new, since the Schur–Horn theorem reduces it to the trace criterion. To solve the resulting non-convex max–min program we develop TRIM, built on minorization–maximization (MM). We apply a tangent-plane minorizer to each of the eigenvalue functions, solve the resulting subproblem on each candidate support in closed form by a polar factor, and accept a support only if it raises the exact objective. TRIM needs no tuning parameters and no relaxation (in the sense of losing tightness), and it does not build the subspace one direction at a time by projecting out what it has already found. It is monotone by construction, terminates finitely, and reports an a posteriori bound on its own optimality gap. It succeeds on general covariances for which one-shot thresholding provably fails, it outperforms the competing estimators on a real corpus at every sparsity level we test, and under a limited feature budget it keeps far more of the weakest component—the one an aggregate objective neglects.

1 INTRODUCTION

Principal component analysis is the default instrument for exploring high-dimensional data, but when the ambient dimension d is comparable to or larger than the sample size n the empirical principal directions are inconsistent [15]. Sparse PCA restores consistency by positing that the leading eigenvectors are supported on a small subset of $s \ll d$ coordinates, which reduces the sample complexity from order d to roughly $s \log d$ and—at least as importantly in practice—makes the recovered directions readable, since each component is a combination of a few named features [16, 31].

Two families of algorithms have grown up around the resulting problem. Combinatorial methods—diagonal thresholding, covariance thresholding, iterative thresholding and the truncated power method—buy their speed with an assumption, since their guarantees rest on hypotheses such as the *spiked identity* model, an isotropic background plus a single rank-one spike, or a local initialization, and without such a hypothesis they carry no assurance at all [1, 4, 19, 30]. Convex methods, chiefly the semidefinite and Fantope relaxations (in the sense of losing tightness), dispense with the spiked assumption at a cost that grows with the ambient dimension, since they carry a $d \times d$ decision variable and order d^2 constraints [7, 29]. State-of-the-art solvers for such programs run in time at least $d^{2\omega}$

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.

in theory, where $\omega \approx 2.37$ is the exponent of matrix multiplication, so no less than $d^{4.5}$, which is impractical at even moderate dimension.

Whether the lightweight guarantees were merely unproven or genuinely false was settled recently, and the answer is the second. There are explicit general covariances, meeting the minimal requirement that the top eigenvector be s-sparse and that the second eigenvalue be bounded away from the first, on which diagonal thresholding, covariance thresholding and a planted-clique-style greedy heuristic each fail with constant probability [17]. Repair for a single component came with the same construction, a *restarted* truncated power method (RTPM) started from every coordinate basis vector with an oversampled truncation level. Beyond one component the picture changes, since a sparse direction arbitrarily well aligned with the top eigenvector can leave behind a residual matrix whose own leading eigenvector is fully dense. The residual problem is then no longer an instance of sparse PCA, so the recursion that extracts components one at a time cannot be closed [14, 20], and a provable combinatorial method for sparse k-PCA has remained open.

This paper takes up the multi-component setting, and our starting point is that the difficulty is not only algorithmic but also a matter of *what is being maximized*. Almost universally, multi-component sparse PCA maximizes an aggregate of captured variance, either the trace $\mathrm{tr}(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V})$ of the projected sample covariance over a row-sparse orthonormal $\mathbf{V}$ [25, 28], or a sequence of single Rayleigh quotients on residual matrices. Aggregation is exactly what fails in the multi-class discriminant analysis literature, where averaging pairwise separations lets the best-separated pair dominate and pushes the hardest pair into overlap, and where max–min reformulations were introduced to repair it [3]. The same imbalance arises in sparse PCA. A trace objective is a sum whose largest term is λ_1 (the leading eigenvalue of the sample covariance), so when the k components must share one feature budget of size s, spending that budget on the strongest component buys more trace than spreading it, and the weakest components are simply sacrificed. We make this exact by exhibiting covariances for which *every* maximizer of the trace criterion returns a subspace one of whose directions carries no variance at all, whereas a worst-case objective yields one in which every one of the k components keeps a fixed share of what it has in the population. Since s is chosen by the practitioner for interpretability and is rarely as large as the true union support, this is the regime that matters in practice, and it is not a corner case.

We therefore study sparse principal subspace estimation under the *worst-case componentwise retention ratio*

$$\widehat{R}(\mathbf{V}) \;=\; \min_{j\in[k]} \; \frac{\lambda_j\Big(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}\Big)}{\lambda_j(\widehat{\mathbf{\Sigma}})}, \qquad \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k, \quad \|\mathbf{V}\|_{2,0} \le s, \tag{1}$$

which reads off, for each of the k leading principal directions, the fraction of its variance that survives the restriction to s features, and reports the worst of them. The criterion is bounded by one, equals one exactly when the sparse subspace is a top-k eigenspace, and for $k = 1$ reduces to ordinary sparse PCA, so it covers the single-component case without changing the problem.

Our algorithmic contribution is a method built on minorization–maximization (MM) [13, 18, 23] that attacks this criterion directly, with no relaxation (in the sense of losing tightness) and no sequential extraction. Each of the k eigenvalue functions $\mathbf{V} \mapsto \lambda_j(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V})$ admits a tight minorizer obtained by expanding a convex quadratic form about the current iterate, and the resulting surrogate is linear in $\mathbf{V}$. On any fixed support, maximizing a linear function over the row-sparse Stiefel manifold is solved in closed form by a polar factor, and a candidate support is accepted only if it raises the exact objective. For $k = 1$ the update is one truncated power step, hard-threshold $\widehat{\mathbf{\Sigma}}\mathbf{v}^t$ and renormalize, and our framework extends it to the subspace setting in which the recursion above breaks. We call the resulting method **TRIM** (Truncated Rayleigh Iterations via Minorization), which is free of tuning parameters (it is worth mentioning that stopping criteria are not considered hyperparameters to tune [23]).

CONTRIBUTIONS

The main contributions of this work are summarized as follows.

- **A criterion that cannot neglect a component.** We propose the worst-case componentwise retention ratio as the objective for sparse principal subspace estimation. To our knowledge no

criterion that takes the worst case over the components themselves has been proposed for this problem before, and we establish the properties that make it usable here. It is attained exactly by a choice of support, so the only genuinely combinatorial decision is which features to keep. It is bounded above, and attains its bound precisely at the target eigenspace, so its value is an absolute measure of quality. Replacing the orthonormality constraint on the estimated basis by an inequality leaves the optimal value unchanged, which is what later allows the inner problem to be solved in closed form on every support. We also show that the apparently natural alternative, taking the worst case over the individual columns of the basis instead of the subspace they span, is vacuous, since the Schur–Horn theorem reduces it to the trace criterion.

- **An algorithm that attacks the criterion directly.** We develop TRIM, which optimizes the resulting max–min program with no relaxation (in the sense of losing tightness) and without extracting the components one at a time. It is monotone by construction, it terminates after finitely many iterations, and it reports at no extra cost how far its answer can be from the optimum. That bound is a dividend of normalizing the criterion, which places every component on one absolute scale.
- **A statistical guarantee under a general covariance.** We prove that any global maximizer of the criterion recovers the leading eigenspace from a number of samples growing logarithmically in the dimension and linearly in the budget, with no spiked assumption, no incoherence condition and no requirement on the initialization.
- **Experimental support on synthetic and real data.** On the single-component problem TRIM matches the strongest combinatorial and convex baselines, so moving to the worst-case criterion costs nothing where the usual objective is already adequate. On the subspace problem under a limited feature budget it recovers far more of the weakest component than the aggregate estimators it replaces. On the 20 Newsgroups corpus it improves held-out worst-component retention at equal sparsity, ahead of every state-of-the-art and benchmark method we compare against.

Notation. Vectors are lowercase boldface, matrices uppercase boldface, and $[n] = \{1, \dots, n\}$. We write $\mathbb{S}^{d\times d}$ for the real symmetric matrices, $\mathbb{S}^{d\times d}_{\succcurlyeq 0}$ for its cone of positive semidefinite matrices, and $\lambda_1(\mathbf{M}) \geq \cdots \geq \lambda_d(\mathbf{M})$ for the eigenvalues in decreasing order; $\mathbf{M}_S$ is the principal submatrix $\mathbf{M}_{S\times S}$. For $\mathbf{V} \in \mathbb{R}^{d\times k}$, $\operatorname{rsupp}(\mathbf{V})$ is its set of non-zero rows and $\|\mathbf{V}\|_{2,0} = |\operatorname{rsupp}(\mathbf{V})|$. The operator, Frobenius and nuclear norms are $\|\cdot\|_{\mathrm{op}}, \|\cdot\|_F, \|\cdot\|_*$, $\langle\cdot,\cdot\rangle$ is the trace inner product, $\preccurlyeq$ the Loewner order, $\operatorname{top}_s(\mathbf{b})$ zeroes all but the s largest entries of $\mathbf{b}$ in magnitude, and $\sin^2\angle(\mathbf{V},\mathbf{V}') = 1 - \|\mathbf{V}'^\top\mathbf{V}\|_F^2/k$. We write $O(\cdot)$ and $\Omega(\cdot)$ for the usual asymptotic upper and lower bounds, and $a \lesssim b$ ($a \gtrsim b$) when $a \leq Cb$ ($a \geq Cb$) for an absolute constant C.

2 Problem formulation

The setting. We work in the general sparse k-PCA setting, which assumes only what well-posedness requires, namely an eigengap and a sparse leading eigenspace.

Setting 1 (General sparse k-PCA)**.** Fix $(s, d, n, k) \in \mathbb{N}^4$ with $(s, k) \in [d]^2$, $\gamma \in (0, \frac{1}{2})$ and $\sigma > 0$. There are unknown orthonormal $\{\mathbf{v}_i \in \mathbb{R}^d\}_{i\in[k]}$ spanning the top-k eigenspace of an unknown $\mathbf{\Sigma} \in \mathbb{S}^{d\times d}_{\succcurlyeq 0}$ with $\lambda_{k+1}(\mathbf{\Sigma})/\lambda_k(\mathbf{\Sigma}) \leq 1 - \gamma$ and $|S^\star| \leq s$, where $S^\star := \bigcup_{i\in[k]} \operatorname{supp}(\mathbf{v}_i)$. We observe $\{\mathbf{x}_i\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathcal{D}$, a σ-sub-Gaussian distribution with covariance $\mathbf{\Sigma}$, and write $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i^\top$.

Taking $\mathbf{\Sigma} \propto (1-\gamma)\mathbf{I}_d + \gamma\mathbf{v}\mathbf{v}^\top$ recovers the spiked identity model, so Setting 1 strictly generalizes it, since the background spectrum may be anything below the gap and need not be isotropic.

The worst-case criterion. Let $\mathcal{V}_s^k := \{\mathbf{V} \in \mathbb{R}^{d\times k} : \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k,\ \|\mathbf{V}\|_{2,0} \leq s\}$ denote the row-sparse Stiefel manifold, and set $\kappa_j := \lambda_j(\widehat{\mathbf{\Sigma}})$ for $j \in [k]$. We study

$$(\mathcal{P}): \qquad \max_{\mathbf{V}\in\mathcal{V}_s^k} \widehat{R}(\mathbf{V}), \qquad \widehat{R}(\mathbf{V}) := \min_{j\in[k]} \frac{\lambda_j(\mathbf{V}^\top\widehat{\mathbf{\Sigma}}\mathbf{V})}{\kappa_j}. \tag{2}$$

The quantity $\lambda_j(\mathbf{V}^\top\widehat{\mathbf{\Sigma}}\mathbf{V})/\kappa_j$ is the fraction of the j-th principal variance of $\widehat{\mathbf{\Sigma}}$ that survives the restriction to the s selected features, and $\widehat{R}$ reports the worst of these k fractions. The normalizers κ_j

are computed once, by a k-truncated eigendecomposition of $\widehat{\mathbf{\Sigma}}$; they are not tuning parameters. For $k = 1$, $(\mathcal{P})$ is ordinary sparse PCA and is therefore NP-hard [21]. The following properties fix the geometry of $(\mathcal{P})$.

Proposition 1 (Exact support reduction, and an inequality constraint that loses nothing)**.** *Let $k \le s \le d$ and $\widehat{\mathbf{\Sigma}} \in \mathbb{S}^{d\times d}_{\succcurlyeq 0}$.*

(i) For every $S \subseteq [d]$ with $|S| \le s$, the top-k eigenvectors of $\widehat{\mathbf{\Sigma}}_S$, embedded in $\mathbb{R}^d$, simultaneously maximize $\lambda_j(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V})$ for every $j \in [k]$ among all $\mathbf{V} \in \mathcal{V}_s^k$ with $\mathrm{rsupp}(\mathbf{V}) \subseteq S$. Consequently

$$\max_{\mathbf{V}\in\mathcal{V}_s^k} \widehat{R}(\mathbf{V}) \;=\; \max_{|S|\le s} \min_{j\in[k]} \frac{\lambda_j(\widehat{\mathbf{\Sigma}}_S)}{\kappa_j}. \tag{3}$$

(ii) $\widehat{R}(\mathbf{V}) \le 1$ for every $\mathbf{V} \in \mathcal{V}_s^k$, with equality if and only if the column space of $\mathbf{V}$ is a top-k eigenspace of $\widehat{\mathbf{\Sigma}}$.

(iii) Replacing the constraint $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_k$ by $\mathbf{V}^\top\mathbf{V} \preccurlyeq \mathbf{I}_k$ leaves the optimal value of $(\mathcal{P})$ unchanged (for the proof, see Appendix E).

Part (i) says the only genuinely combinatorial decision is the choice of support. Once S is fixed the optimal subspace is an ordinary $|S|$-dimensional eigenvalue problem, and—crucially—the *same* $\mathbf{V}$ is optimal for all k eigenvalue functions at once, so the max–min structure creates no conflict within a support. Part (ii) says the criterion is normalized and identifies the target exactly, and part (iii) is what lets us solve the MM subproblem on each support in closed form.

