DIAGONAL ATTENUATION: A FINITE-SAMPLE CORRECTION FOR PCA

Anonymous authors
Paper under double-blind review

ABSTRACT

Principal component analysis (PCA) can rotate away from its population target when a covariance matrix is estimated from limited data. We introduce *diagonal attenuation*, which preserves sample cross-covariances while reducing coordinate-wise sample variances. The method is revealed exactly by averaging a linear full-output reconstruction loss over random input masks; studying the correction directly extends it beyond the range attainable by masking. We isolate the part of the random coupling between retained and omitted population directions that is contributed by sample-variance errors, and show how attenuation can reduce the resulting rotation. Under balanced marginal variances, we derive an explicit expected-risk theorem, uniform over the attenuation path for all sufficiently large finite samples, and obtain the asymptotically risk-minimizing strength. For general covariances, we characterize when attenuation leaves the population PCA subspace unchanged and give a risk theorem that also accounts for changing eigengaps and the population cost when the target moves. Simulations track this tradeoff from exact preservation back to PCA. Across local image patches, speech spectra, and smartphone acceleration, both mask-derived and direct attenuation improve PCA under two fitting-sample budgets, and one of them has the largest mean gain among seven methods in every data–budget cell. The full path selects strengths beyond the mask-derived boundary on 63%–95% of the subsamples.

1 INTRODUCTION

Principal component analysis (PCA) estimates a low-dimensional subspace by diagonalizing a sample covariance matrix. With limited data, sampling fluctuations couple population directions on opposite sides of the rank cutoff and rotate the fitted subspace away from its population target; the resulting error is controlled by both these random couplings and the spectral gaps that resist them (Anderson, 1963; Koltchinskii & Lounici, 2017; Reiß & Wahl, 2020; El Hanchi et al., 2025). A less visible source of rotation is error in the coordinatewise sample variances. These errors form a diagonal matrix in the observed coordinates, but the same matrix generally has off-diagonal entries in the population principal-component basis. Those entries couple a direction retained by PCA with one it omits and can therefore rotate the fitted subspace. This identifies a specific correction: can weakening the sample diagonal reduce PCA's finite-sample reconstruction risk while retaining its informative sample cross-covariances?

Input masking reveals precisely this correction. Averaging a full-output orthogonal-projector reconstruction loss over independent Bernoulli input masks yields PCA applied to

$$\mathbf{T}_{n,\alpha} = \mathbf{S}_n - \alpha \operatorname{diag}(\mathbf{S}_n), \qquad \alpha = \frac{m}{1+m}, \tag{1.1}$$

where $\mathbf{S}_n = n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i^\mathsf{T}$ is the sample covariance of n centered observations $\mathbf{x}_i$, $\operatorname{diag}(\mathbf{S}_n)$ retains its diagonal, and m is the mask rate. This *diagonal attenuation* preserves sample cross-covariances while multiplying sample variances by $1-\alpha$. Figure 1 summarizes the estimator path; Section 2 gives the exact derivation.

Input masking reaches only $0 \le \alpha < 1/2$. We study the full path $0 \le \alpha \le 1$, from ordinary PCA to diagonal-deletion (hollowed) PCA (Cai et al., 2021; Abbe et al., 2022), with strength selected by reconstruction validation. Attenuation changes random couplings across the PCA rank cutoff,

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.

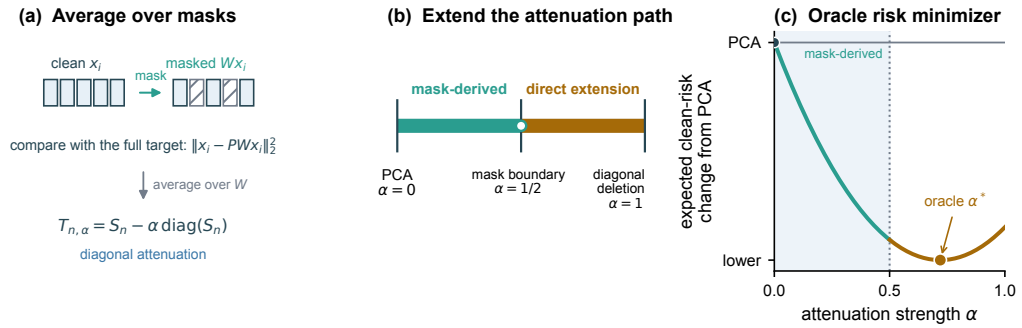

Figure 1: Method overview. **Left:** averaging the full-output reconstruction loss in (2.4) over random input masks yields (1.1). **Center:** input masking produces the first part of the attenuation path, while direct attenuation extends the same correction from PCA to diagonal deletion. **Right:** the oracle strength balances the reduction in sampling-induced rotation against any population cost caused by moving the target, and may lie beyond the mask-derived range.

which we call *sampling leakage*, but may also alter the population subspace and eigengaps. Balanced marginal variances isolate the sampling effect; general covariances expose its tradeoff with population movement. The paper makes three contributions. First, we derive diagonal attenuation exactly from masked-input reconstruction, extend the mask-derived range to a directly tunable PCA estimator, and characterize when the population PCA subspace is preserved or moved. Second, we isolate the sample-diagonal contribution to the standard cross-subspace sampling fluctuation and show how attenuation changes expected reconstruction risk. Under balanced marginal variances, this yields an explicit theorem and the asymptotically risk-minimizing strength; for general covariances, our risk theorems account for changing eigengaps and population bias. Third, simulations verify the predicted risk tradeoff across preserved and moving targets, while fitting-sample-budget studies on local image patches, speech spectra, and smartphone acceleration compare seven methods under common budgets. Both attenuation paths improve PCA in every cell, and one has the largest mean gain in all six cells.

1.1 Related work

PCA and covariance regularization. Classical and modern PCA theory characterizes eigenspace fluctuations and reconstruction risk (Anderson, 1963; Koltchinskii & Lounici, 2017; Reiß & Wahl, 2020; El Hanchi et al., 2025). Shrinkage toward a scalar identity preserves sample eigenvectors (Ledoit & Wolf, 2004), while diagonal-target shrinkage also changes cross-covariances (Schäfer & Strimmer, 2005). Diagonal deletion and HeteroPCA address noisy or incomplete low-rank measurement models (Cai et al., 2021; Abbe et al., 2022; Zhang et al., 2022). Here complete observations come from the covariance whose PCA subspace is the target; we study when partial diagonal attenuation improves its clean reconstruction risk.

Masking and linear reconstruction. Denoising and marginalized autoencoders learn from corrupted inputs or integrate out corruption (Vincent et al., 2008; 2010; Chen et al., 2012; 2014); unmasked linear autoencoders recover PCA (Baldi & Hornik, 1989). Our full-output projector formulation uses ordinary population PCA as the evaluation target and clean reconstruction risk as the loss. Canonical masked autoencoders instead score hidden coordinates (He et al., 2022); Remark E.1 discusses this distinction.

Closest spectral analyses. Ji et al. (2023) relate complementary contrastive masks to diagonal deletion in a spiked noise model. Bisulco et al. (2025) characterize full-output block masking with untied linear encoders and decoders through a generalized eigenproblem. Zurich et al. (2026) analyze full-rank masked regression in proportional asymptotics. Our fixed-rank orthogonal projector instead permits a fixed-dimensional expected-risk comparison with ordinary PCA along a continuous attenuation path. This identifies the specific sampling couplings, eigengaps, and target movement behind the broader spectral-regularization effect of augmentation (Lin et al., 2024). Additional distinctions appear in Section E.4.

1.2 Notation

For a symmetric matrix A, $\lambda_j(A)$ is its jth largest eigenvalue, $\mathrm{diag}(A)$ retains its diagonal, and $\mathrm{off}(A) = A - \mathrm{diag}(A)$. The operator and Frobenius norms are $\|A\|_{\mathrm{op}}$ and $\|A\|_{\mathrm{F}}$, and $\circ$ denotes the elementwise product. The set of rank-d orthogonal projectors is $\mathcal{P}_d$; P_A denotes a leading rank-d projector of A, with a fixed rule for ties. The mask rate is m, and α is the diagonal-attenuation strength. Additional quantities are defined where they first appear.

2 Diagonal attenuation for PCA

2.1 PCA target and clean risk

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be centered observations with covariance $\mathbf{\Sigma}$, and let $\mathcal{P}_d$ contain the rank-d orthogonal projectors, $P = UU^\mathsf{T}$ with $U^\mathsf{T}U = I_d$. Write

$$\mathbf{S}_n = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T}. \tag{2.1}$$

For an independent unmasked observation, the reconstruction risk and its excess over the leading population projector P_* are

$$\mathcal{R}(P) = \mathbb{E}\|\mathbf{x} - P\mathbf{x}\|_2^2 = \mathrm{tr}(\mathbf{\Sigma}) - \mathrm{tr}(P\mathbf{\Sigma}), \tag{2.2}$$

$$\mathcal{E}(P) = \mathcal{R}(P) - \mathcal{R}(P_*) = \mathrm{tr}\{\mathbf{\Sigma}(P_* - P)\}. \tag{2.3}$$

2.2 From input masking to diagonal attenuation

For a random diagonal input mask W, consider full-output reconstruction:

$$\widehat{P}^{\mathrm{mask}} \in \operatorname*{arg\,min}_{P \in \mathcal{P}_d} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_W \|\mathbf{x}_i - PW\mathbf{x}_i\|_2^2. \tag{2.4}$$

Let $W = \mathrm{diag}(w_1, \ldots, w_p)$, where the w_j are independent $\mathrm{Bernoulli}(q)$ variables. Each coordinate is retained with probability q, and $m = 1 - q$ is the mask rate. Averaging over this mask distribution gives the attenuation parameter and matrix

$$\alpha = \frac{1-q}{2-q} = \frac{m}{1+m}, \qquad \mathbf{T}_{n,\alpha} = \mathbf{S}_n - \alpha \cdot \mathrm{diag}(\mathbf{S}_n). \tag{2.5}$$

Mask rates $0 \le m < 1$ correspond exactly to $0 \le \alpha < 1/2$.

Proposition 2.1 (Exact mask marginalization)**.** *Every optimizer of* (2.4) *is a leading rank-d spectral projector of* $\mathbf{T}_{n,\alpha}$*. If* $\lambda_d(\mathbf{T}_{n,\alpha}) > \lambda_{d+1}(\mathbf{T}_{n,\alpha})$*, then*

$$\widehat{P}^{\mathrm{mask}} = P_{\mathbf{T}_{n,\alpha}}. \tag{2.6}$$

Proposition 2.1 shows that masking reveals a deterministic covariance correction: it retains sample cross-covariances and multiplies sample variances by $1 - \alpha$. Because masking reaches only $0 \le \alpha < 1/2$, we extend the same path directly and define

$$\widehat{P}_\alpha = P_{\mathbf{T}_{n,\alpha}}, \qquad 0 \le \alpha \le 1, \tag{2.7}$$

where P_A denotes a leading rank-d projector of a symmetric matrix A. Ordinary PCA corresponds to $\alpha = 0$, while $\alpha = 1$ is diagonal-deletion PCA based on $\mathrm{off}(\mathbf{S}_n)$. The full interval can be searched by held-out reconstruction with PCA included at the endpoint $\alpha = 0$. For a finite candidate grid, an independent-validation implementation gives a Gaussian high-probability fallback guarantee: without post-selection refitting, the selected projector has clean risk no greater than the PCA candidate plus an error that vanishes with the validation sample size (Proposition D.2).

The full-output projector objective differs from masked-target and unconstrained linear reconstruction; see Remark E.1.