The worst case must be taken over the subspace, not the columns. Taking the worst case over the k *columns* of $\mathbf{V}$ instead, i.e. maximizing $\min_j \mathbf{v}_j^\top \widehat{\mathbf{\Sigma}} \mathbf{v}_j$, is vacuous. By the Schur–Horn theorem [11] the basis inside the estimated subspace can always be rotated so that all k diagonal entries of $\mathbf{V}^\top\widehat{\mathbf{\Sigma}}\mathbf{V}$ equal their average, which reduces the column-wise worst case to the trace criterion divided by k (Proposition 8, Appendix E). The subspace-level quantities $\lambda_j(\mathbf{V}^\top\widehat{\mathbf{\Sigma}}\mathbf{V})$ used in (2) are rotation-invariant and do not degenerate this way.

A limited budget separates the two criteria. When the budget is large enough that $s \ge |S^\star|$, the trace criterion and $\widehat{R}$ are both maximized at $S^\star$ in the population, and Theorem 7 shows that the worst-case estimator then recovers the target eigenspace. In practice, however, s is chosen by the practitioner so that components stay readable, and is rarely as large as the unknown union support; in that regime the two criteria diverge sharply.

Proposition 2 (Degeneracy of the trace criterion under a shared budget)**.** *Let $k \ge 2$ and let $\lambda_1 > \lambda_2 \ge \cdots \ge \lambda_k > 0$. Put $\mathbf{v}_j = (\mathbf{e}_{2j-1} + \mathbf{e}_{2j})/\sqrt{2}$ for $j \in [k]$, so that $|S^\star| = 2k$, let $\mathbf{\Sigma} = \sum_{j\in[k]} \lambda_j \mathbf{v}_j\mathbf{v}_j^\top$ and take the budget $s = k$. Then the following hold.*

(i) every maximizer S_{tr} of the trace criterion $S \mapsto \sum_{i\le k} \lambda_i(\mathbf{\Sigma}_S)$ satisfies $\lambda_k(\mathbf{\Sigma}_{S_{\mathrm{tr}}}) = 0$, so the estimated subspace is degenerate, one of its directions carrying no variance at all, and $R = 0$;

(ii) every maximizer of R satisfies $\lambda_j(\mathbf{\Sigma}_S)/\lambda_j = \frac{1}{2}$ for every $j \in [k]$, so $R = \frac{1}{2}$ (for the proof, see Appendix E).

The mechanism is transparent, since pairing two coordinates of one component turns one of them into a hard zero eigenvalue but buys the full λ_j in place of $\lambda_j/2$—a good trade for the trace and a ruinous one for the worst case. Nothing here depends on sampling noise, and Remark 9 carries the same construction to every intermediate budget, not just to $s = k$.

3 Truncated Rayleigh iterations via minorization

Problem $(\mathcal{P})$ is a max–min program over a non-convex set, equivalent by Proposition 1(i) to a combinatorial selection among $\binom{d}{s}$ supports. Our approach is built on the minorization–maximization (MM) principle [13, 18, 23] (for an overview of MM, see Appendix C). At the current iterate $\mathbf{V}^t$

we replace $\widehat{R}$ by a surrogate that lies below it everywhere and touches it at $\mathbf{V}^t$, take its maximizers on candidate supports as proposals, keep one only if it raises $\widehat{R}$, and repeat. Since the surrogate is a global lower bound tight at $\mathbf{V}^t$, any point improving it improves $\widehat{R}$, so its maximizers are well chosen, and the acceptance test keeps the scheme monotone without a step size, a penalty parameter or a line search. Everything rests on the following minorizer.

3.1 A TIGHT MINORIZER FOR EVERY EIGENVALUE FUNCTION

Lemma 3 (Tangent minorizer). *Let $\widehat{\mathbf{\Sigma}} \in \mathbb{S}^{d\times d}_{\succcurlyeq 0}$ and $\mathbf{V}^t \in \mathcal{V}^k_s$, write $\mathbf{M}^t = (\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V}^t$ with eigenpairs $(\mu_1 \geq \cdots \geq \mu_k,\ \mathbf{y}_1, \ldots, \mathbf{y}_k)$, and set $\mathbf{U}_j = [\mathbf{y}_1, \ldots, \mathbf{y}_j] \in \mathbb{R}^{k\times j}$. Then for every $\mathbf{V} \in \mathbb{R}^{d\times k}$ with $\mathbf{V}^\top \mathbf{V} \preccurlyeq \mathbf{I}_k$ and every $j \in [k]$,*

$$\lambda_j\Big(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}\Big) \ \geq\ g_j(\mathbf{V} \,|\, \mathbf{V}^t) := \lambda_{\min}\Big(\mathbf{U}_j^\top \Big[(\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V} + \mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}^t - \mathbf{M}^t\Big] \mathbf{U}_j\Big), \tag{4}$$

with equality at $\mathbf{V} = \mathbf{V}^t$, where $g_j(\mathbf{V}^t \,|\, \mathbf{V}^t) = \mu_j = \lambda_j(\mathbf{M}^t)$. Moreover $g_j(\cdot \,|\, \mathbf{V}^t)$ is concave (for the proof, see Appendix E).

The shape of the proof explains the algorithm. By Courant–Fischer, $\lambda_j(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V})$ is at least $\min_{\|\mathbf{y}\|_2=1} \|\widehat{\mathbf{\Sigma}}^{1/2} \mathbf{V} \mathbf{U}_j \mathbf{y}\|_2^2$; for fixed $\mathbf{y}$ the quantity inside is a convex quadratic in $\mathbf{V}$, hence lies above its tangent plane at $\mathbf{V}^t$, and taking the minimum over $\mathbf{y}$ of those tangent planes gives (4). Substituting it into $(\mathcal{P})$ gives the surrogate program

$$\mathbf{V}^{t+1} \in \arg\max_{\mathbf{V}\in\mathcal{V}^k_s} \ \min_{j\in[k]} \ \frac{g_j(\mathbf{V} \,|\, \mathbf{V}^t)}{\kappa_j}, \tag{5}$$

whose objective is concave in $\mathbf{V}$ but which is still constrained to the non-convex set $\mathcal{V}^k_s$. The next step removes the remaining inner minimum.

3.2 THE LINEARIZED SUBPROBLEM IS SOLVED IN CLOSED FORM ON EVERY SUPPORT

Writing $\lambda_{\min}(\mathbf{N}) = \min\{\langle \mathbf{N}, \mathbf{y}\mathbf{y}^\top \rangle : \|\mathbf{y}\|_2 = 1\}$ and absorbing the index j and the vector $\mathbf{y}$ into a single weight $\mathbf{Y} = \mathbf{U}_j \mathbf{y}\mathbf{y}^\top \mathbf{U}_j^\top \succcurlyeq \mathbf{0}$ with $\operatorname{tr} \mathbf{Y} = 1$, the objective of (5) becomes a pointwise minimum, over a compact weight set $\mathcal{W}^t$, of functions *affine* in $\mathbf{V}$,

$$\min_{j\in[k]} \frac{g_j(\mathbf{V} \,|\, \mathbf{V}^t)}{\kappa_j} \ =\ \min_{(j,\mathbf{Y})\in\mathcal{W}^t} \ \frac{1}{\kappa_j}\Big[\, 2\big\langle \widehat{\mathbf{\Sigma}} \mathbf{V}^t \mathbf{Y},\ \mathbf{V} \big\rangle - \langle \mathbf{Y}, \mathbf{M}^t \rangle\Big]. \tag{6}$$

For any fixed weight the maximization over $\mathbf{V}$ therefore reduces to maximizing a linear function on the row-sparse Stiefel manifold, which has a closed-form solution on every support.

Lemma 4 (Linear maximization on the row-sparse Stiefel manifold). *For any $\mathbf{B} \in \mathbb{R}^{d\times k}$ and $k \leq s \leq d$, $\max_{\mathbf{V}\in\mathcal{V}^k_s} \langle \mathbf{B}, \mathbf{V} \rangle = \max_{|S|\leq s} \|\mathbf{B}_S\|_*$, and an optimal $\mathbf{V}$ is the polar factor $\mathbf{P}\mathbf{Q}^\top$ of $\mathbf{B}_{S^\star}$, where $\mathbf{B}_{S^\star} = \mathbf{P} \operatorname{diag}(\sigma) \mathbf{Q}^\top$ is a thin singular value decomposition. In particular, for $k = 1$ the solution is $\operatorname{top}_s(\mathbf{B}) / \|\operatorname{top}_s(\mathbf{B})\|_2$ (for the proof, see Appendix E).*

This lemma has two consequences. Taking $k = 1$ and $\mathbf{B} = \widehat{\mathbf{\Sigma}} \mathbf{v}^t$—the only weight available when $k = 1$—the update is $\mathbf{v}^{t+1} = \operatorname{top}_s(\widehat{\mathbf{\Sigma}} \mathbf{v}^t) / \|\operatorname{top}_s(\widehat{\mathbf{\Sigma}} \mathbf{v}^t)\|_2$, which is precisely one step of the truncated power method [30], and the exchange moves introduced below take TRIM beyond that method already at $k = 1$ (for the comparison, see Appendix I).

For $k > 1$, by contrast, the inner maximization over S in Lemma 4 is *itself* combinatorial, because the nuclear norm of a submatrix is not additive across rows. The iterative schemes for the trace criterion choose their support by row truncation, keeping the s rows of largest Euclidean norm [25]. Applied to $\mathbf{B}$, this maximizes $\|\mathbf{B}_S\|_F$ in place of $\|\mathbf{B}_S\|_*$, and since $\|\mathbf{B}_S\|_F \leq \|\mathbf{B}_S\|_* \leq \sqrt{k} \|\mathbf{B}_S\|_F$ it solves our subproblem only up to a factor $\sqrt{k}$. Penalized formulations avoid the combinatorial step altogether, but only by replacing the feature budget with a sparsity penalty imposed column by column, whose weight must be chosen and which selects no common set of features [2]. We therefore treat row truncation as a *proposal*, and secure monotonicity by evaluating the exact objective on every support it suggests and accepting only strict improvements, at the price of one $s \times s$ eigendecomposition per candidate.

3.3 THE ALGORITHM

Fix an iterate supported on S^t and let $\mathbf{W}^t = [\mathbf{w}_1^t, \ldots, \mathbf{w}_k^t]$ be its principal directions, $\mathbf{w}_j^t = \mathbf{V}^t \mathbf{y}_j$, so that $(\mathbf{w}_j^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{w}_j^t = \mu_j$. Each weight $\mathbf{Y}$ in (6) produces one linear score and hence one candidate support. We take the vertices of the weight set, two informative interior points of it, and one further move that no choice of $\mathbf{Y}$ produces.

(a) **Componentwise moves** $\mathbf{Y} = \mathbf{y}_j \mathbf{y}_j^\top$, giving the candidate $S = \operatorname{top}_s((\widehat{\mathbf{\Sigma}} \mathbf{w}_j^t)^2)$ for each j, a truncated power step on the j-th principal direction.
(b) **A deficit-weighted move** $\mathbf{Y} = \sum_j (z_j/\kappa_j) \mathbf{y}_j \mathbf{y}_j^\top$ with $z_j \propto \exp(-(\rho_j - \min_i \rho_i)/\operatorname{range}(\rho))$ and $\rho_j = \mu_j/\kappa_j$, steering the budget towards the components furthest from full retention. This move has no analogue in trace-based schemes.
(c) **The Ky Fan move** $\mathbf{Y} \propto \mathbf{I}_k$, reproducing the trace-criterion step, so TRIM can never be beaten by a trace step at any single iteration.
(d) **Exchange moves**, swapping up to q of the least-used coordinates of S^t for the highest-scoring ones outside it, which repair the ties row truncation resolves arbitrarily.

Every candidate is scored by the *exact* objective $\min_j \lambda_j(\widehat{\mathbf{\Sigma}}_S)/\kappa_j$ via Proposition 1(i), and the incumbent is kept unless one of them strictly improves on it.

Restarts and screening. Since $(\mathcal{P})$ is non-convex, the initialization also matters, and we start the iteration from the d seeds $\operatorname{top}_s((\widehat{\mathbf{\Sigma}}_{i:})^2)$, one per coordinate i, and return the best result. For $k > 1$ we add one cascade seed S_{casc}, which is discarded unless it wins on the exact objective. Running all d seeds to convergence is wasteful, so we score them once and keep only the m most promising, m being the number of restarts retained (for the cascade, the screening rule and the computational cost, see Appendix F). A summary of our proposed method is given in Algorithm 1 (for the pseudocode, see Appendix D).

Theorem 5 (Monotone ascent, bounded objective, finite termination)**.** *Let $\{(\mathbf{V}^t, S^t)\}_{t \geq 0}$ be the iterates of one restart of Algorithm 1. Then* (i) *$\widehat{R}(\mathbf{V}^{t+1}) \geq \widehat{R}(\mathbf{V}^t)$ for every t, strictly unless the algorithm has stopped;* (ii) *the objective sequence $\{\widehat{R}(\mathbf{V}^t)\}_{t \geq 0}$ is non-decreasing and bounded above by 1, hence convergent, and it is eventually constant at the value the restart returns;* (iii) *the algorithm terminates after at most $\binom{d}{s}$ iterations; and* (iv) *at termination S^t is a local maximum of the support problem* (3) *with respect to the neighbourhood generated by the moves it uses—in particular no truncated power step, no trace step and no exchange move improves it. Algorithm 1 has no step size, no penalty weight, no smoothing parameter and no convergence tolerance, since the stopping rule is the exact test $R' \leq R$ (for the proof, see Appendix E).*

Monotonicity in (i) is secured by the exact acceptance test that every proposal of the MM surrogate must pass, and the upper bound in (ii) comes from the normalization of the criterion, so the objective sequence is convergent. What makes the convergence finite is (iii), which follows from the fact that $\widehat{R}$ depends on $\mathbf{V}$ only through its support, so the strict increase in (i) forbids revisiting a support. At no extra cost, the construction also yields a bound on how far from optimal the returned support can be.

Corollary 6 (A posteriori optimality gap)**.** *Let $\hat{\mathbf{V}}$ be the output of Algorithm 1 and* OPT *the optimal value of $(\mathcal{P})$. Then $\widehat{R}(\hat{\mathbf{V}}) \leq \mathrm{OPT} \leq 1$, so $1 - \widehat{R}(\hat{\mathbf{V}})$ is a computable upper bound on the optimality gap. Whenever $\widehat{R}(\hat{\mathbf{V}}) = 1$ the output is globally optimal and, by Proposition 1(ii), its column space is a top-k eigenspace of $\widehat{\mathbf{\Sigma}}$ (for the proof, see Appendix E).*

Normalizing by κ_j is what makes this possible, since the criterion then carries an absolute scale and the reported value *is* the bound. That value is also a floor under the retention of every component, which no aggregate of eigenvalues provides, since an aggregate can come close to its own bound while one component is lost.

3.4 STATISTICAL GUARANTEE UNDER A GENERAL COVARIANCE

Corollary 6 concerns the optimization problem; the next theorem concerns the estimator. It says that under Setting 1—no spiked assumption, no incoherence, no initialization condition—the global

maximizer of $(\mathcal{P})$ recovers the target eigenspace as soon as the covariance is accurate *on sparse principal submatrices only*, which is what buys the $s \log d$ rate.

Theorem 7 (Recovery from the sparse operator-norm error)**.** *Assume Setting 1 with $|S^\star| \le s$, write $\lambda_j = \lambda_j(\mathbf{\Sigma})$, and let $\varepsilon := \max_{|S|\le s} \|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_{\mathrm{op}}$ and $\eta := \max_{j\in[k]} |\kappa_j/\lambda_j - 1|$. Let $\hat{\mathbf{V}}$ be any global maximizer of $(\mathcal{P})$ and $\mathbf{V}$ the top-k eigenvectors of $\mathbf{\Sigma}$. If $\varepsilon < \lambda_k$ and $\eta < 1$, then for every $j \in [k]$,*

$$\lambda_j\Big(\hat{\mathbf{V}}^\top \mathbf{\Sigma} \hat{\mathbf{V}}\Big) \ge (1-\theta)\lambda_j - \varepsilon, \qquad \theta := 1 - \frac{(1-\varepsilon/\lambda_k)(1-\eta)}{1+\eta} \le \frac{\varepsilon}{\lambda_k} + 2\eta,$$

and consequently

$$\sin^2 \angle(\hat{\mathbf{V}}, \mathbf{V}) \;\le\; \frac{\theta\lambda_1 + \varepsilon}{\gamma\lambda_k} \;\le\; \frac{1}{\gamma}\left[\frac{\lambda_1}{\lambda_k}\left(\frac{\varepsilon}{\lambda_k} + 2\eta\right) + \frac{\varepsilon}{\lambda_k}\right]. \tag{7}$$

If the normalizers are taken to be the sparse ones $\kappa_j = \max_{|S|\le s} \lambda_j(\widehat{\mathbf{\Sigma}}_S)$, then $\eta \le \varepsilon/\lambda_k$ deterministically (for the proof, see Appendix E).

Under standard sub-Gaussian concentration the resulting sample size scales as $s \log d$, the dependence on s and d of the minimax rate for row-sparse subspaces [28]; the constant is stated in Corollary 11. Theorem 7 concerns the *estimator* defined by $(\mathcal{P})$, as the minimax analyses of row-sparse subspace estimation do [6]. Algorithm 1 provably ascends the objective that defines this estimator, and whenever the reported objective reaches the level that the proof guarantees for the optimum, (7) holds for the returned iterate as well (see Remark 10).

4 Experiments

Section 4.1 treats a single component, Section 4.2 the subspace problem under a limited feature budget, and Section 4.3 a real corpus; the optimization behaviour, the optimality bounds and the cost are studied in Appendix J. These are the covariance models and the text corpus that are standard in this literature.

Methods we compare against. We compare TRIM with the state-of-the-art and benchmark estimators, covering both the combinatorial and the convex family. **DiagThresh** (2009) [15] keeps the s coordinates of largest marginal variance, which identifies the support only when the background is isotropic. **FPS** (2013) [29] is the Fantope relaxation (in the sense of losing tightness), whose work grows with the ambient dimension whatever the budget and whose regularization parameter has to be chosen. **CovThresh** (2016) [8] thresholds the covariance first, but still selects on a statistic computed before any iteration. **Trace criterion** (2020) [25] is the $\ell_{2,0}$-constrained trace maximization, an aggregate objective, and therefore the one whose degeneracy Proposition 2 concerns. **Sequential RTPM** (2026) [17] extracts the components one at a time with the restarted truncated power method, a recursion that is not closed in the sparse setting, since the residual matrix it passes on need not have a sparse leading eigenvector; at $k = 1$ it is RTPM itself. **GreedyCorr** (2026) keeps, for each seed coordinate, the coordinates most correlated with it, committing to the leading direction early, and is among the heuristics that fail on the general covariances motivating the restarted method. For fairness, every method receives the same feature budget and the same sample covariance, and three choices are made on the competitors' behalf, giving the trace criterion TRIM's own search, RTPM the better of its two truncation levels and FPS its best regularization (for fuller descriptions and the protocol, see Appendices B and G).

4.1 A single component, where the criterion costs no accuracy

With $k = 1$ the criterion (2) *is* ordinary sparse PCA, so this experiment tests the optimizer alone. Across the three covariance models TRIM matches or exceeds RTPM from $n = 1000$ onward and recovers the planted direction on the counterexample where diagonal thresholding fails structurally, so moving to the worst-case criterion costs no accuracy at $k = 1$, as Lemma 4 anticipates, since at $k = 1$ the update of TRIM is itself a truncated power step (for the comparison, see Appendix I).

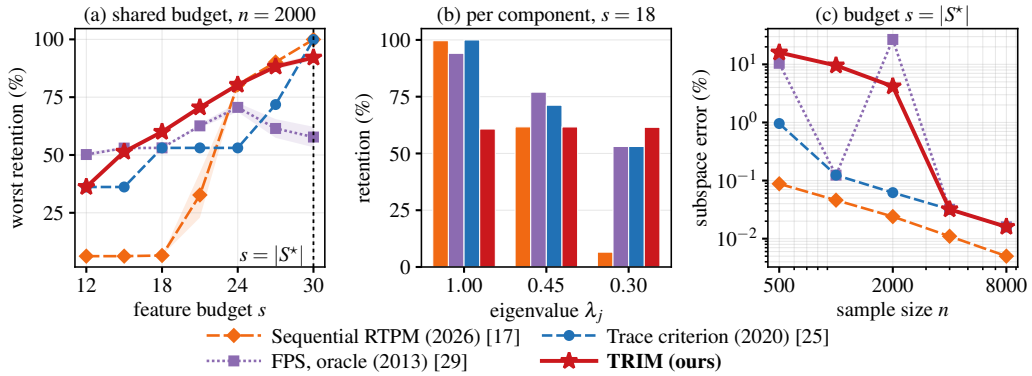

Figure 1: Sparse principal subspace estimation, $d = 300$, $k = 3$, $|S^\star| = 30$.

4.2 Subspace estimation under a limited feature budget

We use the block model of Appendix G.2, with $d = 300$, $k = 3$, planted eigenvalues $(1,\ 0.45,\ 0.30)$ above a floor of 0.02, and three disjoint blocks of 10 coordinates, so $|S^\star| = 30$. Two aspects of the protocol matter. The "Trace criterion" baseline runs TRIM's own search with the Ky Fan value as its objective, so the distance between those two curves is attributable to the objective alone. And FPS is given an oracle-selected regularization parameter, a concession no practitioner could make, so that it is not penalized for a poor default (for both, see Appendix G.3).

Panel (a) varies the shared budget at $n = 2000$. Sequential extraction retains 6.2–6.5% of the weakest component for every $s \le 18$, and the trace criterion stays at 53.1% from $s = 18$ to $s = 24$, while TRIM keeps a usable share of that component throughout, rising from 36.2% at $s = 12$ to 88.2% at $s = 27$. It leads both aggregate estimators over $15 \le s \le 21$, matches the trace criterion at the tightest budget $s = 12$, is level with the better of the two at $s = 24$, and is ahead of every method we compare against at $s = 18$ and $s = 21$. Although oracle-FPS, whose regularization is tuned on the population covariance, is ahead at the two tightest budgets, it is non-monotone in s and falls back to 57.8% at $s = |S^\star|$, where TRIM reaches 92.1%. Once the shared budget covers the true support the aggregate objectives recover, as expected, since both criteria are then maximized at $S^\star$ in the population (Section 2). Below that point, panel (a) is Proposition 2 and Remark 9 playing out at every intermediate budget, not only at $s = k$.

Panel (b) exposes the mechanism at $s = 18$, that is 60% of $|S^\star|$. The trace criterion retains $(100.0,\ 71.2,\ 53.1)\%$ of the three components, sequential extraction $(99.7,\ 61.7,\ 6.5)\%$ and oracle-FPS $(94.1,\ 77.0,\ 53.1)\%$, whereas TRIM retains $(60.8,\ 61.7,\ 61.5)\%$, the flat profile the criterion asks for. The allocations explain the difference, since the trace criterion draws $(10,\ 7,\ 0)$ coordinates from the three blocks and the sequential recursion $(10,\ 6,\ 0)$, while TRIM draws $(6,\ 6,\ 5.4)$, essentially the even split Remark 9 identifies as optimal. The aggregate methods spend the budget where their objective rewards it, on the first two components, and take no coordinate of the third, so the variance their third direction carries is background.

Panel (c) fixes $s = |S^\star|$, the regime covered by Theorem 7, and varies n. At the smallest sample sizes the smallest retained eigenvalue is the hardest spectral quantity to estimate, and TRIM is on a par with the trace criterion once $n \ge 4000$, as the theorem leads one to expect. Figure 1 thus shows the worst-case criterion paying off wherever a shared budget must choose among components.

4.3 20 Newsgroups

We take the 20 Newsgroups corpus with headers, footers and quotes removed, form sublinear TF–IDF features on the $d = 1000$ most frequent non-stopword terms, and split the 18,846 documents in half. Every method is fitted on the first half and *evaluated on the second*, so no method is scored on the sample it optimized. We extract $k = 4$ components under a shared budget $s \in \{20, 30, 45, 60\}$.

Table 1: Held-out worst-component retention on 20 Newsgroups, $k = 4$, in percent.

	worst-component retention (held out)				total EVR
method	$s = 20$	$s = 30$	$s = 45$	$s = 60$	$s = 60$
DiagThresh (2009) [15]	52.6	55.4	58.8	61.0	1.96
FPS (2013) [29]	44.1	44.2	44.2	44.1	1.58
CovThresh (2016) [8]	23.5	23.6	25.3	26.5	0.95
Trace criterion (2020) [25]	<u>55.2</u>	<u>57.8</u>	<u>59.0</u>	<u>62.6</u>	2.03
Sequential RTPM (2026) [17]	51.8	54.4	56.4	58.8	1.91
TRIM (ours)	**55.8**	**59.2**	**59.8**	**63.9**	2.00

Table 1 reports, per budget, the worst componentwise retention on the held-out half (best in bold, second best underlined) and the total explained variance ratio (EVR) at $s = 60$. TRIM is best at every budget, the tightest included, which is the regime Proposition 2 concerns, and since the trace criterion runs the same search, that lead belongs to the objective. Its total explained variance matches the trace criterion's, so the gain is a redistribution toward the component that the trace criterion serves worst. The supports point the same way, since at $s = 60$ each of TRIM's four components is led by a different register, where covariance thresholding spends its support on numerals that co-occur strongly but carry no topical signal (for the recovered terms and their reading, see Appendix H).

5 Limitations

Our problem is non-convex, where no efficient algorithm can guarantee a global maximizer [21]; we establish monotone ascent, a convergent objective and finite termination.

6 Conclusion

We have proposed the worst componentwise retention ratio as the objective for sparse principal subspace estimation, and TRIM as the method that attacks it. The criterion is attained exactly by a choice of support, it carries an absolute scale, and it cannot neglect a component, which an aggregate of captured variance can do whenever k components share a limited feature budget, and which Proposition 2 shows to be a structural failure, not a matter of degree. The algorithm is monotone, finitely terminating, free of tuning and of relaxations (in the sense of losing tightness), and it never projects out what it has already found. Under a limited feature budget it keeps far more of the weakest component than the aggregate objectives it replaces, and on a real corpus it outperforms every estimator we compare against at every sparsity level.

AI use statement

AI-assisted editing was used for language editing and proofreading of the main paper and supplementary materials. The authors take full responsibility for the final content of this paper.

Reproducibility statement

Every number we report is recorded in the supplementary archive that accompanies this submission, which also holds our code and the data. The archive holds our implementation of Algorithm 1, the covariance models of Appendix G.2, the 20 Newsgroups corpus of Section 4.3 with the preprocessing applied to it, the recorded measurements behind every figure and every table, and the scripts that regenerate them, so the whole of Section 4 can be rebuilt from it. Every experiment fixes its seeds explicitly, the synthetic recovery results are averaged over 20 replications for $k = 1$ and 10 for $k > 1$, and shaded bands are one standard error. Appendix G.2 states each covariance model exactly, and Appendix G.3 states the protocol under which every competing estimator was run. Every formal claim is proved in Appendix E, and Appendix G.4 sets out thirteen numerical tests, included in the archive, that check the central ones independently of the proofs. TRIM has no step size, no penalty weight, no smoothing parameter and no tolerance to set, and every experiment reported here was run

on an AMD Ryzen 5 5600H, a six-core processor with a base clock of 3.30 GHz, with no GPU at any point, every timing being measured single-threaded on one core.