2.3 POPULATION TARGET UNDER ATTENUATION

Write

$$D = \mathrm{diag}(\mathbf{\Sigma}), \qquad \mathbf{T}_\alpha = \mathbf{\Sigma} - \alpha D, \qquad P_*^\perp = I_p - P_*. \tag{2.8}$$

Because D may contain directional information, attenuation can move the population PCA target. We measure the resulting population reconstruction cost by

$$\mathsf{B}_{\mathrm{pop}}(\alpha) = \mathcal{E}(P_{\mathbf{T}_\alpha}) = \mathrm{tr}\{\mathbf{\Sigma}(P_* - P_{\mathbf{T}_\alpha})\}. \tag{2.9}$$

This is the reconstruction cost that remains without sampling error. Attenuation mixes the target with its complement through

$$P_*^\perp \mathbf{T}_\alpha P_* = -\alpha P_*^\perp D P_*. \tag{2.10}$$

The block gap compares the smallest retained and largest discarded population Rayleigh quotients:

$$\delta_\alpha = \min_{\|v\|_2=1,\, P_* v = v} v^\mathsf{T} \mathbf{T}_\alpha v - \max_{\|v\|_2=1,\, P_* v = 0} v^\mathsf{T} \mathbf{T}_\alpha v. \tag{2.11}$$

A positive δ_α means that the attenuated population variance retained by P_* remains strictly above that discarded by P_*.

Proposition 2.2 (Zero target-shift characterization)**.** *Fix $\alpha > 0$. The projector P_* is the unique leading rank-d projector of $\mathbf{T}_\alpha$ if and only if*

$$P_*^\perp D P_* = 0, \qquad \delta_\alpha > 0. \tag{2.12}$$

Consequently, $\mathsf{B}_{\mathrm{pop}}(\alpha) = 0$ under these conditions.

The first condition prevents mixing; the second prevents a reversal of spectral ordering. Our risk analysis uses the following standing assumptions.

Assumption 2.1 (Gaussian sampling)**.** The observations satisfy $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} \succ 0$.

Assumption 2.2 (Fixed-dimensional identifiable target)**.** The dimension p and target rank d are fixed as n grows, and $\delta := \lambda_d(\mathbf{\Sigma}) - \lambda_{d+1}(\mathbf{\Sigma}) > 0$.

Assumptions 2.1 and 2.2 are standard in fixed-dimensional PCA analyses (Anderson, 1963; Koltchinskii & Lounici, 2017; Reiß & Wahl, 2020; Yu et al., 2015). Gaussianity gives transparent risk formulas; the supplement extends them to distributions with finite moments. The analysis treats the population mean as known and zero, while in practice the observations are typically centered empirically. Under Gaussian sampling, the centered sample covariance has the same eigenspace distribution as a zero-mean sample covariance based on $n-1$ observations, so the same expected-risk results in Section 3 apply with sample size $n-1$. Additional structure is introduced where the risk analysis requires.

3 WHY ATTENUATION CAN IMPROVE FINITE-SAMPLE PCA

For any symmetric matrix M, the cross-block $P_*^\perp M P_*$ records how M connects the population PCA subspace to its orthogonal complement. We call this block the *cross-subspace leakage*; its individual entries are couplings between retained and omitted directions. For the attenuated sample matrix $\mathbf{T}_{n,\alpha}$, leakage separates into a deterministic population component and a random sampling component. Write $E_n = \mathbf{S}_n - \mathbf{\Sigma}$. Since $P_*^\perp \mathbf{\Sigma} P_* = 0$, diagonal attenuation gives the exact decomposition

$$P_*^\perp \mathbf{T}_{n,\alpha} P_* = \underbrace{-\alpha P_*^\perp D P_*}_{\text{population leakage}} + \underbrace{P_*^\perp \{E_n - \alpha\, \mathrm{diag}(E_n)\} P_*}_{\text{sampling leakage}}. \tag{3.1}$$

The first term determines whether attenuation moves the population target. The second is mean-zero *sampling leakage*; its retained–omitted couplings drive sampling-induced rotation of the fitted subspace. We first isolate this sampling effect under balanced marginal variances, where the population term vanishes and all population eigengaps remain fixed. We then allow unequal marginals, which can change the eigengaps and, when the target moves, introduce population leakage.

3.1 BALANCED MARGINALS: REDUCING SAMPLING-INDUCED ROTATION

Balanced marginal variances isolate the finite-sample effect of attenuation.

Assumption 3.1 (Balanced marginal variances)**.** For some $\tau > 0$, $D = \tau I_p$.

The assumption equalizes the variances of the observed coordinates but does not make the covariance isotropic: correlations, nontrivial principal directions, and unequal eigenvalues remain allowed. At the population level,

$$\mathbf{T}_\alpha = \mathbf{\Sigma} - \alpha\tau I_p. \tag{3.2}$$

Therefore attenuation leaves the population principal subspace and every eigengap unchanged. Any change in the estimated subspace must therefore come from the sample fluctuation matrix $\mathbf{T}_{n,\alpha} - \mathbf{T}_\alpha$, specifically from its couplings across the rank-d cutoff. To see this, choose an orthonormal eigenbasis $u_1, \ldots, u_p$ of $\mathbf{\Sigma}$, ordered so that $\mathbf{\Sigma}u_r = \lambda_r u_r$, and fix $j \le d < k$. For one observation define

$$a_{jk}(\mathbf{x}) = u_k^\mathsf{T}(\mathbf{x}\mathbf{x}^\mathsf{T} - \mathbf{\Sigma})u_j, \qquad b_{jk}(\mathbf{x}) = u_k^\mathsf{T}\,\mathrm{diag}(\mathbf{x}\mathbf{x}^\mathsf{T} - \mathbf{\Sigma})u_j = \sum_{\ell=1}^{p} u_{k\ell}u_{j\ell}(x_\ell^2 - \Sigma_{\ell\ell}), \tag{3.3}$$

where $u_{r\ell}$ is the ℓ-th coordinate of u_r. The variable a_{jk} is one observation's contribution to the unattenuated coupling between the retained direction u_j and the omitted direction u_k. The variable b_{jk} is the part of that contribution arising from coordinatewise variance fluctuations. Both are mean zero. If $\overline{a}_{jk}$ and $\overline{b}_{jk}$ denote their sample averages, then the sampling-leakage matrix collects all residual couplings across the rank cutoff:

$$\overline{Z}_{n,\alpha} := P_*^\perp(\mathbf{T}_{n,\alpha} - \mathbf{T}_\alpha)P_* = \sum_{j\le d<k} (\overline{a}_{jk} - \alpha\overline{b}_{jk})u_k u_j^\mathsf{T}. \tag{3.4}$$

The residual coupling $\overline{a}_{jk} - \alpha\overline{b}_{jk}$ produces a first-order rotation of size $(\overline{a}_{jk} - \alpha\overline{b}_{jk})/(\lambda_j - \lambda_k)$. Its squared contribution to clean risk is weighted by the same eigengap, leaving one eigengap in the denominator. Thus attenuation helps by reducing residual coupling variance. The complete rotation and risk expansion is stated in Lemma G.1 in the supplement. It remains to determine whether attenuation reduces the variance of the residual coupling. Under Gaussian sampling, Isserlis' identity (Isserlis, 1918) gives

$$\mathrm{Cov}\{a_{jk}(\mathbf{x}), b_{jk}(\mathbf{x})\} = 2\lambda_j\lambda_k \sum_{\ell=1}^{p} u_{j\ell}^2 u_{k\ell}^2 \ge 0, \quad \text{and hence} \tag{3.5}$$

$$\mathrm{var}\{a_{jk}(\mathbf{x}) - \alpha b_{jk}(\mathbf{x})\} - \mathrm{var}\{a_{jk}(\mathbf{x})\} = -2\alpha\,\mathrm{Cov}\{a_{jk}(\mathbf{x}), b_{jk}(\mathbf{x})\} + \alpha^2\,\mathrm{var}\{b_{jk}(\mathbf{x})\}. \tag{3.6}$$

The nonnegative overlap term in (3.5) quantifies how strongly the sample-diagonal component aligns with the unattenuated coupling and thus why a small positive attenuation reduces its variance when this overlap is positive; the quadratic term in (3.6) explains why excessive attenuation can eventually increase it. Aggregating these contributions across the rank cutoff, define

$$\begin{aligned} \mathcal{K}(\alpha) &= \sum_{j\le d<k} \frac{\mathrm{var}\{a_{jk}(\mathbf{x}) - \alpha b_{jk}(\mathbf{x})\}}{\lambda_j - \lambda_k}, \\ \mathsf{C}_{\mathbf{\Sigma}} = \sum_{j\le d<k} \frac{\mathrm{Cov}\{a_{jk}(\mathbf{x}), b_{jk}(\mathbf{x})\}}{\lambda_j - \lambda_k}, &\qquad \mathsf{V}_{\mathbf{\Sigma}} = \sum_{j\le d<k} \frac{\mathrm{var}\{b_{jk}(\mathbf{x})\}}{\lambda_j - \lambda_k}. \end{aligned} \tag{3.7}$$

Theorem 3.1 aggregates these pairwise contributions and controls the expected risk uniformly over the full attenuation path.

Theorem 3.1 (Expected risk under balanced marginals)**.** *Under Assumptions 2.1, 2.2 and 3.1, there are finite constants C_0 and N, independent of α, such that*

$$\sup_{0\le\alpha\le 1} \left| n \cdot \mathbb{E}\mathcal{E}(\widehat{P}_\alpha) - \mathcal{K}(\alpha) \right| \le \frac{C_0}{\sqrt{n}}, \qquad n \ge N. \tag{3.8}$$

Moreover, $\mathsf{C}_{\mathbf{\Sigma}}, \mathsf{V}_{\mathbf{\Sigma}} > 0$ and, for every $0 \le \alpha \le 1$,

$$\mathcal{K}(\alpha) - \mathcal{K}(0) = -2\alpha\mathsf{C}_{\mathbf{\Sigma}} + \alpha^2\mathsf{V}_{\mathbf{\Sigma}}. \tag{3.9}$$

Because Theorem 3.1 shows that $\mathcal{K}(\alpha)/n$ approximates the expected clean excess risk, we call $\mathcal{K}(\alpha)$ the n-scaled theoretical risk curve. Its minimizers over the closures of the mask-derived and direct ranges are

$$\alpha^*_{\mathrm{mask,cl}} = \min\left\{\frac{\mathsf{C}_{\mathbf{\Sigma}}}{\mathsf{V}_{\mathbf{\Sigma}}}, \frac{1}{2}\right\}, \qquad \alpha^*_{\mathrm{direct}} = \min\left\{\frac{\mathsf{C}_{\mathbf{\Sigma}}}{\mathsf{V}_{\mathbf{\Sigma}}}, 1\right\}. \tag{3.10}$$

For any $n \geq N$, the theorem guarantees lower expected risk than PCA whenever $2\alpha\mathsf{C}_{\mathbf{\Sigma}} - \alpha^2\mathsf{V}_{\mathbf{\Sigma}} > 2C_0/\sqrt{n}$. Consequently, every fixed $0 < \alpha < \min\{2\mathsf{C}_{\mathbf{\Sigma}}/\mathsf{V}_{\mathbf{\Sigma}}, 1\}$ improves on PCA for all sufficiently large finite samples. A larger mask rate need not be better, and the optimum can lie beyond the mask-derived range. The notation "cl" denotes closure: $\alpha = 1/2$ is approached as the mask rate tends to one, whereas direct attenuation includes the entire interval $[0, 1]$; see Figure 1.

As a concrete example, the rank-one flat-spike model $\mathbf{\Sigma} = \sigma^2 I_p + \theta uu^\mathsf{T}$ with $u = p^{-1/2}\mathbf{1}$, $\sigma^2, \theta > 0$, where $\mathbf{1}$ is the all-ones vector, has balanced marginals. Moreover, $\mathsf{C}_{\mathbf{\Sigma}}/\mathsf{V}_{\mathbf{\Sigma}} \geq 1$, so complete attenuation is optimal on the direct path. Figure 2(a) verifies the theoretical risk curve in a balanced multispike design.

3.2 General covariance: changing eigengaps and population leakage

When marginal variances are unequal, attenuation has two distinct population-level effects. It can change the eigengaps that resist sampling-induced rotation even when the population PCA subspace is unchanged, and it can create population leakage that moves this subspace. We begin with the target-preserving case, allowing the eigengaps to vary with α. Recall the block gap δ_α from (2.11).

Assumption 3.2 (Target preservation)**.** For a compact attenuation interval $\mathcal{I} \subset [0, 1]$ containing zero,

$$P_*^\perp D P_* = 0, \qquad \underline{\delta}_{\mathcal{I}} := \inf_{\alpha \in \mathcal{I}} \delta_\alpha > 0. \tag{3.11}$$

The first condition eliminates population leakage, so $\mathbf{T}_\alpha$ leaves the range of P_* invariant. The positive block gap keeps this subspace above its orthogonal complement in the spectral ordering. Together, the two conditions preserve the leading rank-d population subspace over the attenuation interval. Neither requires equal coordinate variances: Figure 2(b) gives an unequal-marginal example satisfying both.

The general analysis collects the residual couplings into a matrix:

$$\begin{aligned} Z_\alpha(\mathbf{x}) &= P_*^\perp \left[\mathbf{x}\mathbf{x}^\mathsf{T} - \mathbf{\Sigma} - \alpha\,\mathrm{diag}(\mathbf{x}\mathbf{x}^\mathsf{T} - \mathbf{\Sigma})\right] P_*, \\ \overline{Z}_{n,\alpha} &= \frac{1}{n}\sum_{i=1}^n Z_\alpha(\mathbf{x}_i) = P_*^\perp(\mathbf{T}_{n,\alpha} - \mathbf{T}_\alpha)P_*. \end{aligned} \tag{3.12}$$

Its (u_k, u_j) entry is $a_{jk} - \alpha b_{jk}$. Define the space of maps from the target into its orthogonal complement,

$$\mathbb{X} = \{X \in \mathbb{R}^{p\times p} : X = P_*^\perp X P_*\}. \tag{3.13}$$

On this space, define the attenuated gap operator

$$\mathcal{L}_\alpha(X) = X(P_*\mathbf{T}_\alpha P_*) - (P_*^\perp \mathbf{T}_\alpha P_*^\perp)X. \tag{3.14}$$

The positive block gap makes $\mathcal{L}_\alpha$ invertible; in eigenbases of the two attenuated blocks, its inverse divides each cross-block entry by its eigengap, giving the leading rotation

$$X_{n,\alpha} = \mathcal{L}_\alpha^{-1}\overline{Z}_{n,\alpha}. \tag{3.15}$$

To convert rotation into clean risk, weight it by the original covariance gaps using

$$\mathcal{H}_{\mathbf{\Sigma}}(X) = X(P_*\mathbf{\Sigma}P_*) - (P_*^\perp \mathbf{\Sigma} P_*^\perp)X. \tag{3.16}$$

Writing $\langle A, B\rangle_{\mathrm{F}} = \mathrm{tr}(A^\mathsf{T}B)$, the theoretical risk curve is

$$\mathcal{K}(\alpha) = \mathbb{E}\left\langle \mathcal{L}_\alpha^{-1} Z_\alpha(\mathbf{x}), \mathcal{H}_{\mathbf{\Sigma}}\mathcal{L}_\alpha^{-1} Z_\alpha(\mathbf{x})\right\rangle_{\mathrm{F}}. \tag{3.17}$$

This is the coordinate-free counterpart of the balanced pairwise risk; the derivation appears in Section C.2.

Theorem 3.2 (Expected risk under target preservation)**.** *Under Assumptions 2.1, 2.2 and 3.2, there are finite constants C and N, independent of α, such that*

$$\sup_{\alpha\in\mathcal{I}}\left|n\mathbb{E}\mathcal{E}(\widehat{P}_\alpha)-\mathcal{K}(\alpha)\right|\le\frac{C}{\sqrt{n}},\qquad n\ge N. \tag{3.18}$$

For any $n \geq N$, the theorem guarantees lower expected risk than PCA whenever $\mathcal{K}(0) - \mathcal{K}(\alpha) > 2C/\sqrt{n}$. Unlike in the balanced case, this comparison now depends on both the random variation in the residual couplings collected by Z_α and the attenuated eigengaps represented by $\mathcal{L}_\alpha$. When $D = \tau I_p$, $\mathcal{L}_\alpha = \mathcal{L}_0 = \mathcal{H}_{\mathbf{\Sigma}}$, and (3.17) reduces exactly to the pairwise formula (3.7).

Target preservation need not be exact for attenuation to improve finite-sample PCA. When $P_*^\perp D P_* \neq 0$, attenuation creates population leakage, and the leading population subspace of $\mathbf{T}_\alpha$ moves away from P_*. Uniformly over a compact interval $\mathcal{I} \subset [0, 1]$ containing zero with $\inf_{\alpha\in\mathcal{I}}\{\lambda_d(\mathbf{T}_\alpha) - \lambda_{d+1}(\mathbf{T}_\alpha)\} > 0$, Theorem C.2 gives

$$\mathbb{E}\mathcal{E}(\widehat{P}_\alpha)=\mathsf{B}_{\mathrm{pop}}(\alpha)+\frac{\mathcal{K}_{\mathrm{move}}(\alpha)}{n}+O(n^{-3/2}), \tag{3.19}$$

where $\mathsf{B}_{\mathrm{pop}}(\alpha)$ is the clean reconstruction cost of moving the population subspace, as defined in (2.9). The term $\mathcal{K}_{\mathrm{move}}(\alpha)/n$ is the signed finite-sample correction around that moved subspace: sampling can rotate the estimator either toward or away from P_*. It reduces to $\mathcal{K}(\alpha)/n$ when the target is preserved. The theoretical risk curve formed by the two displayed terms favors attenuation over PCA when

$$n\mathsf{B}_{\mathrm{pop}}(\alpha)<\mathcal{K}(0)-\mathcal{K}_{\mathrm{move}}(\alpha). \tag{3.20}$$

Thus exact target preservation is unnecessary for a leading-order gain. For the exact risk comparison, the unscaled risk gain must also dominate the $O(n^{-3/2})$ remainder.

In Figure 2(c), increasing population leakage moves the risk-minimizing strength from the direct-only range through the mask-derived range and finally to $\alpha = 0$, which recovers PCA. Across these regimes, the benefit is determined by the sampling leakage remaining after attenuation, the eigengaps that translate this leakage into rotation, and any population cost from moving the target.

4 Numerical Studies

This section presents numerical studies to examine the finite-sample performance of the proposed method. Controlled Gaussian experiments first ask whether the theoretical risk describes the benefit of attenuation as the population target changes from preserved to moving. Fixed fitting-sample-budget studies on three real-data modalities then compare mask-derived and direct attenuation with PCA, correlation PCA, HeteroPCA, and coordinate- and block-masked-target estimators under common budgets. Gaussianity makes the signed risk comparison explicit; the finite-moment theorem and non-Gaussian stress test in the supplement show that the same sampling mechanism is not tied to Gaussian observations.

4.1 Finite-sample risk across preserved and moving targets

Settings and evaluation. We compare a balanced Hadamard-loading design ($p = 16, d = 3$), an unequal-marginal target-preserving design ($p = 4, d = 2$), and a bivariate rotating covariance with eigenvalues 5 and 1. The last design continuously increases population leakage while fixing the spectrum. Panels (a) and (b) average 3,000 independent Gaussian data sets of size $n = 256$ and compare $n\mathbb{E}\mathcal{E}(\widehat{P}_\alpha)$ with the theoretical curve, without fitted parameters. Panel (c) uses 3,000 independent data sets of size $n = 64$ at each of 13 movement levels and compares the empirical and theoretical risk-minimizing strengths on a common attenuation grid. Exact covariances and grids appear in Section G.2.

Results. The theoretical risk curve follows the Monte Carlo path throughout the attenuation interval in both target-preserving designs. The best simulated attenuation lowers the scaled PCA risk by 16.3% in the balanced design and by 63.4% at $\alpha = 0.90$ in the unequal design. Along the moving-target path, the strength minimizing the theoretical risk differs from the Monte Carlo mean-risk minimizer

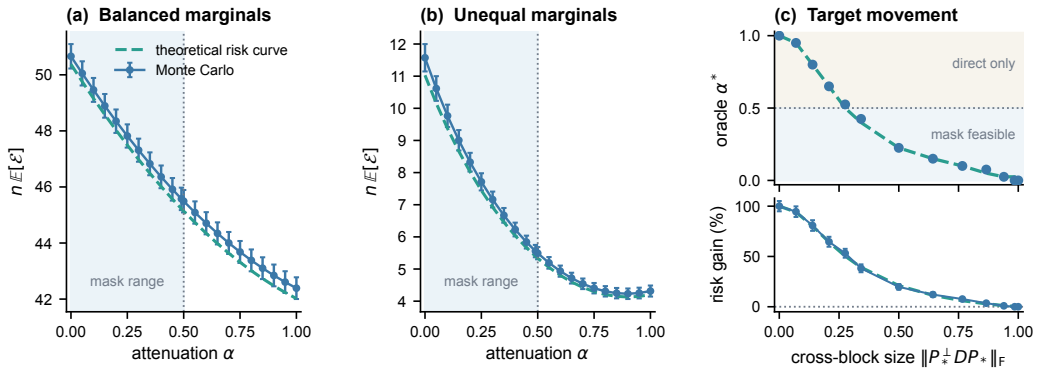

Figure 2: Testing the risk theory from preserved to moving targets. **(a)** Balanced marginal variances and **(b)** unequal marginal variances whose diagonal preserves the PCA target. Points and bars are Monte Carlo means and 95% intervals; dashed curves are the theoretical risk curve from (3.17); the vertical scale is the n-scaled expected clean excess risk. Blue shading marks the mask-derived range $0 \le \alpha < 1/2$, with $\alpha = 1/2$ shown as its limiting endpoint. **(c)** For the bivariate covariance rotated by ϕ with eigenvalues 5 and 1 ((G.5)), the upper axis compares the strengths selected by the theoretical risk curve (line) and Monte Carlo mean risk (points), and the lower axis reports the corresponding risk reductions relative to PCA. The horizontal coordinate is $\|P^{\perp}_{*,\phi} D_\phi P_{*,\phi}\|_{\mathrm{F}}$, where $P_{*,\phi}$ is the leading projector of $\mathbf{\Sigma}_\phi$ and $D_\phi = \mathrm{diag}(\mathbf{\Sigma}_\phi)$. Larger values indicate greater population leakage.

by at most one grid step, 0.025, at all 13 movement levels. As the population coupling grows, the preferred strength decreases from $\alpha = 1$, passes through the mask-derived range, and eventually reaches $\alpha = 0$, recovering classical PCA. The corresponding Monte Carlo risk reduction is nearly 100% under exact target preservation and remains 38.2% when the preferred strength first enters the mask-derived range, before tapering to zero at the PCA endpoint.

4.2 Fixed-budget studies on image patches and ordered signals

Data and fitting budgets. We use Optdigits image patches ($p = 9, d = 2$), Phoneme log-periodograms ($p = 256, d = 8$), and Human Activity Recognition (HAR) acceleration-magnitude series ($p = 128, d = 16$) (Alpaydin & Kaynak, 1998; Hastie et al., 2009; Reyes-Ortiz et al., 2013). All methods receive the same $N \in \{64, 128\}$ observations. Tunable methods share five validation folds, select parameters by reconstruction error, and are refit on all N observations; all methods are evaluated on the official test set. Speakers and subjects remain within one fold; labels are unused. Optdigits uses ten patch-location seeds with 100 training subsamples per seed; the signal studies use 100 subsamples. Full construction details appear in Section G.3.