REFERENCES

[1] Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):2877–2921, 2009.

[2] Konstantinos Benidis, Ying Sun, Prabhu Babu, and Daniel P. Palomar. Orthogonal sparse PCA and covariance estimation via Procrustes reformulation. *IEEE Transactions on Signal Processing*, 64(23):6211–6226, 2016.

[3] Wei Bian and Dacheng Tao. Max-min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5): 1037–1050, 2011.

[4] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[5] Jian-Feng Cai, Zhuozhi Xian, and Jiaxi Ying. Fast and provable algorithms for sparse PCA with improved sample complexity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 6319–6340, 2025.

[6] T. Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[7] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3): 434–448, 2007.

[8] Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. *Journal of Machine Learning Research*, 17(141):1–41, 2016.

[9] Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

[10] Yuqing He, Guanyi Wang, and Yu Yang. Sparse principal component analysis with non-oblivious adversarial perturbations. *arXiv preprint arXiv:2411.05332*, 2024.

[11] Alfred Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3):620–630, 1954.

[12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

[13] David R. Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[14] Arun Jambulapati, Syamantak Kumar, Jerry Li, Shourya Pandey, Ankit Pensia, and Kevin Tian. Black-box k-to-1-PCA reductions: Theory and applications. In *Conference on Learning Theory*, volume 247 of *PMLR*, pp. 2564–2607, 2024.

[15] Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[16] Ian T. Jolliffe, Nickolay T. Trendafilov, and Mudassir Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3): 531–547, 2003.

[17] Syamantak Kumar, Purnamrita Sarkar, Kevin Tian, and Peiyuan Zhang. Combinatorial sparse PCA beyond the spiked identity model. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *PMLR*, 2026.

[18] Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

[19] Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[20] Lester W. Mackey. Deflation methods for sparse PCA. In *Advances in Neural Information Processing Systems*, 2008.

[21] Malik Magdon-Ismail. NP-hardness and inapproximability of sparse PCA. *Information Processing Letters*, 126:35–38, 2017.

[22] Gleb Novikov. Sparse PCA beyond covariance thresholding. *arXiv preprint arXiv:2302.10158*, 2023.

[23] Ying Sun, Prabhu Babu, and Daniel P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2017.

[24] Kean Ming Tan, Zhaoran Wang, Han Liu, and Tong Zhang. Sparse generalized eigenvalue problem: optimal statistical rates via truncated Rayleigh flow. *Journal of the Royal Statistical Society: Series B*, 80(5):1057–1086, 2018.

[25] Lai Tian, Feiping Nie, Rong Wang, and Xuelong Li. Learning feature sparse principal subspace. In *Advances in Neural Information Processing Systems*, 2020.

[26] David Vävinggren, Francis Bach, André M. H. Teixeira, Dave Zachariah, and Antônio H. Ribeiro. A robust optimization approach to sparse principal component analysis. *arXiv preprint arXiv:2606.03553*, 2026.

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

[28] Vincent Q. Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[29] Vincent Q. Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems*, 2013.

[30] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(28):899–925, 2013.

[31] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

Supplementary Material

A Related work

We summarize the threads this paper touches.

Combinatorial estimators and the spiked identity model. The lightest sparse PCA estimators pick a support by a one-shot statistic—diagonal thresholding [1, 15] or covariance thresholding [4, 8, 22]—or refine it iteratively, as in iterative thresholding [19], the truncated power method [30], truncated Rayleigh flow [24] and its recent refinements [5]. Every one of these analyses buys its guarantee with an assumption the data need not satisfy, such as a spiked identity covariance or a local initialization, and that hypothesis cannot simply be dropped, since there are general covariances against which the one-shot estimators fail outright [17]. The same work supplies the restarted truncated power method (RTPM) we compare against throughout.

Subspaces, sequential extraction and row sparsity. Row and column subspace sparsity are distinguished in [28], and row sparsity—one shared support for all k components—is what makes the output a genuine feature selection, and is the notion we adopt. The $\ell_{2,0}$-constrained *trace* problem is studied in [25]; that objective is an aggregate, and its degeneracy is what Proposition 2 analyzes. The alternative to joint optimization is to extract the components one at a time [14, 20], and in the sparse setting that recursion is not closed, since the residual matrix it passes on need not have a sparse leading eigenvector, which is our reason for avoiding it.

Robust formulations of sparse PCA. Sparse PCA under non-oblivious adversarial perturbations [10] maximizes variance against an adversary who corrupts the samples within a norm ball, and AdvPCA [26] induces sparsity by optimizing a reconstruction objective against bounded worst-case corruptions of the latent space. In both, the inner minimization runs over *perturbations of the data*, and the quantity being maximized remains an aggregate of explained variance; shrink the uncertainty set to a point and each returns to an ordinary aggregate objective. The inner minimization in (2) runs over no adversary and no perturbation. It runs over the k *components themselves*, and asks which of them is served worst by the chosen support. The two constructions therefore answer different questions, and only the second is sensitive to the degeneracy of Proposition 2, since a robust criterion built on an aggregate inherits that degeneracy along with the aggregate, however large its uncertainty set is made.

B The compared methods

This appendix describes what each of the methods we compare against does. The description is of the estimator itself; how each one was run is stated in Appendix G.3.

DiagThresh (2009) [15]. Diagonal thresholding selects the support by a single pass over the diagonal of the sample covariance, keeping the s coordinates of largest marginal variance $\widehat{\mathbf{\Sigma}}_{ii}$, and then returns the top-k eigenvectors of the resulting $s \times s$ principal submatrix. The reasoning behind it holds only under a spiked identity covariance, where the coordinates carrying the spike are exactly those whose variance exceeds the isotropic background. Under a general covariance that reasoning fails in both directions, since a coordinate can have large marginal variance without belonging to the leading eigenvector, and a coordinate in the support can be masked by background variance elsewhere. Its lightness is bought at the price of a model assumption the data need not satisfy.

FPS (2013) [29]. The Fantope relaxation (in the sense of losing tightness) replaces the rank-k projection matrix by a matrix variable constrained to the Fantope, the convex hull of rank-k projections, and adds an ℓ_1 penalty to induce sparsity. The resulting problem is convex and is solved by ADMM; the solution is then converted back into a subspace by row truncation followed by a top-k eigendecomposition on the retained support. Two costs come with that convexity. The decision variable is $d \times d$, so the work grows with the ambient dimension whatever the budget, and the regularization parameter has to be chosen, with the sparsity level reached only indirectly through it.

Convexity removes the difficulty at the price of exactness, since the optimum of the convex program need not be an optimum of the problem it replaces.

CovThresh (2016) [8]. Covariance thresholding sets to zero the entries of the sample covariance that fall below a threshold, takes the leading eigenvectors of the thresholded matrix, keeps the s coordinates on which they weigh most, and then takes the top-k eigenvectors of the *unthresholded* submatrix on that support. Thresholding suppresses the small spurious correlations that dominate the covariance when d is large, which widens the range of sparsity levels a one-shot rule survives, but it does not change what kind of rule it is. Selection still rests on a single statistic computed before any iteration, so the estimator inherits the structural limitation of the one-shot family, and it is defeated by a background whose entries survive thresholding without carrying topical signal.

Trace criterion (2020) [25]. This is the $\ell_{2,0}$-constrained trace maximization, that is the maximization of $\mathrm{tr}(\mathbf{V}^\top \widehat{\mathbf{\Sigma}}\mathbf{V})$ over row-sparse orthonormal $\mathbf{V}$. Its objective is an aggregate, the sum of the k retained eigenvalues, and a sum is maximized by spending a shared budget wherever it buys the most variance, which under a limited budget means the leading direction, so the weakest components are sacrificed. This is the property Proposition 2 concerns. Optimizing an aggregate jointly over all k components still neglects what the aggregate neglects, even though it spares the method the difficulty a sequential recursion runs into.

RTPM and Sequential RTPM (2026) [17]. The restarted truncated power method runs truncated power iterations from every coordinate basis vector, with a truncation level that may exceed the budget, and returns the iterate of largest $\mathbf{v}^\top \widehat{\mathbf{\Sigma}}\mathbf{v}$, truncated to the budget. Restarting from every coordinate removes the dependence on a good initialization, and the method provably succeeds under a general covariance, which the earlier combinatorial estimators do not, but the guarantee is stated for a single component only. To obtain k components we use it inside the standard recursion that extracts them one at a time, applying it to successive residuals $\mathbf{P}\widehat{\mathbf{\Sigma}}\mathbf{P}$ within the shared budget. That recursion is not closed in the sparse setting, since the residual matrix need not have a sparse leading eigenvector, so the guarantee does not survive the step from one component to several. For $k = 1$ the two coincide.

GreedyCorr (2026) [17]. A planted-clique-style greedy heuristic, seeded from every coordinate, which keeps for each seed the s coordinates whose rows of $\widehat{\mathbf{\Sigma}}$ are most correlated with the row of the seed, and returns the seed whose support captures the most variance. Selecting by correlation with a single seed commits to the leading direction early and never revises that commitment, and the heuristic is among the estimators shown to fail on the general covariances that motivate the restarted truncated power method. We include it as the natural greedy alternative to the thresholding rules.

C A PRELIMINARY ON THE MINORIZATION–MAXIMIZATION (MM) PRINCIPLE

This appendix reviews the minorization–maximization (MM) principle and its extension to max–min objectives, which is the form our criterion takes. MM is a general principle of optimization, not a particular algorithm, and what follows fixes the notation in which we use it.

C.1 THE GENERAL MM FRAMEWORK

MM addresses constrained maximization problems of the form

$$\max_{\mathbf{X}\in\mathcal{X}} \; f(\mathbf{X}), \tag{S1}$$

where f is the objective, $\mathbf{X}$ the optimization variable and $\mathcal{X}$ the feasible set. Instead of attacking (S1) directly, MM solves a sequence of simpler problems. At iteration t, given the current iterate $\mathbf{X}^t$, one constructs a surrogate $g(\mathbf{X} \mid \mathbf{X}^t)$, called a *minorizer* of f, which is required to satisfy

$$g(\mathbf{X} \mid \mathbf{X}^t) \;\le\; f(\mathbf{X}), \qquad \forall\, \mathbf{X} \in \mathcal{X}, \tag{S2}$$

$$g(\mathbf{X}^t \mid \mathbf{X}^t) \;=\; f(\mathbf{X}^t). \tag{S3}$$

Condition (S2) makes the surrogate a global lower bound on the objective, and condition (S3) makes it touch the objective at the current iterate. The next iterate is obtained by maximizing the surrogate over the feasible set,

$$\mathbf{X}^{t+1} \in \arg\max_{\mathbf{X}\in\mathcal{X}} g\big(\mathbf{X} \mid \mathbf{X}^t\big), \tag{S4}$$

and the two steps are repeated until the stopping rule is met. Figure S1 illustrates the behaviour of the scheme.

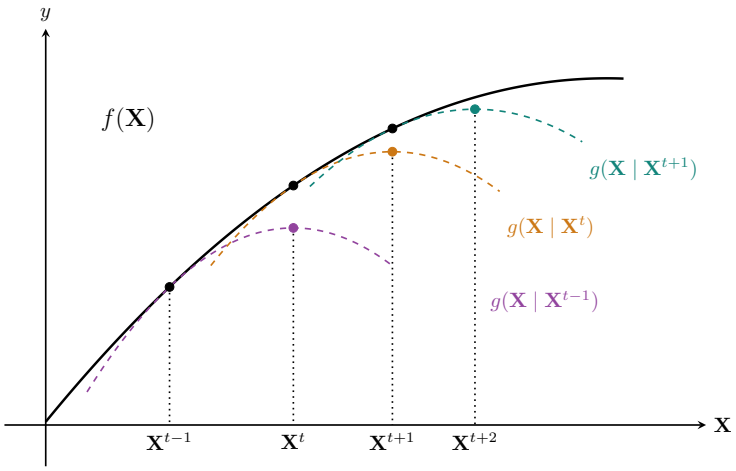

Figure S1: Behaviour of the MM scheme.

What this construction buys is monotone ascent, with no step size, no penalty weight and no line search. Writing the relations in order,

$$f(\mathbf{X}^{t+1}) \;\geq\; g(\mathbf{X}^{t+1} \mid \mathbf{X}^t) \;\geq\; g(\mathbf{X}^t \mid \mathbf{X}^t) \;=\; f(\mathbf{X}^t), \tag{S5}$$

where the first inequality is the lower-bound property (S2), the second is the maximization step (S4), and the equality is the tangency condition (S3). The sequence $\{f(\mathbf{X}^t)\}$ is therefore non-decreasing. The minorizer is *tight* at $\mathbf{X}^t$ by construction, and that tightness is what distinguishes MM from a relaxation (in the sense of losing tightness), namely a surrogate whose optimal value need not agree with that of the problem it replaces.

C.2 An MM framework for max–min problems

The principle extends to max–min objectives of the form

$$\max_{\mathbf{X}\in\mathcal{X}} \Big\{ f(\mathbf{X}) := \min_{i\in[K]} f_i(\mathbf{X}) \Big\}, \tag{S6}$$

in which the objective is the pointwise minimum of a family $\{f_i\}_{i\in[K]}$. This is the shape of the problem studied in this paper, with the f_i the normalized eigenvalue functions and $K = k$.