Methods and evaluation. We compare PCA, mask-derived and direct attenuation, correlation PCA, HeteroPCA, and coordinate- and block-masked-target projectors (two-start numerical fits). Both attenuation paths preserve coordinate scales and include PCA; their candidate grids cover $[0, 0.49]$ and $[0, 1]$, respectively. Correlation PCA standardizes coordinates, HeteroPCA imputes the diagonal, and masked-target methods score hidden coordinates and tune their mask rate including zero. Structured masks use adjacent pixel rows, frequency bins, or time points. Grids and optimization details appear in Sections F.2 and G.3. For replicate b, let $\widehat{R}_{b,\mathrm{test}}$ be a method's test reconstruction mean squared error and let $\widehat{R}^{\mathrm{PCA}}_{b,\mathrm{test}}$ be the test PCA error. We compute $R^{(d)}_{\mathrm{test}}$ by centering the official test observations and summing the smallest $p - d$ eigenvalues of their empirical covariance; equivalently, it is the reconstruction error of the best rank-d affine subspace fitted to the test sample. We report the percentage of PCA's test excess reconstruction risk removed,

$$G_b = 100\, \frac{\widehat{R}^{\mathrm{PCA}}_{b,\mathrm{test}} - \widehat{R}_{b,\mathrm{test}}}{\widehat{R}^{\mathrm{PCA}}_{b,\mathrm{test}} - R^{(d)}_{\mathrm{test}}}. \tag{4.1}$$

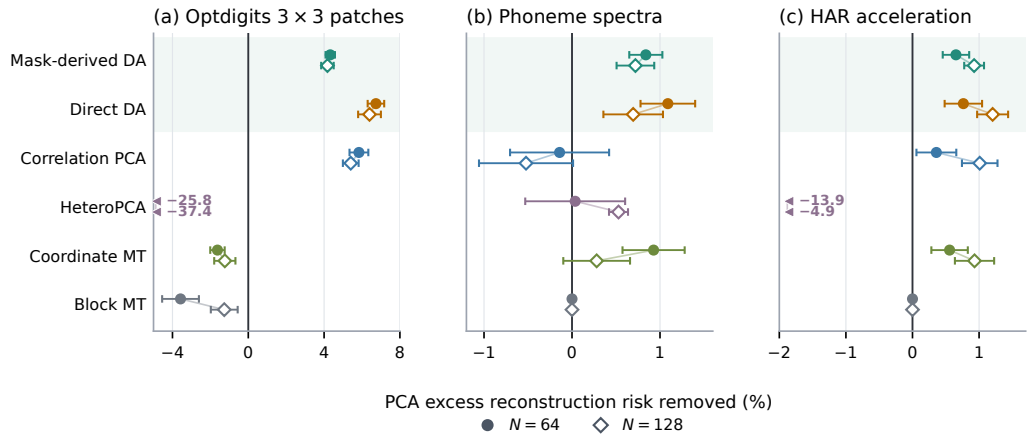

Figure 3: Seven-method fixed fitting-sample-budget comparison across three real-data modalities. Each row gives the mean percentage of PCA excess reconstruction risk removed by one alternative; positive values favor that method. Filled circles and open diamonds correspond to $N = 64$ and 128, respectively, and the shaded rows are the two attenuation paths. DA denotes diagonal attenuation and MT denotes masked-target training. The 95% intervals are conditional on the observed data and official split. Optdigits intervals use ten patch-location means, each averaging 100 training-subsample fits; Phoneme and HAR intervals use 100 training subsamples. Within each fit, all methods use the same sampled observations, validation folds, and official test set. HeteroPCA means beyond the displayed range are marked at the left boundary and labeled by their values.

The test reference in the denominator only expresses the final error reduction on a common scale, which is not used for model selection.

Results. Figure 3 compares PCA with all six alternatives. Both attenuation paths have positive mean gains with intervals above zero in all six data–budget cells. The image experiment produces the largest improvements. At $N = 64$ and 128, mask-derived attenuation removes 4.32% and 4.18% of PCA's test excess reconstruction risk, while direct attenuation removes 6.74% and 6.40%. The direct path selects $\alpha > 1/2$ on 94%–95% of the image-patch fits and 63%–80% of the ordered-signal fits, so the mask-derived boundary is active in every data set. Table 4 in the supplement gives the corresponding numerical comparison. One of the two attenuation paths has the largest mean gain in every data–budget cell: direct attenuation leads in five cells, and mask-derived attenuation leads for Phoneme at $N = 128$. Correlation PCA is competitive on the image patches and HAR, and coordinate masked-target training helps on three of the four ordered-signal cells, but neither matches this consistency across modalities. Exact intervals, selected strengths, and optimization details are reported in Section F.2.

5 Discussion

Input masking reveals a covariance correction that preserves sample cross-covariances while attenuating sample variances. Extending it beyond the mask-derived range gives a directly tunable path from PCA to diagonal deletion. Its benefit comes from reducing sampling-induced rotation, balanced against changing eigengaps and any population target movement. The controlled experiments track this tradeoff, and the three real-data studies demonstrate the value of searching the full path.

The main risk theory is fixed dimensional and Gaussian; the supplement extends the balanced formula under finite-moment conditions and distinguishes masked-target and unconstrained reconstruction objectives. The empirical evidence covers local image, frequency, and temporal coordinates. Structured attenuation beyond a common diagonal strength is a natural direction for future work.

AI USE STATEMENT

Generative-AI tools assisted with literature discovery and synthesis, proof cross-checking, code review, and language editing. The authors verified the cited primary sources, derivations, and computations and take full responsibility for the manuscript.

REFERENCES

Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of PCA and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022.

Ethem Alpaydin and Cihan Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. Dataset, CC BY 4.0.

T. W. Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Rajendra Bhatia. *Matrix Analysis*. Springer, New York, 1997.

Anthony Bisulco, Rahul Ramesh, Randall Balestriero, and Pratik Chaudhari. From linearity to non-linearity: How masked autoencoders capture spatial correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16441–16450, 2025.

Changxiao Cai, Gen Li, Yuejie Chi, H. Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49 (2):944–967, 2021.

Minmin Chen, Zhixiang Eddie Xu, Kilian Q. Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 767–774. Omnipress, 2012.

Minmin Chen, Kilian Weinberger, Fei Sha, and Yoshua Bengio. Marginalized denoising autoencoders for nonlinear representations. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1476–1484. PMLR, 2014.

Edgar Dobriban, William Leeb, and Amit Singer. PCA from noisy, linearly reduced data: The diagonal case. *arXiv preprint arXiv:1611.10333*, 2016.

Ayoub El Hanchi, Murat A. Erdogdu, and Chris J. Maddison. A geometric analysis of PCA. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition, 2009.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *Journal of Machine Learning Research*, 24(330):1–78, 2023.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7918–7928, 2023.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Chi-Heng Lin, Chiraag Kaushik, Eva L. Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

Arnu Pretorius, Steve Kroon, and Herman Kamper. Learning dynamics of linear denoising autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4141–4150. PMLR, 2018.

Markus Reiß and Martin Wahl. Nonasymptotic upper bounds for the reconstruction error of PCA. *The Annals of Statistics*, 48(2):1098–1123, 2020.

Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human activity recognition using smartphones. UCI Machine Learning Repository, 2013. Dataset, CC BY 4.0.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article 32, 2005.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

Stefan Wager, Sida Wang, and Percy S. Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

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

Arie Wortsman Zurich, Federica Gerace, Bruno Loureiro, and Yue M. Lu. A random matrix theory of masked self-supervised regression. *arXiv preprint arXiv:2601.23208*, 2026.

SUPPLEMENTARY MATERIAL

Overview of the supplement. The supplement follows the argument of the main text. Section A proves the exact mask marginalization that constructs the estimator. Section B develops its population geometry, and Section C gives the proofs of the target-preserving, target-moving, and balanced-marginal risk results. Section D extends the balanced theorem beyond Gaussian observations and establishes the validation fallback guarantee. Section E compares the full-output projector with masked-target and unconstrained linear reconstruction objectives. Finally, Section F reports the non-Gaussian stress test, full real-data tables, and implementation details for the empirical comparisons.

Guide to the proofs of the main results. The following list maps each main result to its proof, in main-text order.

(i) Proposition 2.1 is proved in Section A by expanding the mask-averaged reconstruction loss and reducing its optimization to the leading eigenspace of $\mathbf{T}_{n,\alpha}$.

(ii) Proposition 2.2 and the population-bias bound (B.1) are proved in Section B. The first uses the $P_* \oplus P_*^{\perp}$ block decomposition; the second integrates the derivative of the population projector along the attenuation path.

(iii) Lemma G.1 is proved in Section C.4.1. Spectral-projector commutation converts each residual cross-entry into a rotation coordinate, and the common projector expansion converts those coordinates into clean reconstruction risk.

(iv) Theorem 3.1 is proved in Section C.4.2. Gaussian pairwise moments make the variance of each retained-to-omitted residual random coupling a quadratic function of α; their eigengap-weighted sum gives the theoretical risk and its optimal mask-derived and direct strengths.

(v) Theorem 3.2 is proved in Section C.2, using the common clean-risk expansion in Section C.1. The proof converts cross-subspace covariance noise into projector rotation and then into expected clean reconstruction risk.

(vi) The target-moving expansion (3.19) is formalized and proved in Section C.3. It expands around $P_{\mathbf{T}_\alpha}$ and adds the population cost of moving away from P_*.

CONTENTS

A Mask marginalization and estimator construction . . . 13

B Population geometry under attenuation . . . 13

B.1 Exact preservation with unequal marginals . . . 13

B.2 Population bias along the attenuation path . . . 14

C Proofs of the finite-sample risk results . . . 14

C.1 A common clean-risk expansion . . . 14

C.2 Target-preserving expected risk . . . 15

C.3 Total risk around a moving population target . . . 16

C.4 Balanced marginals and the explicit risk curve . . . 18

C.4.1 Pairwise form under balanced marginals . . . 18

C.4.2 From pairwise moments to the risk curve . . . 19

D Robustness and validation selection . . . 20

D.1 Risk expansion beyond Gaussian sampling . . . 20

D.2 Validation selection with a PCA fallback . . . 21

E Alternative masking and reconstruction objectives . . . 22

E.1 The general block masked-target objective . . . 22

E.2 Masked-target random slopes at the PCA target . . . 23

E.3 Linear reconstruction beyond the projector constraint . . . 24

E.4 Additional related context . . . 26

F Additional numerical and real-data studies . . . 27

F.1 Robustness beyond Gaussian observations . . . 27

F.2 Fixed-budget real-data comparisons . . . 27

G Additional details for the conference version . . . 29

G.1 Balanced residual coupling and rotation 29
G.2 Gaussian experiment construction and evaluation 29
G.3 Real-data construction and selection protocol 30
G.4 Numerical comparison across fitting budgets 30

A Mask marginalization and estimator construction

This section proves the exact reduction from masked-input reconstruction to diagonal attenuation used in Proposition 2.1.

Proof of Proposition 2.1. For any symmetric projector P and diagonal mask W,

$$\|x - PWx\|_2^2 = x^\mathsf{T}x - 2x^\mathsf{T}PWx + x^\mathsf{T}WPWx = x^\mathsf{T}x - \mathrm{tr}\left[P\{Wxx^\mathsf{T} + xx^\mathsf{T}W - Wxx^\mathsf{T}W\}\right]. \tag{A.1}$$

The second equality uses the fact that the two scalar cross terms are equal. For independent Bernoulli-q mask coordinates,

$$\mathbb{E}(Wxx^\mathsf{T}) = qxx^\mathsf{T},$$
$$\mathbb{E}(Wxx^\mathsf{T}W) = q^2xx^\mathsf{T} + q(1-q)\,\mathrm{diag}(xx^\mathsf{T}).$$

Taking the expectation in (A.1) therefore produces

$$2qxx^\mathsf{T} - q^2xx^\mathsf{T} - q(1-q)\,\mathrm{diag}(xx^\mathsf{T}) = q(2-q)\left\{xx^\mathsf{T} - \frac{1-q}{2-q}\,\mathrm{diag}(xx^\mathsf{T})\right\}.$$

The positive prefactor $q(2-q)$ does not affect the optimizer for $q > 0$. After averaging over observations, minimizing the loss is equivalent to maximizing $\mathrm{tr}(P\mathbf{T}_{n,\alpha})$ over $P \in \mathcal{P}_d$. Ky Fan's variational principle (Bhatia, 1997) gives a leading rank-d projector of $\mathbf{T}_{n,\alpha}$, proving Proposition 2.1 and the mapping $\alpha = (1-q)/(2-q) = m/(1+m)$. □

B Population geometry under attenuation

Diagonal attenuation can preserve the population PCA target exactly or move it along the path $\mathbf{T}_\alpha = \mathbf{\Sigma} - \alpha \cdot \mathrm{diag}(\mathbf{\Sigma})$. This section first characterizes exact preservation, including the unequal-marginal example from the main text, and then quantifies the population cost when preservation is only approximate.

B.1 Exact preservation with unequal marginals

Balanced marginal variances are sufficient but not necessary for exact target preservation. To verify the unequal design in (G.4), note that its diagonal is $D = \mathrm{diag}(4, 4, 2.5, 2.5)$. Let $u_1 = (1, 1, 0, 0)^\mathsf{T}/\sqrt{2}$ and $u_2 = (0, 0, 1, 1)^\mathsf{T}/\sqrt{2}$. The leading population subspace is $\mathrm{span}\{u_1, u_2\}$, and D preserves this subspace and its complement. Direct calculation gives $\inf_{0\le\alpha\le1} \delta_\alpha = 7/2 - 3/\sqrt{2} > 1.37$, so the PCA target is retained along the full attenuation path.

Proof of Proposition 2.2. If $P_*DP_*^\perp = 0$, symmetry of D implies that both $\mathrm{range}(P_*)$ and its orthogonal complement are invariant under D. They are already invariant under $\mathbf{\Sigma}$, so they are invariant under $\mathbf{T}_\alpha = \mathbf{\Sigma} - \alpha D$. This matrix is block diagonal in the decomposition. The condition $\delta_\alpha > 0$ places every eigenvalue of the target block above every eigenvalue of the residual block. Its leading rank-d projector is therefore uniquely P_*. Conversely, if P_* is the unique leading projector of $\mathbf{T}_\alpha$, it is invariant under both $\mathbf{T}_\alpha$ and $\mathbf{\Sigma}$. Since $\alpha > 0$, subtracting their cross-block equations gives $P_*DP_*^\perp = 0$; uniqueness gives $\delta_\alpha > 0$. □

B.2 Population bias along the attenuation path

To control the population cost, follow $\mathbf{T}_t = \mathbf{\Sigma} - tD$, which begins at $P_{\mathbf{T}_0} = P_*$. If the cutoff eigengap remains positive along this path, define

$$\underline{g}_\alpha := \inf_{0\le t\le\alpha}\{\lambda_d(\mathbf{T}_t) - \lambda_{d+1}(\mathbf{T}_t)\} > 0, \qquad \mathsf{B}_{\mathrm{pop}}(\alpha) \le \frac{\alpha^2}{\underline{g}_\alpha}\sup_{0\le t\le\alpha}\|P^{\perp}_{\mathbf{T}_t} D P_{\mathbf{T}_t}\|_{\mathrm{F}}^2. \tag{B.1}$$

Thus population movement remains small when the diagonal creates little cross-subspace coupling along the path and the cutoff eigengap stays separated.

Proof of the population-bias identity and (B.1). We follow the attenuation path directly. Write $P_t = P_{\mathbf{T}_t}$ and $B(t) = \mathrm{tr}\{\mathbf{\Sigma}(P_* - P_t)\}$. The positive cutoff gap makes the cluster projector P_t differentiable, even if eigenvalues repeat within the leading or residual block. At each t, choose orthonormal eigenbases $\mathbf{T}_t v_j(t) = \mu_j(t) v_j(t)$ for the two blocks. The standard spectral-projector derivative, which is basis invariant within repeated eigenspaces, gives

$$P_t' = -\sum_{j\le d<k}\frac{v_k(t)^{\mathsf{T}} D v_j(t)}{\mu_j(t) - \mu_k(t)}\{v_k(t)v_j(t)^{\mathsf{T}} + v_j(t)v_k(t)^{\mathsf{T}}\}.$$

Because P_t commutes with $\mathbf{T}_t$, $\mathrm{tr}(\mathbf{T}_t P_t') = 0$. Using $\mathbf{\Sigma} = \mathbf{T}_t + tD$ therefore yields

$$B'(t) = -\mathrm{tr}(\mathbf{\Sigma} P_t') = -t\,\mathrm{tr}(D P_t') = 2t\sum_{j\le d<k}\frac{\{v_k(t)^{\mathsf{T}} D v_j(t)\}^2}{\mu_j(t) - \mu_k(t)}.$$

Since $B(0) = 0$, integration gives the exact identity

$$\mathsf{B}_{\mathrm{pop}}(\alpha) = 2\int_0^\alpha t\sum_{j\le d<k}\frac{\{v_k(t)^{\mathsf{T}} D v_j(t)\}^2}{\mu_j(t) - \mu_k(t)}\,dt. \tag{B.2}$$

Every denominator is at least $\underline{g}_\alpha$, while the sum of the squared numerators is $\|(I_p - P_t)DP_t\|_{\mathrm{F}}^2$. Bounding the integrand by its supremum and using $2\int_0^\alpha t\,dt = \alpha^2$ proves (B.1). In particular, if $\mathbf{\Sigma} u_k = \lambda_k u_k$, continuity of the integrand at zero gives

$$\mathsf{B}_{\mathrm{pop}}(\alpha) = \alpha^2\sum_{j\le d<k}\frac{(u_k^{\mathsf{T}} D u_j)^2}{\lambda_j - \lambda_k} + O(\alpha^3).$$

This local expression shows explicitly why distant residual directions do not require a separate tail penalty: their larger eigengaps reduce their rotation. □

The next section uses this population decomposition in the expected-risk theory: exact preservation removes the first term, while the target-moving result retains it and quantifies the remaining sampling contribution around $P_{\mathbf{T}_\alpha}$.

C Proofs of the finite-sample risk results

The main text first uses individual retained-to-discarded rotations to build intuition and then treats the full subspace. Here we provide a common projector perturbation argument, which justifies both the target-preserving and target-moving risk theorems. We then specialize the target-preserving coefficient to balanced marginal variances, where the sampling-leakage calculation becomes explicit.

C.1 A common clean-risk expansion

We first establish a two-matrix projector expansion that applies whether the population target is preserved or moves.

Lemma C.1 (Clean risk around a possibly moving target)**.** *Let A and B be symmetric matrices, let $P = P_A$, and suppose that the cutoff gap of A is at least $\gamma > 0$. Let $P^\perp = I_p - P$ denote the complementary projector and set $\mathbb{X} = \{X : X = P^\perp X P\}$. On $\mathbb{X}$, define*

$$\mathcal{L}_A(X) = X(PAP) - (P^\perp A P^\perp)X, \qquad \mathcal{H}_B(X) = X(PBP) - (P^\perp B P^\perp)X. \tag{C.1}$$

For symmetric F, define

$$Y_{1,F} = \mathcal{L}_A^{-1}(P^\perp F P), \tag{C.2}$$

$$Y_{2,F} = \mathcal{L}_A^{-1}\{P^\perp F P^\perp Y_{1,F} - Y_{1,F} P F P\}. \tag{C.3}$$

There is a finite constant $C_{A,B,d}$ such that every symmetric F with $\|F\|_{\mathrm{op}} \le \gamma/4$ satisfies

$$\begin{aligned}\mathrm{tr}\{B(P - P_{A+F})\} &= -2\langle P^\perp B P, Y_{1,F}\rangle_{\mathrm{F}} + \langle Y_{1,F}, \mathcal{H}_B(Y_{1,F})\rangle_{\mathrm{F}} \\ &\quad - 2\langle P^\perp B P, Y_{2,F}\rangle_{\mathrm{F}} + R^{\mathrm{mov}}_{A,B}(F),\end{aligned} \tag{C.4}$$

$$|R^{\mathrm{mov}}_{A,B}(F)| \le C_{A,B,d}\|F\|^3_{\mathrm{op}}. \tag{C.5}$$

If $P^\perp B P = 0$, this reduces to

$$\mathrm{tr}\{B(P - P_{A+F})\} = \langle X_F, \mathcal{H}_B(X_F)\rangle_{\mathrm{F}} + R_{A,B}(F), \tag{C.6}$$

$$X_F = Y_{1,F}, \qquad |R_{A,B}(F)| \le C_{A,B,d}\|F\|^3_{\mathrm{op}}. \tag{C.7}$$

The constant can be chosen uniformly over a compact family of pairs (A, B) whose norms are bounded and whose relevant cutoff gaps are bounded below.

Proof of Lemma C.1. The gap makes $\mathcal{L}_A$ invertible on $\mathbb{X}$, with inverse norm at most $1/\gamma$. Weyl's inequality also keeps the leading cluster of $A + F$ separated. Its range is therefore the graph of a unique map $Y : \mathrm{range}(P) \to \mathrm{range}(P^\perp)$ with $\|Y\|_{\mathrm{op}} = O(\|F\|_{\mathrm{op}}/\gamma)$. The block eigenvalue equation for this graph is the Riccati equation

$$\mathcal{L}_A(Y) = P^\perp F P + P^\perp F P^\perp Y - Y P F P - Y P F P^\perp Y. \tag{C.8}$$

Applying $\mathcal{L}_A^{-1}$, substituting the first-order approximation into the two quadratic terms, and using the preceding bound gives

$$Y = Y_{1,F} + Y_{2,F} + O(\|F\|^3_{\mathrm{op}}). \tag{C.9}$$

All constants here depend only on the fixed matrices and the gap.

Relative to $\mathrm{range}(P) \oplus \mathrm{range}(P^\perp)$, the projector onto the graph of Y has blocks

$$P_{A+F} = \begin{pmatrix} (I + Y^\mathsf{T} Y)^{-1} & (I + Y^\mathsf{T} Y)^{-1} Y^\mathsf{T} \\ Y(I + Y^\mathsf{T} Y)^{-1} & Y(I + Y^\mathsf{T} Y)^{-1} Y^\mathsf{T} \end{pmatrix}. \tag{C.10}$$

Expanding these blocks to second order and using (C.9) gives

$$\mathrm{tr}\{B(P - P_{A+F})\} = -2\langle P^\perp B P, Y_{1,F} + Y_{2,F}\rangle_{\mathrm{F}} + \langle Y_{1,F}, \mathcal{H}_B(Y_{1,F})\rangle_{\mathrm{F}} + O(\|F\|^3_{\mathrm{op}}),$$

This proves (C.4). When $P^\perp B P = 0$, the two off-diagonal terms vanish and (C.6) follows. The graph, Riccati, and inverse bounds are uniform over the compact family stated in the lemma. □

C.2 Target-preserving expected risk

We now specialize the common expansion to attenuation paths that preserve the population PCA target.

Proof of Theorem 3.2. For $\alpha \in \mathcal{I}$, put

$$A_\alpha = \mathbf{T}_\alpha, \qquad F_{n,\alpha} = E_n - \alpha\,\mathrm{diag}(E_n), \qquad E_n = \mathbf{S}_n - \mathbf{\Sigma}. \tag{C.11}$$

By Assumption 3.2, A_α and $\mathbf{\Sigma}$ have the same leading projector $P = P_*$, and the cutoff gap of A_α is at least $\underline{\delta}_{\mathcal{I}}$. The closure of $\{A_\alpha : \alpha \in \mathcal{I}\}$ is compact; continuity of the block gap and the strictly positive infimum preserve the same projector and gap bound on this closure. Hence Lemma C.1 applies with one uniform remainder constant whenever $\|F_{n,\alpha}\|_{\mathrm{op}} \le \underline{\delta}_{\mathcal{I}}/4$.