Let $g_i(\mathbf{X} \mid \mathbf{X}^t)$ be a minorizer of each individual f_i, so that

$$g_i(\mathbf{X} \mid \mathbf{X}^t) \;\leq\; f_i(\mathbf{X}), \qquad \forall\, \mathbf{X} \in \mathcal{X}, \tag{S7}$$

$$g_i(\mathbf{X}^t \mid \mathbf{X}^t) \;=\; f_i(\mathbf{X}^t). \tag{S8}$$

A surrogate for the max–min objective is then obtained as the pointwise minimum of the individual surrogates,

$$g\big(\mathbf{X} \mid \mathbf{X}^t\big) := \min_{i\in[K]} g_i\big(\mathbf{X} \mid \mathbf{X}^t\big). \tag{S9}$$

This is a valid minorizer of f. The lower-bound property follows from

$$g(\mathbf{X} \mid \mathbf{X}^t) = \min_i g_i(\mathbf{X} \mid \mathbf{X}^t) \;\leq\; \min_i f_i(\mathbf{X}) = f(\mathbf{X}), \tag{S10}$$

and the tangency condition from

$$g(\mathbf{X}^t \mid \mathbf{X}^t) = \min_i g_i(\mathbf{X}^t \mid \mathbf{X}^t) = \min_i f_i(\mathbf{X}^t) = f(\mathbf{X}^t). \tag{S11}$$

Applying the update (S4) with the surrogate (S9) therefore inherits the ascent property (S5) for the max–min objective.

Two features of this construction matter for what follows. The composite surrogate is again tight, so no relaxation (in the sense of losing tightness) is introduced at any stage. And the inner minimum over i survives into the surrogate, so the subproblem (S4) remains a max–min problem in its own right, and the way it is handled is where the substance of the method lies.

A fuller account of MM, and in particular of the standard techniques for constructing minorizers, is given in [13, 18, 23].

D THE TRIM ALGORITHM IN PSEUDOCODE

Algorithm 1 collects the iteration described in Section 3. The candidate supports (a)–(d) are the moves listed there, the restarts and the screening rule are those of Appendix F, and every candidate is scored by the exact objective $\min_j \lambda_j(\widehat{\mathbf{\Sigma}}_S)/\kappa_j$, which is legitimate by Proposition 1(i).

Algorithm 1 TRIM (Truncated Rayleigh Iterations via Minorization)

Require: $\widehat{\mathbf{\Sigma}} \in \mathbb{S}^{d\times d}_{\succcurlyeq 0}$, budget s, components k, cap T

$\kappa_j \leftarrow \lambda_j(\widehat{\mathbf{\Sigma}}),\ j \in [k]$ ▷ normalizers, once
$\mathcal{S}_0 \leftarrow \{\mathrm{top}_s((\widehat{\mathbf{\Sigma}}_{i:})^2) : i \in [d]\} \cup \{S_{\mathrm{casc}}\}$ ▷ restarts
for each $S^0 \in \mathcal{S}_0$ **do**
  $S \leftarrow S^0;\ \mathbf{V} \leftarrow$ top-k eigenvectors of $\widehat{\mathbf{\Sigma}}_S$
  $R \leftarrow \min_j \lambda_j(\widehat{\mathbf{\Sigma}}_S)/\kappa_j$
  for $t = 1, \dots, T$ **do**
    $\mathbf{W} \leftarrow$ principal directions of $\mathbf{V};\ \mathbf{G} \leftarrow \widehat{\mathbf{\Sigma}}\mathbf{W}$ ▷ $O(dsk)$
    $\mathcal{C} \leftarrow$ candidate supports (a)–(d) from $\mathbf{G}$ and S
    $S' \leftarrow \arg\max_{C\in\mathcal{C}}\ \min_j \lambda_j(\widehat{\mathbf{\Sigma}}_C)/\kappa_j$ ▷ exact scoring
    $R' \leftarrow \min_j \lambda_j(\widehat{\mathbf{\Sigma}}_{S'})/\kappa_j$
    if $R' \le R$ **then break** ▷ local optimum reached
    end if
    $S \leftarrow S';\ R \leftarrow R';\ \mathbf{V} \leftarrow$ top-k eigenvectors of $\widehat{\mathbf{\Sigma}}_S$
  end for
end for
return the pair $(\mathbf{V}, S)$ with the largest R over all restarts

E PROOFS

Throughout, $\widehat{\mathbf{\Sigma}} \in \mathbb{S}^{d\times d}_{\succcurlyeq 0}$, $k \le s \le d$, and $\mathcal{V}^k_s = \{\mathbf{V} \in \mathbb{R}^{d\times k} : \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k,\ \|\mathbf{V}\|_{2,0} \le s\}$. We use the Cauchy interlacing theorem and Weyl's inequality [12], and Ky Fan's variational principle [9].

E.1 PROOF OF PROPOSITION 1

(i). Fix $S \subseteq [d]$ with $|S| \le s$ and let $\mathbf{V} \in \mathcal{V}^k_s$ satisfy $\mathrm{rsupp}(\mathbf{V}) \subseteq S$. Deleting the zero rows gives $\widetilde{\mathbf{V}} \in \mathbb{R}^{|S|\times k}$ with $\widetilde{\mathbf{V}}^\top\widetilde{\mathbf{V}} = \mathbf{I}_k$ and $\mathbf{V}^\top\widehat{\mathbf{\Sigma}}\mathbf{V} = \widetilde{\mathbf{V}}^\top\widehat{\mathbf{\Sigma}}_S\widetilde{\mathbf{V}}$. The map $\widetilde{\mathbf{V}} \mapsto \widetilde{\mathbf{V}}^\top\widehat{\mathbf{\Sigma}}_S\widetilde{\mathbf{V}}$ is a compression of $\widehat{\mathbf{\Sigma}}_S$ to a k-dimensional subspace, so Cauchy interlacing gives

$$\lambda_j\Big(\widetilde{\mathbf{V}}^\top\widehat{\mathbf{\Sigma}}_S\widetilde{\mathbf{V}}\Big) \le \lambda_j(\widehat{\mathbf{\Sigma}}_S) \qquad \text{for every } j \in [k]. \tag{S12}$$

Taking $\widetilde{\mathbf{V}}^\star$ to be the top-k eigenvectors of $\widehat{\mathbf{\Sigma}}_S$ yields $\widetilde{\mathbf{V}}^{\star\top}\widehat{\mathbf{\Sigma}}_S\widetilde{\mathbf{V}}^\star = \mathrm{diag}(\lambda_1(\widehat{\mathbf{\Sigma}}_S), \dots, \lambda_k(\widehat{\mathbf{\Sigma}}_S))$, so (S12) holds with equality simultaneously for all $j \in [k]$. Hence $\widetilde{\mathbf{V}}^\star$ maximizes every $\lambda_j(\cdot)$ at once, and in particular maximizes any non-decreasing function of the vector $(\lambda_1, \dots, \lambda_k)$, including $\min_j \lambda_j(\cdot)/\kappa_j$. Optimizing over S then gives (3). ■

(ii). Applying (S12) with $S = [d]$ gives $\lambda_j(\mathbf{V}^\top\widehat{\mathbf{\Sigma}}\mathbf{V}) \le \lambda_j(\widehat{\mathbf{\Sigma}}) = \kappa_j$ for all $j \in [k]$, hence $\widehat{R}(\mathbf{V}) \le 1$. Suppose $\widehat{R}(\mathbf{V}) = 1$. Then $\lambda_j(\mathbf{V}^\top\widehat{\mathbf{\Sigma}}\mathbf{V}) = \kappa_j$ for every $j \in [k]$ (each ratio is at

most 1 and the minimum equals 1), so $\mathrm{tr}(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}) = \sum_{j \le k} \lambda_j(\widehat{\mathbf{\Sigma}})$. By Ky Fan's principle [9], $\max\{\mathrm{tr}(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}) : \mathbf{V}^\top \mathbf{V} = \mathbf{I}_k\} = \sum_{j \le k} \lambda_j(\widehat{\mathbf{\Sigma}})$ with the maximum attained exactly at those $\mathbf{V}$ whose column space is a top-k eigenspace of $\widehat{\mathbf{\Sigma}}$; hence the column space of $\mathbf{V}$ is such an eigenspace. Conversely, if it is, then $\lambda_j(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}) = \lambda_j(\widehat{\mathbf{\Sigma}})$ for all $j \in [k]$ and $\widehat{R}(\mathbf{V}) = 1$. ■

(iii). Let $\mathbf{V}$ satisfy $\mathbf{V}^\top \mathbf{V} \preccurlyeq \mathbf{I}_k$ and $\|\mathbf{V}\|_{2,0} \le s$. Write a polar-type factorization $\mathbf{V} = \mathbf{Q}\mathbf{C}$ where $\mathbf{Q} \in \mathbb{R}^{d \times k}$ has orthonormal columns with $\mathrm{rsupp}(\mathbf{Q}) \subseteq \mathrm{rsupp}(\mathbf{V})$ and $\mathbf{C} \in \mathbb{R}^{k \times k}$ satisfies $\mathbf{C}^\top \mathbf{C} = \mathbf{V}^\top \mathbf{V} \preccurlyeq \mathbf{I}_k$, so $\|\mathbf{C}\|_{\mathrm{op}} \le 1$ (such a factorization exists, since we may take $\mathbf{Q}$ to be an orthonormal basis of a k-dimensional subspace of the coordinate subspace $\mathbb{R}^{\mathrm{rsupp}(\mathbf{V})}$ containing the range of $\mathbf{V}$, which is possible since $|\,\mathrm{rsupp}(\mathbf{V})| \le s$ and $k \le s$). Then with $\mathbf{N} := \mathbf{Q}^\top \widehat{\mathbf{\Sigma}} \mathbf{Q} \succcurlyeq 0$ we have $\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V} = \mathbf{C}^\top \mathbf{N} \mathbf{C} = \mathbf{B}^\top \mathbf{B}$ for $\mathbf{B} := \mathbf{N}^{1/2} \mathbf{C}$, so for every $j \in [k]$,

$$\lambda_j\Big(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}\Big) = \sigma_j(\mathbf{B})^2 \le \big(\sigma_j(\mathbf{N}^{1/2})\, \|\mathbf{C}\|_{\mathrm{op}}\big)^2 \le \lambda_j(\mathbf{N}) = \lambda_j\Big(\mathbf{Q}^\top \widehat{\mathbf{\Sigma}} \mathbf{Q}\Big),$$

where the first inequality is the standard bound $\sigma_j(\mathbf{A}_1 \mathbf{A}_2) \le \sigma_j(\mathbf{A}_1) \|\mathbf{A}_2\|_{\mathrm{op}}$. Since $\mathbf{Q} \in \mathcal{V}_s^k$, every point $\mathbf{V}$ of the inequality-constrained set is dominated componentwise by a feasible point $\mathbf{Q}$, and the optimal values coincide. ■

E.2 Proof of Proposition 8

Proposition 8 (The column-wise worst case reduces to the trace criterion)**.** *For every $\widehat{\mathbf{\Sigma}} \in \mathbb{S}_{\succcurlyeq 0}^{d \times d}$ and $k \le s \le d$,*

$$\max_{\mathbf{V} \in \mathcal{V}_s^k} \min_{j \in [k]} \mathbf{v}_j^\top \widehat{\mathbf{\Sigma}} \mathbf{v}_j \;=\; \frac{1}{k} \max_{|S| \le s} \sum_{i \le k} \lambda_i(\widehat{\mathbf{\Sigma}}_S) \;=\; \frac{1}{k} \max_{\mathbf{V} \in \mathcal{V}_s^k} \mathrm{tr}\Big(\mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}\Big).$$

Fix S with $|S| \le s$ and let $\mathbf{V}$ range over matrices in $\mathcal{V}_s^k$ with $\mathrm{rsupp}(\mathbf{V}) \subseteq S$. For any such $\mathbf{V}$ and any orthogonal $\mathbf{O} \in \mathbb{R}^{k \times k}$, the matrix $\mathbf{V}\mathbf{O}$ is again feasible with the same column space, and $(\mathbf{V}\mathbf{O})^\top \widehat{\mathbf{\Sigma}} (\mathbf{V}\mathbf{O}) = \mathbf{O}^\top \mathbf{M} \mathbf{O}$ with $\mathbf{M} = \mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}$. Write $\mathrm{diag}(\mathbf{N})$ for the vector of diagonal entries of $\mathbf{N}$ and $\prec$ for majorization. By the Schur–Horn theorem [11], as $\mathbf{O}$ ranges over the orthogonal group the attainable diagonals $\mathrm{diag}(\mathbf{O}^\top \mathbf{M} \mathbf{O})$ are exactly the vectors majorized by the eigenvalue vector $\lambda(\mathbf{M})$. The constant vector $(\mathrm{tr}\,\mathbf{M}/k, \ldots, \mathrm{tr}\,\mathbf{M}/k)$ is majorized by every such $\lambda(\mathbf{M})$, hence attainable, and no attainable diagonal has minimum entry larger than the mean. Therefore

$$\max_{\mathbf{O}} \min_{j \in [k]} \big(\mathbf{O}^\top \mathbf{M} \mathbf{O}\big)_{jj} = \frac{\mathrm{tr}\,\mathbf{M}}{k}.$$

Maximizing over $\mathbf{V}$ supported on S and using Ky Fan's principle on $\widehat{\mathbf{\Sigma}}_S$ gives $\max \mathrm{tr}\,\mathbf{M} = \sum_{i \le k} \lambda_i(\widehat{\mathbf{\Sigma}}_S)$, and maximizing over S completes the proof. ■

E.3 Proof of Proposition 2

With $\mathbf{v}_j = (\mathbf{e}_{2j-1} + \mathbf{e}_{2j})/\sqrt{2}$ the vectors $\{\mathbf{v}_j\}_{j \in [k]}$ are orthonormal with disjoint supports $P_j := \{2j-1, 2j\}$, so $\mathbf{\Sigma} = \sum_j \lambda_j \mathbf{v}_j \mathbf{v}_j^\top$ has eigenvalues $\lambda_1 > \cdots > \lambda_k > 0$ and $\lambda_{k+1}(\mathbf{\Sigma}) = 0$. Let $S \subseteq [d]$ with $|S| = s = k$. Since $\mathbf{\Sigma}_{ab} = 0$ whenever a, b lie in different pairs, $\mathbf{\Sigma}_S$ is block diagonal with respect to the partition of S induced by $\{P_j\}$. Let $A := \{j : P_j \subseteq S\}$ and $B := \{j : |S \cap P_j| = 1\}$, with $a = |A|$, $b = |B|$, so $2a + b = k$ (coordinates outside $\bigcup_j P_j$ contribute zero blocks and may be assumed absent). The block for $j \in A$ is $\frac{\lambda_j}{2} \mathbf{1}\mathbf{1}^\top$, with eigenvalues λ_j and 0; the block for $j \in B$ is the scalar $\lambda_j/2$. Hence the multiset of eigenvalues of $\mathbf{\Sigma}_S$ is

$$\{\lambda_j : j \in A\} \;\cup\; \{0 \text{ with multiplicity } a\} \;\cup\; \{\lambda_j/2 : j \in B\}, \tag{S13}$$

of total cardinality $2a + b = k$.