The linear rotation term in that lemma is

$$X_{n,\alpha} = \mathcal{L}_\alpha^{-1}(P^\perp F_{n,\alpha} P) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\alpha^{-1} Z_\alpha(\mathbf{x}_i), \tag{C.12}$$

where the summands are independent and centered. Cross terms therefore vanish in expectation, giving the exact identity

$$n\mathbb{E}\,\langle X_{n,\alpha}, \mathcal{H}_{\mathbf{\Sigma}}(X_{n,\alpha})\rangle_{\mathrm{F}} = \mathcal{K}(\alpha). \tag{C.13}$$

For fixed p, Gaussian covariance concentration gives $\mathbb{E}\|E_n\|_{\mathrm{op}}^r = O(n^{-r/2})$ for every fixed r. Since $\|\mathrm{diag}(E_n)\|_{\mathrm{op}} \le \|E_n\|_{\mathrm{op}}$ and $0 \le \alpha \le 1$,

$$\sup_{\alpha\in\mathcal{I}} \|F_{n,\alpha}\|_{\mathrm{op}} \le 2\|E_n\|_{\mathrm{op}}. \tag{C.14}$$

On the event $\|E_n\|_{\mathrm{op}} \le \underline{\delta}_{\mathcal{I}}/8$, the cubic remainder in Lemma C.1 thus has expectation $O(n^{-3/2})$, uniformly in α. The complement has exponentially small probability. Both the clean excess risk and the quadratic leading term have finite Gaussian moments, so Cauchy–Schwarz makes their contribution on that complement exponentially smaller than $n^{-3/2}$. Combining these facts with (C.13) gives

$$\sup_{\alpha\in\mathcal{I}} \left| \mathbb{E}\mathcal{E}(\widehat{P}_\alpha) - \frac{\mathcal{K}(\alpha)}{n} \right| = O(n^{-3/2}).$$

Multiplying by n and enlarging the constant to cover finitely many initial sample sizes proves (3.18). □

C.3 Total risk around a moving population target

We now remove target preservation. Let $P_\alpha = P_{\mathbf{T}_\alpha}$ and suppose that the cutoff gap of $\mathbf{T}_\alpha$ remains positive over a compact interval $\mathcal{I} \subset [0,1]$ containing zero:

$$g_{\mathcal{I}} = \inf_{\alpha\in\mathcal{I}} \{\lambda_d(\mathbf{T}_\alpha) - \lambda_{d+1}(\mathbf{T}_\alpha)\} > 0. \tag{C.15}$$

The theorem separates the deterministic cost of moving from P_* from the finite-sample reconstruction effect of sampling-induced rotation around P_α. The centered first-order sampling rotation has mean zero. At second order, both its variance and its average displacement matter because clean reconstruction risk is measured relative to P_* rather than P_α. For one observation, write the centered attenuated fluctuation as

$$\Xi_\alpha(x) = xx^{\mathsf{T}} - \mathbf{\Sigma} - \alpha\,\mathrm{diag}(xx^{\mathsf{T}} - \mathbf{\Sigma}). \tag{C.16}$$

On the cross-subspace maps $\mathbb{X}_\alpha = \{X : X = P_\alpha^\perp X P_\alpha\}$, define

$$\mathcal{L}_\alpha^{\mathrm{mov}}(X) = X(P_\alpha \mathbf{T}_\alpha P_\alpha) - (P_\alpha^\perp \mathbf{T}_\alpha P_\alpha^\perp)X, \tag{C.17}$$

$$\mathcal{H}_{\mathbf{\Sigma},\alpha}(X) = X(P_\alpha \mathbf{\Sigma} P_\alpha) - (P_\alpha^\perp \mathbf{\Sigma} P_\alpha^\perp)X. \tag{C.18}$$

The first two graph perturbations generated by $\Xi_\alpha(x)$ are

$$Y_{1,\alpha}(x) = \{\mathcal{L}_\alpha^{\mathrm{mov}}\}^{-1}\{P_\alpha^\perp \Xi_\alpha(x) P_\alpha\}, \tag{C.19}$$

$$Y_{2,\alpha}(x) = \{\mathcal{L}_\alpha^{\mathrm{mov}}\}^{-1}\left\{P_\alpha^\perp \Xi_\alpha(x) P_\alpha^\perp Y_{1,\alpha}(x) - Y_{1,\alpha}(x) P_\alpha \Xi_\alpha(x) P_\alpha\right\}. \tag{C.20}$$

Define the signed sampling coefficient

$$\mathcal{K}_{\mathrm{move}}(\alpha) = \mathbb{E}\left[\langle Y_{1,\alpha}, \mathcal{H}_{\mathbf{\Sigma},\alpha}(Y_{1,\alpha})\rangle_{\mathrm{F}} - 2\left\langle P_\alpha^\perp \mathbf{\Sigma} P_\alpha, Y_{2,\alpha}\right\rangle_{\mathrm{F}}\right]. \tag{C.21}$$

The second term is absent under target preservation. When the population target moves, it records whether the average second-order sampling displacement points toward or away from the clean PCA subspace; consequently, $\mathcal{K}_{\mathrm{move}}(\alpha)$ need not be nonnegative by itself.

Theorem C.2 (Total risk with population target movement)**.** *Under Assumptions 2.1 and 2.2 and (C.15), there are finite constants C and N such that*

$$\sup_{\alpha\in\mathcal{I}} \left| \mathbb{E}\mathcal{E}(\widehat{P}_\alpha) - \mathsf{B}_{\mathrm{pop}}(\alpha) - \frac{\mathcal{K}_{\mathrm{move}}(\alpha)}{n} \right| \le \frac{C}{n^{3/2}}, \qquad n \ge N. \tag{C.22}$$

Proof of Theorem C.2. For $\alpha \in \mathcal{I}$, write

$$A_\alpha = \mathbf{T}_\alpha, \qquad F_{n,\alpha} = \frac{1}{n}\sum_{i=1}^{n} \Xi_\alpha(\mathbf{x}_i), \qquad P_\alpha = P_{A_\alpha}. \tag{C.23}$$

The excess risk has the exact decomposition

$$\mathcal{E}(\widehat{P}_\alpha) = \mathsf{B}_{\mathrm{pop}}(\alpha) + \mathrm{tr}\{\mathbf{\Sigma}(P_\alpha - P_{A_\alpha + F_{n,\alpha}})\}. \tag{C.24}$$

Apply Lemma C.1 to the second term with $A = A_\alpha$, $B = \mathbf{\Sigma}$, and $P = P_\alpha$. Let $Y_{1,n,\alpha}$ and $Y_{2,n,\alpha}$ denote the two graph terms in Equations (C.2) and (C.3). Linearity gives

$$Y_{1,n,\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_{1,\alpha}(\mathbf{x}_i), \qquad \mathbb{E}Y_{1,n,\alpha} = 0. \tag{C.25}$$

The quadratic graph term expands into a double sum. Independence and centering remove every summand with distinct observation indices, leaving

$$\mathbb{E}Y_{2,n,\alpha} = \frac{1}{n}\mathbb{E}Y_{2,\alpha}(\mathbf{x}). \tag{C.26}$$

The same argument applied to the clean quadratic term gives

$$\mathbb{E}\left\langle Y_{1,n,\alpha}, \mathcal{H}_{\mathbf{\Sigma},\alpha}(Y_{1,n,\alpha})\right\rangle_{\mathrm{F}} = \frac{1}{n}\mathbb{E}\left\langle Y_{1,\alpha}, \mathcal{H}_{\mathbf{\Sigma},\alpha}(Y_{1,\alpha})\right\rangle_{\mathrm{F}}. \tag{C.27}$$

The expectation of the linear graph term is zero. Combining Equations (C.26) and (C.27) with (C.4) therefore identifies the complete second-order expectation as $\mathcal{K}_{\mathrm{move}}(\alpha)/n$.

It remains to control the remainder uniformly. The gap condition (C.15) and compactness of $\{A_\alpha : \alpha \in \mathcal{I}\}$ give one graph-expansion constant for the whole interval. Moreover,

$$\sup_{\alpha\in\mathcal{I}} \|F_{n,\alpha}\|_{\mathrm{op}} \le 2\|\mathbf{S}_n - \mathbf{\Sigma}\|_{\mathrm{op}}. \tag{C.28}$$

The Gaussian moment and complement-event argument used in (C.14) consequently bounds the expected cubic remainder by $O(n^{-3/2})$, uniformly in α. Substitution into (C.24) proves (C.22). When $P_\alpha = P_*$, we have $P_\alpha^\perp \mathbf{\Sigma} P_\alpha = 0$; the second graph contribution vanishes and $\mathcal{K}_{\mathrm{move}}(\alpha) = \mathcal{K}(\alpha)$, recovering Theorem 3.2. □

The theorem gives a direct comparison with PCA:

$$\mathbb{E}\mathcal{E}(\widehat{P}_\alpha) - \mathbb{E}\mathcal{E}(\widehat{P}_0) = \mathsf{B}_{\mathrm{pop}}(\alpha) + \frac{\mathcal{K}_{\mathrm{move}}(\alpha) - \mathcal{K}(0)}{n} + O(n^{-3/2}). \tag{C.29}$$

Thus a fixed strength is beneficial to first order at sample size n when

$$n\mathsf{B}_{\mathrm{pop}}(\alpha) < \mathcal{K}(0) - \mathcal{K}_{\mathrm{move}}(\alpha). \tag{C.30}$$

If $\mathsf{B}_{\mathrm{pop}}(\alpha) > 0$ for a fixed α, this inequality eventually fails as n grows. More generally, if $\mathsf{B}_{\mathrm{pop}}$ has the unique zero $\alpha = 0$ on $\mathcal{I}$, every oracle minimizer of the expected risk converges to zero. This is the regime in which attenuation is a genuinely finite-sample correction rather than a fixed population transformation.

The small-α form makes the return to PCA more explicit. With $\mathbf{\Sigma}u_j = \lambda_j u_j$, the population term satisfies

$$\mathsf{B}_{\mathrm{pop}}(\alpha) = \alpha^2 \mathsf{C}_D + O(\alpha^3), \qquad \mathsf{C}_D = \sum_{j\le d<k} \frac{(u_k^\mathsf{T} D u_j)^2}{\lambda_j - \lambda_k}. \tag{C.31}$$

If $\mathsf{C}_D > 0$ and $\mathsf{R}_{\mathrm{move}} := -\mathcal{K}'_{\mathrm{move}}(0) > 0$, then the leading approximation in (C.29) is minimized locally at

$$\widetilde{\alpha}_n = \frac{\mathsf{R}_{\mathrm{move}}}{2\mathsf{C}_D\, n} + o(n^{-1}). \tag{C.32}$$

Equation (C.32) concerns the minimizer of the displayed leading approximation; the exact oracle convergence to zero follows directly from the uniform expansion above.

We next return to balanced marginals, where the target-preserving risk coefficient can be evaluated explicitly.

C.4 Balanced marginals and the explicit risk curve

We now make the target-preserving risk coefficient explicit under balanced marginal variances.

C.4.1 Pairwise form under balanced marginals

Under balanced marginals, recall the pairwise variables $a_{jk}(\mathbf{x})$ and $b_{jk}(\mathbf{x})$ from (3.3). To lighten notation in this subsection, write $a_{jk} = a_{jk}(\mathbf{x})$ and $b_{jk} = b_{jk}(\mathbf{x})$. Attenuation replaces the sample average of a_{jk} by that of $a_{jk} - \alpha b_{jk}$. For the moment calculations below, put

$$w_{jk} = u_j \circ u_k, \qquad s_{jk} = \|w_{jk}\|_2^2, \qquad h_{jk} = 2w_{jk}^{\mathsf{T}}(\boldsymbol{\Sigma} \circ \boldsymbol{\Sigma})w_{jk}. \tag{C.33}$$

Here s_{jk} measures how strongly the two population directions load on the same observed coordinates. Isserlis' Gaussian fourth-moment identity (Isserlis, 1918) gives

$$\operatorname{var}(a_{jk}) = \lambda_j\lambda_k, \quad \operatorname{Cov}(a_{jk}, b_{jk}) = 2\lambda_j\lambda_k s_{jk}, \quad \operatorname{var}(b_{jk}) = h_{jk}. \tag{C.34}$$

Proof of Lemma G.1. Fix $\alpha \in [0,1]$ and write $F_{n,\alpha} = \mathbf{T}_{n,\alpha} - \mathbf{T}_\alpha$ and $\Delta P_\alpha = \widehat{P}_\alpha - P_*$. Gaussian covariance concentration and the fixed cutoff gap, together with the Davis–Kahan theorem, give

$$\|F_{n,\alpha}\|_{\mathrm{op}} = O_{\mathbb{P}}(n^{-1/2}), \qquad \|\Delta P_\alpha\|_{\mathrm{op}} = O_{\mathbb{P}}(n^{-1/2}). \tag{C.35}$$

Because $\widehat{P}_\alpha$ is a spectral projector of $\mathbf{T}_{n,\alpha}$, the two matrices commute. Substituting $\mathbf{T}_{n,\alpha} = \mathbf{T}_\alpha + F_{n,\alpha}$ and $\widehat{P}_\alpha = P_* + \Delta P_\alpha$, and using $\mathbf{T}_\alpha P_* = P_* \mathbf{T}_\alpha$, yields

$$\begin{aligned} 0 &= \mathbf{T}_{n,\alpha}\widehat{P}_\alpha - \widehat{P}_\alpha \mathbf{T}_{n,\alpha} \\ &= (\mathbf{T}_\alpha P_* - P_* \mathbf{T}_\alpha) + (\mathbf{T}_\alpha \Delta P_\alpha - \Delta P_\alpha \mathbf{T}_\alpha) \\ &\quad + (F_{n,\alpha}P_* - P_* F_{n,\alpha}) + (F_{n,\alpha}\Delta P_\alpha - \Delta P_\alpha F_{n,\alpha}) \\ &= \mathbf{T}_\alpha \Delta P_\alpha - \Delta P_\alpha \mathbf{T}_\alpha + F_{n,\alpha}P_* - P_* F_{n,\alpha} + O_{\mathbb{P}}(n^{-1}). \end{aligned} \tag{C.36}$$

The first parenthesis vanishes because P_* is a spectral projector of $\mathbf{T}_\alpha$; the last is $O_{\mathbb{P}}(n^{-1})$ by (C.35). Under balance, $\mathbf{T}_\alpha u_r = (\lambda_r - \alpha\tau)u_r$. Hence

$$\begin{aligned} u_k^{\mathsf{T}}(\mathbf{T}_\alpha \Delta P_\alpha - \Delta P_\alpha \mathbf{T}_\alpha)u_j &= -(\lambda_j - \lambda_k)u_k^{\mathsf{T}} \Delta P_\alpha u_j, \\ u_k^{\mathsf{T}}(F_{n,\alpha}P_* - P_* F_{n,\alpha})u_j &= u_k^{\mathsf{T}} F_{n,\alpha} u_j, \end{aligned}$$

where the second equality uses $P_* u_j = u_j$ and $u_k^{\mathsf{T}} P_* = 0$. Premultiplying (C.36) by u_k^{T} and postmultiplying by u_j therefore gives

$$0 = -(\lambda_j - \lambda_k)u_k^{\mathsf{T}} \Delta P_\alpha u_j + u_k^{\mathsf{T}} F_{n,\alpha} u_j + O_{\mathbb{P}}(n^{-1}).$$

Since $u_k^{\mathsf{T}} F_{n,\alpha} u_j = \overline{a}_{jk} - \alpha \overline{b}_{jk}$, solving the preceding identity and using (G.1), then summing over the finitely many pairs $j \le d < k$, proves (G.2); the matrix remainder is $O_{\mathbb{P}}(n^{-1})$ in Frobenius norm because p is fixed.

Finally, adding the scalar matrix $\alpha\tau I_p$ does not change spectral projectors. Hence $\widehat{P}_\alpha$ is also the leading projector of $\boldsymbol{\Sigma} + F_{n,\alpha}$. We may therefore apply Lemma C.1 with $A = B = \boldsymbol{\Sigma}$ and $P = P_*$. Because $P_*^{\perp} \boldsymbol{\Sigma} P_* = 0$, the simplified conclusion (C.6) applies. To evaluate its quadratic term, expand the cross-block perturbation in the population eigenbasis:

$$P_*^{\perp} F_{n,\alpha} P_* = \sum_{j \le d < k} (u_k^{\mathsf{T}} F_{n,\alpha} u_j) u_k u_j^{\mathsf{T}}.$$

For each basis matrix $u_k u_j^{\mathsf{T}}$, the operator in (C.1) satisfies

$$\mathcal{L}_{\boldsymbol{\Sigma}}(u_k u_j^{\mathsf{T}}) = \mathcal{H}_{\boldsymbol{\Sigma}}(u_k u_j^{\mathsf{T}}) = (\lambda_j - \lambda_k) u_k u_j^{\mathsf{T}}.$$

Consequently, writing $f_{kj} = u_k^{\mathsf{T}} F_{n,\alpha} u_j$,

$$X_{F_{n,\alpha}} = \mathcal{L}_{\boldsymbol{\Sigma}}^{-1}(P_*^{\perp} F_{n,\alpha} P_*) = \sum_{j \le d < k} \frac{f_{kj}}{\lambda_j - \lambda_k} u_k u_j^{\mathsf{T}},$$

$$\left\langle X_{F_{n,\alpha}}, \mathcal{H}_{\boldsymbol{\Sigma}}(X_{F_{n,\alpha}}) \right\rangle_{\mathrm{F}} = \sum_{j \le d < k} \frac{f_{kj}^2}{\lambda_j - \lambda_k},$$

where the last identity uses the Frobenius orthonormality of $\{u_k u_j^\mathsf{T} : j \le d < k\}$. Thus (C.6) gives

$$\mathcal{E}(\widehat{P}_\alpha) = \sum_{j \le d < k} \frac{(u_k^\mathsf{T} F_{n,\alpha} u_j)^2}{\lambda_j - \lambda_k} + O_{\mathbb{P}}(\|F_{n,\alpha}\|_{\mathrm{op}}^3).$$

Using (C.35) and the same cross-entry identity proves (G.3). □

The moment identities above show how attenuation changes each retained-to-omitted coupling before these contributions are aggregated across the rank cutoff.

Aggregating across the rank cutoff gives

$$\mathsf{A}_{\mathbf{\Sigma}} = \sum_{j \le d < k} \frac{\lambda_j \lambda_k s_{jk}}{\lambda_j - \lambda_k}, \qquad \mathsf{B}_{\mathbf{\Sigma}} = \sum_{j \le d < k} \frac{h_{jk}}{\lambda_j - \lambda_k}, \tag{C.37}$$

$$G(\alpha) = \sum_{j \le d < k} \frac{\lambda_j \lambda_k - 4\alpha \lambda_j \lambda_k s_{jk} + \alpha^2 h_{jk}}{\lambda_j - \lambda_k}. \tag{C.38}$$

Consequently, $\mathcal{K}(\alpha) = G(\alpha)$, $\mathsf{C}_{\mathbf{\Sigma}} = 2\mathsf{A}_{\mathbf{\Sigma}}$, and $\mathsf{V}_{\mathbf{\Sigma}} = \mathsf{B}_{\mathbf{\Sigma}}$. These identities connect the readable main-text tradeoff to the pairwise quantities used in the proofs below.

C.4.2 From pairwise moments to the risk curve

Proof of Theorem 3.1. Under balance,

$$\mathbf{T}_{n,\alpha} = \mathbf{\Sigma} - \alpha\tau I_p + E_{n,\alpha}, \qquad E_{n,\alpha} = E_n - \alpha \operatorname{diag}(E_n), \qquad E_n = \mathbf{S}_n - \mathbf{\Sigma}.$$

The scalar shift preserves all population eigenvectors and gaps. Equal-rank projectors have equal trace, so replacing $\mathbf{\Sigma}$ by $\mathbf{\Sigma} - \alpha\tau I_p$ also leaves the clean excess risk unchanged. After removing this shift, apply Lemma C.1 with $A = B = \mathbf{\Sigma}$ and $F = E_{n,\alpha}$. In the eigenbasis of $\mathbf{\Sigma}$, its quadratic term is the sum below. Thus, on $\|E_{n,\alpha}\|_{\mathrm{op}} \le \delta/4$,

$$\mathcal{E}(\widehat{P}_\alpha) = \sum_{j \le d < k} \frac{(u_j^\mathsf{T} E_{n,\alpha} u_k)^2}{\lambda_j - \lambda_k} + \mathrm{Rem}_{n,\alpha}, \qquad |\mathrm{Rem}_{n,\alpha}| \le C_{\mathrm{rem}} \|E_{n,\alpha}\|_{\mathrm{op}}^3, \tag{C.39}$$

Here C_{rem} depends only on the fixed spectrum and cutoff gap. The formula is invariant to basis rotations inside a repeated-eigenvalue block, because the corresponding squared cross-block entries share the same denominator.

Because $u_j^\mathsf{T} E_{n,\alpha} u_k$ is the sample average of $a_{jk} - \alpha b_{jk}$, the moment identities in (C.34) give

$$n\mathbb{E}(u_j^\mathsf{T} E_{n,\alpha} u_k)^2 = \mathrm{var}(a_{jk} - \alpha b_{jk}) = \lambda_j \lambda_k - 4\alpha \lambda_j \lambda_k s_{jk} + \alpha^2 h_{jk}. \tag{C.40}$$

Equation (C.40) gives the expectation of the quadratic term in (C.39) as $G(\alpha)/n$. The uniform good-event and complement argument in the proof of Theorem 3.2 applies unchanged because $\sup_{0 \le \alpha \le 1} \|E_{n,\alpha}\|_{\mathrm{op}} \le 2\|E_n\|_{\mathrm{op}}$. Consequently,

$$\sup_{0 \le \alpha \le 1} \left| \mathbb{E}\mathcal{E}(\widehat{P}_\alpha) - \frac{G(\alpha)}{n} \right| = O(n^{-3/2}),$$

which proves (3.8) after enlarging its constant for finitely many initial sample sizes. For a fixed α with $\Delta_\alpha = G(0) - G(\alpha) > 0$,

$$n\{\mathbb{E}\mathcal{E}(\widehat{P}_0) - \mathbb{E}\mathcal{E}(\widehat{P}_\alpha)\} \ge \Delta_\alpha - \frac{2C}{\sqrt{n}},$$

which yields an eventual finite-sample improvement whenever the theoretical risk curve is smaller than at PCA. Summing (C.40) over cross-boundary pairs gives (C.38). These perturbation and concentration steps are consistent with classical and nonasymptotic PCA analyses (Anderson, 1963; Koltchinskii & Lounici, 2017; Reiß & Wahl, 2020).

It remains to verify strict positivity and optimize the quadratic. If every s_{jk} across the cutoff were zero, then every observed coordinate would lie entirely in either the selected or rejected eigenspace.

Its variance would consequently be at least λ_d in the first case and at most λ_{d+1} in the second. Both kinds of coordinates must occur, which contradicts their common variance and the positive cutoff gap. Hence $\mathsf{A}_{\mathbf{\Sigma}} > 0$. For a pair with $s_{jk} > 0$, $w_{jk} \neq 0$; positive definiteness of $\mathbf{\Sigma} \circ \mathbf{\Sigma}$ then gives $h_{jk} > 0$, so $\mathsf{B}_{\mathbf{\Sigma}} > 0$. Expanding $G(\alpha) - G(0)$ proves (3.9). This completes the proof of Theorem 3.1. Differentiating the same strictly convex quadratic and clipping its minimizer to the two closed ranges then gives (3.10). □

The flat-spike benchmark. This standard spiked-covariance model gives a particularly transparent positive case. For the rank-one flat spike $\mathbf{\Sigma} = \sigma^2 I_p + \theta u u^{\mathsf{T}}$ with $u = p^{-1/2}\mathbf{1}$, every residual eigenvector u_k satisfies

$$s_{1k} = \frac{1}{p}, \qquad h_{1k} = \frac{2}{p}\left(\sigma^4 + \frac{2\sigma^2\theta}{p}\right), \qquad \frac{\mathsf{C}_{\mathbf{\Sigma}}}{\mathsf{V}_{\mathbf{\Sigma}}} = \frac{2\mathsf{A}_{\mathbf{\Sigma}}}{\mathsf{B}_{\mathbf{\Sigma}}} = \frac{\sigma^2 + \theta}{\sigma^2 + 2\theta/p} \geq 1. \tag{C.41}$$

The derivative $-4\mathsf{A}_{\mathbf{\Sigma}} + 2\alpha\mathsf{B}_{\mathbf{\Sigma}}$ is therefore nonpositive for every $0 \leq \alpha \leq 1$, proving the flat-spike statement in the main text.

D Robustness and validation selection

This section extends the balanced-marginal risk theorem beyond Gaussian sampling and gives a formal guarantee for selecting attenuation by held-out reconstruction while retaining PCA as a candidate.