(i). By (S13) the trace criterion is $F(S) = \sum_{i \le k} \lambda_i(\mathbf{\Sigma}_S) = \sum_{j \in A} \lambda_j + \frac{1}{2} \sum_{j \in B} \lambda_j$. Consider the two configurations S_1 with $A = \{1\}$ and $B = \{2, \dots, k-1\}$, and S_0 with $A = \emptyset$ and $B = [k]$. Both are feasible ($2 \cdot 1 + (k-2) = k$ and $b = k$), and

$$F(S_1) - F(S_0) = \lambda_1 + \tfrac{1}{2}\sum_{j=2}^{k-1} \lambda_j - \tfrac{1}{2} \sum_{j=1}^{k} \lambda_j = \tfrac{1}{2}(\lambda_1 - \lambda_k) > 0,$$

because $\lambda_1 > \lambda_k$. Hence no maximizer of F has $a = 0$, i.e. every maximizer has $a \ge 1$. By (S13), $a \ge 1$ forces 0 into the spectrum of $\mathbf{\Sigma}_S$, and since that spectrum has exactly k elements we get $\lambda_k(\mathbf{\Sigma}_S) = 0$ and therefore $R = \min_j \lambda_j(\mathbf{\Sigma}_S)/\lambda_j = 0$.

(ii). If $a \ge 1$ then $R = 0$ by the previous paragraph. If $a = 0$ then $b = k$, so $B = [k]$ and by (S13) the eigenvalues of $\mathbf{\Sigma}_S$ are $\lambda_j/2$, $j \in [k]$, already in decreasing order. Thus $\lambda_j(\mathbf{\Sigma}_S)/\lambda_j = \frac{1}{2}$ for every j and $R = \frac{1}{2} > 0$. Hence the maximizers of R are exactly the supports with $a = 0$, and they attain $R = \frac{1}{2}$. ■

Remark 9 (Intermediate budgets). The same construction with blocks of size $r \ge 2$, that is $\mathbf{v}_j = r^{-1/2} \sum_{i \in P_j} \mathbf{e}_i$ with $|P_j| = r$ disjoint, shows the same phenomenon at every budget $s < kr$. A support that takes c_j coordinates from P_j contributes eigenvalues $\lambda_j c_j / r$ and $c_j - 1$ zeros, so the trace criterion is $\sum_j \lambda_j c_j / r$, which is maximized by concentrating the budget on the largest λ_j, while R, which equals $\min_j c_j / r$ whenever the values $\lambda_j c_j / r$ keep the order of the λ_j, is maximized by spreading it as evenly as possible. Figure 1(a) is the empirical counterpart with $r = 10$ and $k = 3$.

E.4 Proof of Lemma 3

Write $\mathbf{M} = \mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}$ and $\mathbf{M}^t = (\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V}^t$ with the eigendecomposition and notation of the statement. By the Courant–Fischer theorem applied to $\mathbf{M}$ and the particular j-dimensional subspace $\mathrm{range}(\mathbf{U}_j)$,

$$\lambda_j(\mathbf{M}) \;=\; \max_{\substack{\mathbf{U} \in \mathbb{R}^{k \times j} \\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}_j}} \lambda_{\min}\big(\mathbf{U}^\top \mathbf{M} \mathbf{U}\big) \;\ge\; \lambda_{\min}\big(\mathbf{U}_j^\top \mathbf{M} \mathbf{U}_j\big) \;=\; \min_{\|\mathbf{y}\|_2 = 1} \big\| \widehat{\mathbf{\Sigma}}^{1/2} \mathbf{V} \mathbf{U}_j \mathbf{y} \big\|_2^2. \tag{S14}$$

For each fixed unit $\mathbf{y}$, the map $\mathbf{V} \mapsto \|\widehat{\mathbf{\Sigma}}^{1/2} \mathbf{V} \mathbf{U}_j \mathbf{y}\|_2^2$ is a convex quadratic (a composition of a linear map with a squared Euclidean norm), so it dominates its first-order Taylor expansion at $\mathbf{V}^t$, giving

$$\begin{aligned} \big\| \widehat{\mathbf{\Sigma}}^{1/2} \mathbf{V} \mathbf{U}_j \mathbf{y} \big\|_2^2 \;&\ge\; \big\| \widehat{\mathbf{\Sigma}}^{1/2} \mathbf{V}^t \mathbf{U}_j \mathbf{y} \big\|_2^2 + 2 \big\langle \widehat{\mathbf{\Sigma}} \mathbf{V}^t \mathbf{U}_j \mathbf{y} \mathbf{y}^\top \mathbf{U}_j^\top,\; \mathbf{V} - \mathbf{V}^t \big\rangle \\ &=\; \mathbf{y}^\top \mathbf{U}_j^\top \Big[(\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V} + \mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}^t - \mathbf{M}^t \Big] \mathbf{U}_j \mathbf{y}, \end{aligned} \tag{S15}$$

where the second line collects terms and uses $\langle \widehat{\mathbf{\Sigma}} \mathbf{V}^t \mathbf{U}_j \mathbf{y} \mathbf{y}^\top \mathbf{U}_j^\top, \mathbf{V} \rangle = \mathbf{y}^\top \mathbf{U}_j^\top (\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V} \mathbf{U}_j \mathbf{y}$ together with the symmetrization $2\mathbf{y}^\top \mathbf{U}_j^\top (\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V} \mathbf{U}_j \mathbf{y} = \mathbf{y}^\top \mathbf{U}_j^\top [(\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V} + \mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}^t] \mathbf{U}_j \mathbf{y}$. Taking the minimum over unit $\mathbf{y}$ on both sides of (S15) and combining with (S14) gives (4). Note that only convexity of the quadratic was used, so the bound holds for all $\mathbf{V} \in \mathbb{R}^{d \times k}$, in particular on the set $\{\mathbf{V} : \mathbf{V}^\top \mathbf{V} \preccurlyeq \mathbf{I}_k\}$ of Proposition 1(iii).

For tightness, set $\mathbf{V} = \mathbf{V}^t$, and the bracket becomes $\mathbf{M}^t + \mathbf{M}^t - \mathbf{M}^t = \mathbf{M}^t$ and $\mathbf{U}_j^\top \mathbf{M}^t \mathbf{U}_j = \mathrm{diag}(\mu_1, \dots, \mu_j)$, whose smallest eigenvalue is $\mu_j = \lambda_j(\mathbf{M}^t)$. Finally, $g_j(\cdot | \mathbf{V}^t)$ is concave because $\lambda_{\min}$ is a concave function on $\mathbb{S}^{j \times j}$ and the inner map $\mathbf{V} \mapsto \mathbf{U}_j^\top [(\mathbf{V}^t)^\top \widehat{\mathbf{\Sigma}} \mathbf{V} + \mathbf{V}^\top \widehat{\mathbf{\Sigma}} \mathbf{V}^t - \mathbf{M}^t] \mathbf{U}_j$ is affine. ■

E.5 Proof of Lemma 4

Fix S with $|S| \le s$ and consider $\max\{\langle \mathbf{B}_S, \mathbf{V} \rangle : \mathbf{V}^\top \mathbf{V} = \mathbf{I}_k,\ \mathrm{rsupp}(\mathbf{V}) \subseteq S\}$. Deleting zero rows, this is $\max\{\langle \widetilde{\mathbf{B}}, \widetilde{\mathbf{V}} \rangle : \widetilde{\mathbf{V}}^\top \widetilde{\mathbf{V}} = \mathbf{I}_k\}$ with $\widetilde{\mathbf{B}} \in \mathbb{R}^{|S| \times k}$, whose value is $\|\widetilde{\mathbf{B}}\|_*$ with maximizer the polar factor of $\widetilde{\mathbf{B}}$; this is the classical orthogonal Procrustes solution [12]. Maximizing over S gives the stated value. For $k = 1$ the nuclear norm is the Euclidean norm, so $\max_{|S| \le s} \|\mathbf{b}_S\|_2$ is attained by the s largest entries in magnitude, and the polar factor is the normalization. ■

E.6 Proof of Theorem 5

(i). By construction, line 13 of Algorithm 1 sets $S \leftarrow S'$ only when $R' > R$, since line 11 ends the loop otherwise, where R' and R are the values of the exact objective $\min_j \lambda_j(\widehat{\boldsymbol{\Sigma}}_\cdot)/\kappa_j$ at S' and S. By Proposition 1(i) these are exactly $\widehat{R}$ evaluated at the corresponding optimal $\mathbf{V}$'s, so $\widehat{R}(\mathbf{V}^{t+1}) > \widehat{R}(\mathbf{V}^t)$ whenever the algorithm does not stop. (The candidate set $\mathcal{C}$ is generated from the MM surrogate by Lemmas 3–4, and the exact acceptance test guarantees that every accepted candidate raises the objective.)

(ii). The sequence $\{\widehat{R}(\mathbf{V}^t)\}_{t\geq 0}$ is non-decreasing by (i). It is bounded above by 1, since $\widehat{R}(\mathbf{V}) \leq 1$ for every $\mathbf{V} \in \mathcal{V}_s^k$ by Proposition 1(ii), and the iterates are feasible by construction, since each $\mathbf{V}^t$ is the matrix of top-k eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{S^t}$ with $|S^t| \leq s$, hence lies in $\mathcal{V}_s^k$. A non-decreasing real sequence bounded above converges, so $\widehat{R}(\mathbf{V}^t) \to R^\infty$ for some $R^\infty \leq 1$. By (iii) the loop exits after finitely many iterations, at some index $t^\star$, after which no update is made; the sequence is therefore constant from $t^\star$ onwards and $R^\infty = \widehat{R}(\mathbf{V}^{t^\star})$, the value the restart returns. The convergence is thus finite.

(iii). $\widehat{R}$ depends on $\mathbf{V}$ only through $\mathrm{rsupp}(\mathbf{V})$, by Proposition 1(i). Since the iterates have strictly increasing objective by (i), no support is visited twice, and there are at most $\binom{d}{s}$ supports; the loop therefore exits after finitely many iterations, and it exits through the test $R' \leq R$ whenever T is large enough.

(iv). At termination no candidate in $\mathcal{C}$ strictly improves the objective. The set $\mathcal{C}$ contains the truncated power step on each principal direction (move (a)), the deficit-weighted step (move (b)), the trace step (move (c)) and the exchange moves (d), so S^t is a local maximum with respect to the neighbourhood they generate. ■

E.7 Proof of Corollary 6

$\hat{\mathbf{V}}$ is feasible for $(\mathcal{P})$, so $\widehat{R}(\hat{\mathbf{V}}) \leq \mathrm{OPT}$; and $\mathrm{OPT} \leq 1$ by Proposition 1(ii). If $\widehat{R}(\hat{\mathbf{V}}) = 1$ both inequalities are equalities, so $\hat{\mathbf{V}}$ is optimal, and Proposition 1(ii) identifies its column space. ■

E.8 Proof of Theorem 7

Write $\lambda_j = \lambda_j(\boldsymbol{\Sigma})$ and let $S^\star = \bigcup_{i\leq k} \mathrm{supp}(\mathbf{v}_i)$, so $|S^\star| \leq s$ by assumption.

Step 1, the optimum is large. Let $\mathbf{V}^\star \in \mathcal{V}_s^k$ collect the top-k eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{S^\star}$ embedded in $\mathbb{R}^d$. Because the top-k eigenvectors of $\boldsymbol{\Sigma}$ are supported on $S^\star$, we have $\lambda_j(\boldsymbol{\Sigma}_{S^\star}) = \lambda_j$ for $j \in [k]$, and Weyl's inequality applied to $\widehat{\boldsymbol{\Sigma}}_{S^\star} = \boldsymbol{\Sigma}_{S^\star} + (\widehat{\boldsymbol{\Sigma}}_{S^\star} - \boldsymbol{\Sigma}_{S^\star})$ gives $\lambda_j(\widehat{\boldsymbol{\Sigma}}_{S^\star}) \geq \lambda_j - \varepsilon$. Hence, using $\kappa_j \leq (1+\eta)\lambda_j$,

$$\widehat{R}(\hat{\mathbf{V}}) \;\geq\; \widehat{R}(\mathbf{V}^\star) \;\geq\; \min_{j\in[k]} \frac{\lambda_j - \varepsilon}{(1+\eta)\lambda_j} \;\geq\; \frac{1-\varepsilon/\lambda_k}{1+\eta}, \tag{S16}$$

where the last step uses $\varepsilon/\lambda_j \leq \varepsilon/\lambda_k$ for $j \leq k$.

Step 2, transfer to the population. Let $\hat{S} = \mathrm{rsupp}(\hat{\mathbf{V}})$, so $|\hat{S}| \leq s$. By definition of $\widehat{R}$ and $\kappa_j \geq (1-\eta)\lambda_j$,

$$\lambda_j\Big(\hat{\mathbf{V}}^\top \widehat{\boldsymbol{\Sigma}} \hat{\mathbf{V}}\Big) \;\geq\; \widehat{R}(\hat{\mathbf{V}})\,\kappa_j \;\geq\; \widehat{R}(\hat{\mathbf{V}})(1-\eta)\lambda_j.$$

Because $\hat{\mathbf{V}}$ is supported on $\hat{S}$ with $|\hat{S}| \leq s$, we have $\|\hat{\mathbf{V}}^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\hat{\mathbf{V}}\|_{\mathrm{op}} \leq \|\widehat{\boldsymbol{\Sigma}}_{\hat{S}} - \boldsymbol{\Sigma}_{\hat{S}}\|_{\mathrm{op}} \leq \varepsilon$, so Weyl's inequality gives

$$\lambda_j\Big(\hat{\mathbf{V}}^\top \boldsymbol{\Sigma} \hat{\mathbf{V}}\Big) \;\geq\; \lambda_j\Big(\hat{\mathbf{V}}^\top \widehat{\boldsymbol{\Sigma}} \hat{\mathbf{V}}\Big) - \varepsilon \;\geq\; \frac{(1-\varepsilon/\lambda_k)(1-\eta)}{1+\eta}\lambda_j - \varepsilon \;=\; (1-\theta)\lambda_j - \varepsilon, \tag{S17}$$

which is the first display of the theorem. The bound $\theta \leq \varepsilon/\lambda_k + 2\eta$ follows from $1 - \frac{(1-x)(1-\eta)}{1+\eta} \leq x + \frac{2\eta}{1+\eta} \leq x + 2\eta$ for $x, \eta \in [0,1)$.

Step 3, from eigenvalues to angles. Let $\mathbf{V}, \mathbf{V}_\perp$ be the top-k eigenvectors of $\mathbf{\Sigma}$ and an orthonormal basis of the orthogonal complement, and put $\alpha^2 := \|\mathbf{V}_\perp^\top \hat{\mathbf{V}}\|_F^2 = k \sin^2 \angle(\hat{\mathbf{V}}, \mathbf{V})$. Decomposing $\hat{\mathbf{V}}$ in the two blocks,

$$\operatorname{tr}\Big(\hat{\mathbf{V}}^\top \mathbf{\Sigma} \hat{\mathbf{V}}\Big) = \big\langle \mathbf{\Sigma}_\parallel, \mathbf{V}^\top \hat{\mathbf{V}} \hat{\mathbf{V}}^\top \mathbf{V} \big\rangle + \big\langle \mathbf{\Sigma}_\perp, \mathbf{V}_\perp^\top \hat{\mathbf{V}} \hat{\mathbf{V}}^\top \mathbf{V}_\perp \big\rangle \;\le\; \sum_{j \le k} \lambda_j - \alpha^2(\lambda_k - \lambda_{k+1}),$$

because the first term is at most $\sum_{j \le k} \lambda_j - \alpha^2 \lambda_k$ (the trace of $\mathbf{V}^\top \hat{\mathbf{V}} \hat{\mathbf{V}}^\top \mathbf{V}$ is $k - \alpha^2$ and its eigenvalues lie in $[0, 1]$, so by Ky Fan the term is at most $\sum_{j \le k} \lambda_j$ minus $\alpha^2 \lambda_k$), while the second is at most $\alpha^2 \lambda_{k+1}$. On the other hand, summing (S17) over $j \in [k]$,

$$\operatorname{tr}\Big(\hat{\mathbf{V}}^\top \mathbf{\Sigma} \hat{\mathbf{V}}\Big) \;\ge\; (1 - \theta) \sum_{j \le k} \lambda_j - k\varepsilon.$$

Combining the two displays and using $\sum_{j \le k} \lambda_j \le k \lambda_1$,

$$\alpha^2(\lambda_k - \lambda_{k+1}) \;\le\; \theta \sum_{j \le k} \lambda_j + k\varepsilon \;\le\; k(\theta \lambda_1 + \varepsilon).$$

Setting 1 gives $\lambda_{k+1} \le (1 - \gamma)\lambda_k$, hence $\lambda_k - \lambda_{k+1} \ge \gamma \lambda_k$ and $\alpha^2 \le k(\theta \lambda_1 + \varepsilon)/(\gamma \lambda_k)$, and dividing by k, since $\alpha^2 = k \sin^2 \angle(\hat{\mathbf{V}}, \mathbf{V})$, yields (7).

Step 4, sparse normalizers. If $\kappa_j = \max_{|S| \le s} \lambda_j(\widehat{\mathbf{\Sigma}}_S)$, then on one hand $\kappa_j \ge \lambda_j(\widehat{\mathbf{\Sigma}}_{S^\star}) \ge \lambda_j - \varepsilon$, and on the other, for the maximizing S, $\lambda_j(\widehat{\mathbf{\Sigma}}_S) \le \lambda_j(\mathbf{\Sigma}_S) + \varepsilon \le \lambda_j + \varepsilon$ by Weyl and interlacing. Hence $|\kappa_j - \lambda_j| \le \varepsilon$ and $\eta \le \varepsilon / \lambda_k$. ■

Remark 10 (The recovery bound needs no global maximizer). Step 2 of the proof uses only the value $\widehat{R}(\hat{\mathbf{V}})$ and never the fact that $\hat{\mathbf{V}}$ is a global maximizer. Consequently, if Algorithm 1 returns an iterate with $\widehat{R}(\hat{\mathbf{V}}) \ge (1 - \varepsilon/\lambda_k)/(1 + \eta)$, the level that (S16) guarantees for the optimum, then (7) holds verbatim for the returned iterate, so the recovery guarantee of Theorem 7 extends to every output of the algorithm that reaches this level.

Corollary 11 (Sample complexity)**.** *Under Setting 1 with sparse normalizers there is an absolute constant C such that, for any $\delta \in (0, 1)$ and $\alpha \in (0, 1)$, if*

$$n \;\ge\; C \frac{\sigma^4}{\lambda_k^2} \cdot \frac{\lambda_1^2}{\alpha^2 \gamma^2 \lambda_k^2} \cdot \big(s \log d + \log(1/\delta)\big),$$

then with probability at least $1 - \delta$ every global maximizer of $(\mathcal{P})$ satisfies $\sin^2 \angle(\hat{\mathbf{V}}, \mathbf{V}) \le \alpha$.

E.9 Proof of Corollary 11

By a standard sub-Gaussian covariance concentration bound [27] applied to a fixed S with $|S| \le s$, and a union bound over the $\binom{d}{s} \le (ed/s)^s$ such subsets, there is an absolute constant C_0 with

$$\varepsilon \;\le\; C_0 \, \sigma^2 \left(\sqrt{\frac{s \log d + \log(1/\delta)}{n}} + \frac{s \log d + \log(1/\delta)}{n} \right)$$

with probability at least $1 - \delta$ [17]. In the regime $n \ge s \log d + \log(1/\delta)$ the first term dominates up to a constant, so $\varepsilon \le \tau \lambda_k$ as soon as $n \ge C_1(\sigma^2/(\tau \lambda_k))^2 (s \log d + \log(1/\delta))$. With sparse normalizers $\eta \le \varepsilon/\lambda_k \le \tau$, so $\theta \le 3\tau$ and (7) gives $\sin^2 \angle(\hat{\mathbf{V}}, \mathbf{V}) \le (3\tau \lambda_1 + \tau \lambda_k)/(\gamma \lambda_k) \le 4\tau \lambda_1/(\gamma \lambda_k)$. Choosing $\tau = \alpha \gamma \lambda_k / (4\lambda_1)$ and substituting gives the stated sample size, after absorbing numerical constants into C. ■

F The cascade proposal

Algorithm 1 restarts from the d coordinate seeds $\operatorname{top}_s((\widehat{\mathbf{\Sigma}}_{i:})^2)$, which keep the s coordinates of largest absolute covariance with coordinate i, that is, the supports one truncated power step reaches from the coordinate basis vectors at which RTPM starts [17]. For $k > 1$ we add one further restart, obtained from a cascade that extracts the components one at a time and is used purely as a proposal.

1. Set $\mathbf{P}_1 = \mathbf{I}_d$, and for $i = 1, \ldots, k$ run TRIM with $k = 1$ on $\mathbf{P}_i \widehat{\boldsymbol{\Sigma}} \mathbf{P}_i$ to obtain a sparse direction $\mathbf{u}_i$, normalize $\mathbf{u}_i \leftarrow \mathbf{P}_i \mathbf{u}_i / \|\mathbf{P}_i \mathbf{u}_i\|_2$, and set $\mathbf{P}_{i+1} = \mathbf{P}_i - \mathbf{u}_i \mathbf{u}_i^\top$.
2. Allocate the budget across the k proposal directions, starting from an even split $c_j \approx s/k$, greedily transfer one unit from c_b to c_a whenever doing so increases the *exact* objective $\min_j \lambda_j(\widehat{\boldsymbol{\Sigma}}_S)/\kappa_j$ of the union support $S = \bigcup_j \operatorname{top}_{c_j}(\mathbf{u}_j^2)$, until no transfer helps.

Cost of one restart. Forming $\widehat{\boldsymbol{\Sigma}}$ costs $O(nd^2)$ once. Within a restart, computing $\mathbf{G} = \widehat{\boldsymbol{\Sigma}}\mathbf{W}$ costs $O(dsk)$ since $\mathbf{V}$ has only s non-zero rows, and the moves (a)–(d) produce $O(kq^2)$ candidates, where q is the exchange radius, whose scoring costs $O(q^2 s^2 k^2)$ with a k-truncated eigensolver on each $s \times s$ submatrix. Over m restarts and T iterations this is $O(mT(dsk + q^2 s^2 k^2))$, against $O(nd^2 T)$ for RTPM, which restarts d times on a dense $d \times d$ product, and $O(d^3)$ per iteration for the Fantope SDP, whose ADMM solver eigendecomposes a dense $d \times d$ matrix at every step. Screening the d coordinate restarts down to the m most promising by their initial objective, which we do in every experiment with $k > 1$, bounds the number of restarts by m whatever the dimension.

The point of step 2 is that the known deficiency of the recursion, namely that the residual matrix need not have a sparse leading eigenvector, makes the cascade unreliable as an *algorithm* while leaving it a useful *initializer*, since the union support it produces is scored by the same exact objective as every other candidate and is discarded unless it wins. This is why TRIM inherits the reach of the cascade without inheriting its failure mode.

G Experimental details and additional results

G.1 Settings used for TRIM

The screening level m, the number of restarts retained, and the iteration cap T are budgets on computation, not parameters to tune. TRIM returns the best of its restarts and ascends monotonically within each, so enlarging m or T can never lower the returned objective, and every run in Figure S3(a) stops through the exact test well before its cap. The values recorded in the released drivers are $m = 15$ and $T = 25$ for the subspace experiments, $m = 20$ and $T = 20$ on 20 Newsgroups, and $m = 25$ and $T = 25$ for the convergence and runtime study. On 20 Newsgroups the coordinate seeds are taken at 250 evenly spaced coordinates, for TRIM and for every compared method that restarts from coordinates. At $k = 1$ all d coordinate restarts are run, with $T = 40$, and the ten best supports they reach are refined by the exchange moves. The ascent trajectories of Figure S3(a) are run with $T = 30$. The exchange radius is $q = 4$ everywhere, and the candidate moves (a)–(d) of Section 3 are active in every run.

G.2 Covariance models

All synthetic covariances are constructed to satisfy Setting 1. In every case $\boldsymbol{\Sigma}$ is reported exactly; samples are drawn as $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_i \mathbf{x}_i \mathbf{x}_i^\top$.

Model 1 (spiked identity). $\boldsymbol{\Sigma} = (1-\gamma)\mathbf{I}_d + \gamma \mathbf{v}\mathbf{v}^\top$ with $\mathbf{v}$ flat on $[s]$; $\gamma = 0.3$, so $\lambda_1 = 1$ and $\lambda_2 = 0.7$.

Model 2, diagonal-thresholding counterexample. $\boldsymbol{\Sigma} = \mathbf{v}\mathbf{v}^\top + \nu \sum_{i>s} \mathbf{e}_i \mathbf{e}_i^\top$ with $\mathbf{v}$ flat on $[s]$ and $\nu = 0.5$. Then $\boldsymbol{\Sigma}_{ii} = 1/s$ on $\operatorname{supp}(\mathbf{v})$ but $\nu = 0.5$ off it, so the s largest marginal variances miss the support entirely, while $\lambda_1 = 1 > 0.5 = \lambda_2$ so the model is well posed. This construction defeats diagonal thresholding by design [17]. We choose the background level so that the instance stays statistically feasible at the sample sizes we use, since a random $s \times s$ principal submatrix of the background block has top eigenvalue $\approx \nu(1 + \sqrt{s/n})^2$, which must stay below $\lambda_1 = 1$.

Model 2, general covariance. $\boldsymbol{\Sigma} = \mathbf{v}\mathbf{v}^\top + \mathbf{Q}\operatorname{diag}(\mu)\mathbf{Q}^\top + 0.05\,\mathbf{I}_d$ where $\mathbf{v}$ has i.i.d. Gaussian entries on $[s]$, normalized to unit length, $\mathbf{Q}$ is a random orthonormal basis of a $d/4$-dimensional subspace orthogonal to $\mathbf{v}$, and $\mu_i \sim \operatorname{Unif}(0.3g, g)$ with $g = 0.85$, a dense background spectrum pressed right up against the eigengap.

Block subspace model ($k > 1$). $\mathbf{\Sigma} = \sum_{j \le k} \lambda_j \mathbf{v}_j \mathbf{v}_j^\top + \nu \sum_{i \notin S^\star} \mathbf{e}_i \mathbf{e}_i^\top + 0.02 \mathbf{I}_d$ with $d = 300$, $k = 3$, $\lambda = (1, 0.45, 0.30)$, $\nu = 0.15$, and $\mathbf{v}_j$ flat on disjoint blocks P_j of size 10 whose union is $S^\star$, so $|S^\star| = 30$. This is Remark 9 with $r = 10$. The range of budgets in Figure 1(a) moves through the regime in which the trace criterion must choose which components to sacrifice.

G.3 How the baselines were run

Appendix B describes what each compared method does. All of them are given the same feature budget s and the same $\widehat{\mathbf{\Sigma}}$, and three choices are made on the competitors' behalf. The "Trace criterion" baseline runs TRIM's own search, with the same restarts, screening, cascade proposal, moves (a)–(d) and exact acceptance test, and differs from TRIM only in its objective, the Ky Fan value $\sum_{j \le k} \lambda_j(\widehat{\mathbf{\Sigma}}_S)$, so that any gap between the two is attributable to the *objective* alone. RTPM, alone and inside the sequential recursion, is run both at the truncation level its authors use when they compare it with other estimators and at the oversampled level their analysis prescribes, and every figure and table reports, cell by cell, the better of the two.

In the subspace experiment FPS is also granted an oracle-selected regularization parameter, chosen as the one that maximizes the *population* worst retention. We grant it because the theoretically motivated default selects only the leading block on this covariance, and reporting that default would penalize FPS for a poor choice, not for its criterion. No practitioner has access to the population covariance, so the curve labelled oracle-FPS is an upper bound on what that estimator could achieve in practice.

G.4 Numerical verification of the formal statements

Propositions 1, 2 and 8, Lemma 3 and Theorem 5(i) are additionally checked numerically by the script `verify_props.py` in the released code, which runs thirteen independent tests, all of which pass. Proposition 2 is verified by brute-force enumeration of all $\binom{d}{k}$ supports for $k \in \{2, 3, 4\}$, confirming that every trace maximizer attains $R = 0$ while $\max_S R = \frac{1}{2}$. Proposition 8 is verified by maximizing $\min_j (\mathbf{O}^\top \mathbf{M} \mathbf{O})_{jj}$ over 20,000 random rotations and comparing against $\mathrm{tr}(\mathbf{M})/k$. Proposition 1(i) is verified by checking that the top-k eigenvectors of $\widehat{\mathbf{\Sigma}}_S$ dominate 4,000 random orthonormal frames in all k eigenvalues simultaneously, and part (iii) by sampling 20,000 points with $\mathbf{V}^\top \mathbf{V} \preccurlyeq \mathbf{I}_k$ and confirming that none beats the orthonormal optimum. Lemma 3 is verified by evaluating both sides of (4) at 3,000 random supports and at $\mathbf{V}^t$ itself, and Theorem 5(i) by confirming that the objective is non-decreasing along each of 20 independent runs.

H Leading terms of the recovered components

Table S1 gives the eight highest-loading terms of each component for every method at the two extreme budgets, with the components of each method listed in the order in which it returns them.

The table complements Table 1 with what the supports contain. At $s = 60$ each of TRIM's four components is led by a different register, a general discussion vocabulary, a request and reply register (`thanks`, `advance`, `hi`), a religion cluster (`god`, `jesus`, `believe`) and an address cluster (`edu`, `email`), whereas the trace criterion mixes religion and games within a single component. Covariance thresholding spends much of its support on numerals (`10`, `15`, `12`) and on a few stems (`work`, `year`), and FPS draws numerals into every component at $s = 60$, so their supports carry little topical signal. Extracting the components one at a time places the terms `written` and `writing` in two of its four components at $s = 20$, so part of its tightest budget goes to a single family of stems.

Table S1: Highest-loading terms, 20 Newsgroups, $k = 4$.

s	method	eight highest-loading terms per component
20	DiagThresh (2009) [15]	1. know, don, think, just, does, people, like, edu 2. thanks, does, know, think, just, people, windows, edu 3. edu, think, thanks, god, like, drive, people, don 4. does, just, thanks, god, like, edu, use, windows
20	FPS (2013) [29]	1. thanks, does, edu, just, wouldn, world, work, won 2. does, edu, just, work, world, word, wonder, thanks 3. edu, just, thanks, does, work, world, word, women 4. thanks, just, does, edu, world, won, work, words
20	CovThresh (2016) [8]	1. year, yes, 00, young, works, working, world, worth 2. work, works, working, yes, 00, worked, worth, world 3. yes, wrong, word, work, year, world, 00, words 4. wrong, 00, yes, wouldn, year, words, word, working
20	Trace criterion (2020) [25]	1. thanks, think, don, people, just, advance, god, edu 2. know, does, thanks, don, like, advance, just, hi 3. does, god, just, like, thanks, think, don, jesus 4. edu, think, don, email, mail, people, does, god
20	Sequential RTPM (2026) [17]	1. don, know, think, just, people, written, writing, write 2. thanks, does, know, advance, windows, don, think, just 3. god, jesus, does, believe, say, thanks, know, advance 4. game, games, team, year, good, wrong, written, writing
20	**TRIM (ours)**	1. don, know, think, just, people, like, does, say 2. does, know, think, just, edu, people, good, don 3. edu, think, don, does, people, like, want, ve 4. god, just, like, think, people, believe, know, good
60	DiagThresh (2009) [15]	1. don, think, just, people, thanks, like, know, say 2. know, does, thanks, windows, mail, like, help, program 3. does, god, windows, use, just, know, problem, dos 4. game, god, games, edu, thanks, team, use, people

s	method	eight highest-loading terms per component
60	FPS (2013) [29]	1. thanks, does, edu, just, 10, 11, 15, 50 2. does, edu, just, 10, work, 15, 11, wrong 3. edu, just, thanks, does, wrong, work, 10, 16 4. thanks, does, just, edu, 10, 15, 11, 12
60	CovThresh (2016) [8]	1. 10, 15, 12, 00, 11, 16, 14, 13 2. year, years, yes, 00, 16, 12, 11, work 3. work, works, working, year, yes, worked, 10, 00 4. yes, wrong, year, word, work, world, years, 10
60	Trace criterion (2020) [25]	1. thanks, think, don, people, just, windows, god, advance 2. know, does, thanks, don, like, advance, just, windows 3. god, game, does, think, like, jesus, just, games 4. game, use, windows, does, just, games, team, know
60	Sequential RTPM (2026) [17]	1. thanks, know, does, advance, hi, mail, windows, think 2. don, think, just, people, know, like, thanks, does 3. god, jesus, believe, christ, christian, bible, faith, life 4. game, games, team, year, hockey, good, play, baseball
60	**TRIM (ours)**	1. don, think, people, just, thanks, god, like, say 2. know, thanks, does, advance, like, hi, mail, don 3. god, does, just, like, jesus, think, don, believe 4. edu, think, don, email, mail, people, does, god

I SINGLE-COMPONENT EXPERIMENTS IN DETAIL

With $k = 1$ the criterion (2) *is* ordinary sparse PCA, so this experiment tests the optimizer alone. Figure S2 reports $\cos^2$ alignment with the planted direction at $d = 200$, $s = 16$, over 20 replications.

On the spiked identity model (panel a) TRIM matches or exceeds RTPM from $n = 1000$ onward, and the margin is widest in the transition region, at 77.1% against 61.4% for $n = 2000$ and 96.2% against 92.5% for $n = 4000$; at the two smallest sample sizes the two are indistinguishable, neither recovering the planted direction. Both restart from every coordinate, and the exchange moves (d) find supports that a truncated power iteration cannot reach once it has settled. On the diagonal-thresholding counterexample (panel b) diagonal thresholding returns alignment exactly zero at every sample size, a structural failure that no sample size repairs, while TRIM and RTPM both recover perfectly from $n = 1000$. On a general covariance (panel c) the two are indistinguishable and approach full alignment, while diagonal thresholding stays between 76% and 85%. Although FPS leads at the smallest sample sizes, as a convex relaxation (in the sense of losing tightness) with a $d \times d$ variable is expected to, TRIM overtakes it on the spiked identity model by $n = 4000$ (96.2% against 92.7%) and matches it on a general covariance from $n = 1000$, at about a fifth of the cost (0.54 s against 2.64 s at $n = 2000$). TRIM leads GreedyCorr on the spiked identity model by a wide margin (77.1% against 45.2% at $n = 2000$ and 96.2% against 57.6% at $n = 4000$) and stays ahead at every sample size on the general covariance; on the counterexample the two are on a par from $n = 1000$, where both recover perfectly, and at $n = 500$ the greedy heuristic is the more reliable of

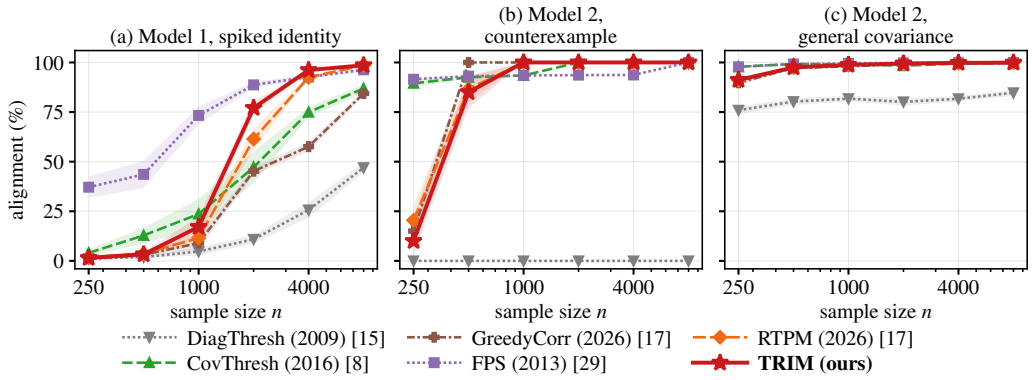

Figure S2: Single-component recovery, $d = 200$, $s = 16$.

the two, with TRIM and RTPM alike recovering in most replications but not in all. Moving to the worst-case criterion thus costs no accuracy at $k = 1$, as Lemma 4 anticipates, since there the update of TRIM is itself a truncated power step.

J Convergence, optimality bounds and runtime in detail

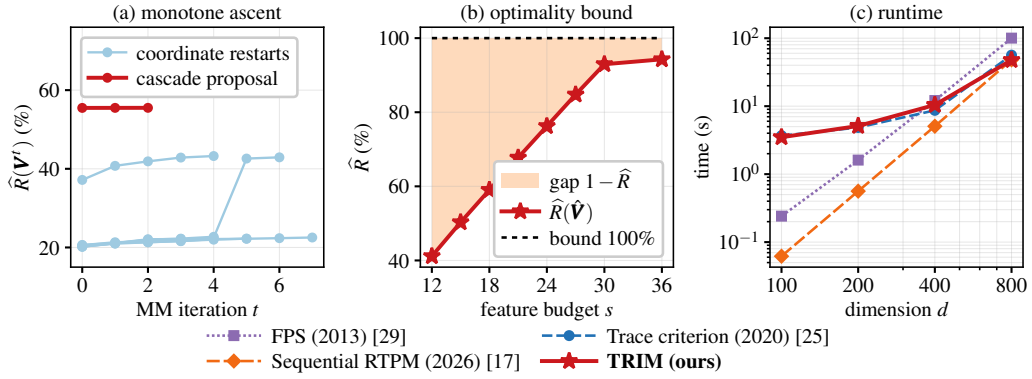

Figure S3: Optimization behaviour and cost of Algorithm 1.

Panel (a) traces the objective from seven restarts. Every trajectory is non-decreasing, as Theorem 5(i) requires, and each terminates through the exact test $R' \le R$ after between three and eight iterations, well before the cap $T = 30$, so part (iii) is not merely a finiteness statement in principle. The cascade proposal enters at $\widehat{R} = 55.5\%$, already above the 43.2% that the best coordinate restart reaches.

Panel (b) plots the optimality bound of Corollary 6. The shaded region is the gap $1 - \widehat{R}(\hat{\mathbf{V}})$, narrowing monotonically from 58.9% at $s = 12$ to 5.7% at $s = 36$. This bound is computed from the returned solution alone, with no knowledge of $\boldsymbol{\Sigma}$ and no reference solution. The panel shows that the criterion supplies an absolute scale on which the optimality of every component is judged at once, which no aggregate of eigenvalues provides.

Panel (c) gives the cost, and the picture improves for TRIM as the problem grows. At $d = 800$ TRIM is the fastest of the four methods (47.7 s), ahead of the trace criterion, which runs the same search and costs about as much at every dimension (56.3 s), and $2.1\times$ faster than the Fantope SDP (100.3 s), with a gap in seconds that widens with d because the SDP carries a $d \times d$ variable while screening bounds TRIM's restarts. At the smallest dimensions every method finishes within a few seconds (3.5 s for TRIM at $d = 100$), where TRIM's fixed costs are large relative to the problem size and runtime is not the binding constraint.

Table S2: Wall-clock seconds for one fit, $n = 2000$, $k = 3$, $s = 24$, single-threaded.

method	$d = 100$	$d = 200$	$d = 400$	$d = 800$
FPS (2013) [29]	<u>0.24</u>	<u>1.61</u>	12.11	100.30
Trace criterion (2020) [25]	3.68	4.86	<u>8.68</u>	56.31
Sequential RTPM (2026) [17]	**0.06**	**0.56**	**5.07**	<u>48.04</u>
TRIM (ours)	3.49	5.11	10.44	**47.66**

None of the baselines solves the worst-case problem, so Table S2 sets TRIM against methods that pursue other objectives, and it shows the worst-case criterion reached at a cost comparable to theirs, which is further evidence that the approach is practical.