D.1 Risk expansion beyond Gaussian sampling

The preceding risk theory uses Gaussian fourth moments to obtain an explicit sign. The perturbation mechanism itself is more general: it depends on the actual covariance between each rotation fluctuation and its diagonal component. The next result records this robustness after the Gaussian theory has been completed.

For a distribution ν with covariance $\mathbf{\Sigma}$, define

$$\mathsf{C}_\nu = \sum_{j \leq d < k} \frac{\mathrm{Cov}(a_{jk}, b_{jk})}{\lambda_j - \lambda_k}, \qquad \mathsf{V}_\nu = \sum_{j \leq d < k} \frac{\mathrm{var}(b_{jk})}{\lambda_j - \lambda_k}, \tag{D.1}$$

where a_{jk} and b_{jk} are given in (3.3).

Theorem D.1 (Moment extension of the attenuation risk)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent copies of a mean-zero random vector with covariance $\mathbf{\Sigma}$, $\mathbb{E}\|\mathbf{x}\|_2^8 < \infty$, and balanced marginal variances. Keep p, d fixed and assume a positive cutoff gap. With $a_{jk}, b_{jk}, \mathsf{C}_\nu$, and V_ν as in (3.3) and (D.1), define*

$$\mathsf{G}_\nu(0) = \sum_{j \leq d < k} \frac{\mathrm{var}(a_{jk})}{\lambda_j - \lambda_k},$$
$$\mathsf{G}_\nu(\alpha) = \mathsf{G}_\nu(0) - 2\alpha\mathsf{C}_\nu + \alpha^2\mathsf{V}_\nu. \tag{D.2}$$

For every $0 \leq \bar{\alpha} \leq 1$, there are finite constants K_ν, N_ν such that

$$\sup_{0 \leq \alpha \leq \bar{\alpha}} \left| n\mathbb{E}\mathcal{E}(\widehat{P}_\alpha) - \mathsf{G}_\nu(\alpha) \right| \leq \frac{K_\nu}{\sqrt{n}}, \qquad n \geq N_\nu. \tag{D.3}$$

Proof of Theorem D.1. For a retained–discarded pair, both quantities in (3.3) are centered and

$$u_j^{\mathsf{T}} E_{n,\alpha} u_k = \frac{1}{n}\sum_{i=1}^{n}(a_{jk,i} - \alpha b_{jk,i}).$$

Consequently,

$$n\mathbb{E}(u_j^{\mathsf{T}} E_{n,\alpha} u_k)^2 = \mathrm{var}(a_{jk}) - 2\alpha\,\mathrm{Cov}(a_{jk}, b_{jk}) + \alpha^2\,\mathrm{var}(b_{jk}). \tag{D.4}$$

Summing (D.4) in the leading quadratic of (C.39) gives exactly $\mathsf{G}_\nu(\alpha)/n$.

The eighth-moment condition and fixed dimension imply $\mathbb{E}\|E_n\|_{\mathrm{op}}^r = O(n^{-r/2})$ for $r = 3, 4$. This follows by applying a fixed-dimensional Rosenthal inequality to the entries of $\mathbf{x}_i\mathbf{x}_i^\mathsf{T} - \mathbf{\Sigma}$ and then using equivalence of matrix norms. Since $\|E_{n,\alpha}\|_{\mathrm{op}} \le 2\|E_n\|_{\mathrm{op}}$ uniformly on $0 \le \alpha \le 1$, the cubic remainder in (C.39) contributes $O(n^{-3/2})$ on the perturbative event. Markov's inequality with the fourth moment gives probability $O(n^{-2})$ to its complement. Excess risk is bounded, and the leading quadratic on that complement is controlled by $K\|E_n\|_{\mathrm{op}}^2$; Cauchy–Schwarz makes both complementary contributions $O(n^{-2})$. Multiplication by n proves the uniform bound (D.3). □

For an interval $[a, b]$, let $\Pi_{[a,b]}(t) = \min\{b, \max\{a, t\}\}$ denote Euclidean projection, or clipping, onto that interval. The quadratic also gives the practical consequences. If $\mathsf{V}_\nu = 0$, every b_{jk} is almost surely zero, so Cauchy–Schwarz forces $\mathsf{C}_\nu = 0$ and attenuation leaves the leading risk unchanged. Otherwise, $\mathsf{C}_\nu > 0$ implies $\mathsf{V}_\nu > 0$; the direct leading-risk oracle is $\Pi_{[0,1]}(\mathsf{C}_\nu/\mathsf{V}_\nu)$, and the mask-path oracle is its limiting projection onto $[0, 1/2]$. A fixed level improves the leading risk whenever $0 < \alpha < 2\mathsf{C}_\nu/\mathsf{V}_\nu$, subject also to $\alpha < 1/2$ on the mask path. More precisely, if $\Delta_{\nu,\alpha} = \mathsf{G}_\nu(0) - \mathsf{G}_\nu(\alpha) > 0$, then

$$n \ge N_\nu \quad \text{and} \quad n > \left(\frac{2K_\nu}{\Delta_{\nu,\alpha}}\right)^2 \quad \Longrightarrow \quad \mathbb{E}\mathcal{E}(\widehat{P}_\alpha) < \mathbb{E}\mathcal{E}(\widehat{P}_0). \tag{D.5}$$

D.2 Validation selection with a PCA fallback

The attenuation path contains ordinary PCA at $\alpha = 0$. The next result turns this nesting into a finite-sample safety statement for a simple sample-splitting implementation. Candidate projectors are fitted once on a training sample and then selected on an independent validation sample; no post-selection refitting is needed for the guarantee.

Proposition D.2 (Independent-validation fallback guarantee)**.** *Let $\mathcal{A} \subset [0, 1]$ be a finite set containing 0, and condition on rank-d projectors $\{\widehat{P}_\alpha : \alpha \in \mathcal{A}\}$ fitted from a training sample. Let $Y_1, \dots, Y_{n_v}$ be an independent validation sample from $\mathcal{N}(0, \mathbf{\Sigma})$ and define*

$$\widehat{\mathcal{R}}_v(P) = \frac{1}{n_v}\sum_{i=1}^{n_v}\|Y_i - PY_i\|_2^2, \qquad \widehat{\alpha}_v \in \arg\min_{\alpha\in\mathcal{A}} \widehat{\mathcal{R}}_v(\widehat{P}_\alpha). \tag{D.6}$$

For $0 < \eta < 1$, put $M = |\mathcal{A}|$, $t = \log(2M/\eta)$, and

$$\varepsilon_v(t) = 2\|\mathbf{\Sigma}\|_{\mathrm{F}}\sqrt{\frac{t}{n_v}} + 2\|\mathbf{\Sigma}\|_{\mathrm{op}}\frac{t}{n_v}. \tag{D.7}$$

With conditional probability at least $1 - \eta$,

$$\mathcal{R}(\widehat{P}_{\widehat{\alpha}_v}) \le \min_{\alpha\in\mathcal{A}} \mathcal{R}(\widehat{P}_\alpha) + 2\varepsilon_v(t) \le \mathcal{R}(\widehat{P}_0) + 2\varepsilon_v(t). \tag{D.8}$$

Thus the selected estimator approaches an exact PCA fallback as the validation size grows, while retaining any improvement available elsewhere on the path.

Proof of Proposition D.2. Conditional on the training sample, fix a candidate P and write $Y_i = \mathbf{\Sigma}^{1/2}Z_i$ with $Z_i \sim \mathcal{N}(0, I_p)$. Then

$$\|Y_i - PY_i\|_2^2 = Z_i^\mathsf{T} B_P Z_i, \qquad B_P = \mathbf{\Sigma}^{1/2}(I_p - P)\mathbf{\Sigma}^{1/2} \succeq 0,$$

and $\mathbb{E}(Z_i^\mathsf{T} B_P Z_i) = \mathrm{tr}(B_P) = \mathcal{R}(P)$. The Gaussian quadratic-form inequality of Laurent & Massart (2000), applied to the block-diagonal average of the n_v forms, gives

$$\Pr\left(\left|\widehat{\mathcal{R}}_v(P) - \mathcal{R}(P)\right| > 2\|B_P\|_{\mathrm{F}}\sqrt{\frac{t}{n_v}} + 2\|B_P\|_{\mathrm{op}}\frac{t}{n_v}\right) \le 2e^{-t}.$$

Because $0 \preceq I_p - P \preceq I_p$, we have $0 \preceq B_P \preceq \mathbf{\Sigma}$. Eigenvalue monotonicity under the Loewner order therefore gives $\|B_P\|_{\mathrm{F}} \le \|\mathbf{\Sigma}\|_{\mathrm{F}}$ and $\|B_P\|_{\mathrm{op}} \le \|\mathbf{\Sigma}\|_{\mathrm{op}}$. A union bound over the M candidates therefore makes all validation risks uniformly accurate to $\varepsilon_v(t)$ with probability at least $1 - \eta$. On this event, if $\alpha^* \in \arg\min_{\alpha\in\mathcal{A}} \mathcal{R}(\widehat{P}_\alpha)$, then

$$\mathcal{R}(\widehat{P}_{\widehat{\alpha}_v}) \le \widehat{\mathcal{R}}_v(\widehat{P}_{\widehat{\alpha}_v}) + \varepsilon_v(t) \le \widehat{\mathcal{R}}_v(\widehat{P}_{\alpha^*}) + \varepsilon_v(t) \le \mathcal{R}(\widehat{P}_{\alpha^*}) + 2\varepsilon_v(t).$$

Since $0 \in \mathcal{A}$, the final inequality in (D.8) follows. □

The result is stated for independent validation because this permits a direct finite-sample guarantee. The real-data studies use the data-efficient analogue: all candidate methods share the same grouped cross-validation folds within each fixed fitting budget, and the selected estimator is then refit on that budget.

E ALTERNATIVE MASKING AND RECONSTRUCTION OBJECTIVES

Remark E.1 (Relation to masked and linear autoencoders). Our full-output orthogonal-projector objective has the PCA subspace as its unmasked target. Canonical masked autoencoders score only masked targets, and unconstrained linear autoencoders optimize a broader class of maps. The distinctions are developed in Sections E.2 and E.3 (Baldi & Hornik, 1989; He et al., 2022).

E.1 THE GENERAL BLOCK MASKED-TARGET OBJECTIVE

The canonical masked-target projector loss is

$$\mathcal{L}_{\mathrm{MT}}(P) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_W \|(I_p - W)(\mathbf{x}_i - PW\mathbf{x}_i)\|_2^2, \qquad P \in \mathcal{P}_d. \tag{E.1}$$

It evaluates reconstruction error only on coordinates hidden from the encoder, whereas the full-output loss in (2.4) evaluates error on the entire clean vector.

The matched empirical comparison in Section F.2 includes both coordinate masks and masks that hide contiguous coordinate groups. Let $\mathcal{G} = \{G_1, \ldots, G_G\}$ be a partition of the coordinates. For a matrix A, let $\mathcal{D}_{\mathcal{G}}(A)$ retain its within-group diagonal blocks and put $\mathcal{O}_{\mathcal{G}}(A) = A - \mathcal{D}_{\mathcal{G}}(A)$. A group mask has the form $W = \mathrm{blockdiag}(w_1 I_{G_1}, \ldots, w_G I_{G_G})$, where the w_g are independent Bernoulli-q variables. Write $R = I - W$ and $D_P = \mathcal{D}_{\mathcal{G}}(P)$.

The calculation used for Proposition 2.1 also gives the full-output group-masked matrix

$$\mathbf{T}_{n,\alpha,\mathcal{G}} = \mathbf{S}_n - \alpha \mathcal{D}_{\mathcal{G}}(\mathbf{S}_n), \qquad \alpha = \frac{1-q}{2-q}. \tag{E.2}$$

Thus group masking attenuates within-group covariance blocks rather than only coordinate variances. The masked-target loss has a different reduction.

Proposition E.1 (Exact symmetric masked-target objective)**.** *For $m = 1 - q$ and every $P \in \mathcal{P}_d$, the group version of* (E.1) *satisfies*

$$\mathcal{L}_{\mathrm{MT}}(P) = m\,\mathrm{tr}(\mathbf{S}_n) + mq\left[-2\,\langle \mathbf{S}_n, \mathcal{O}_{\mathcal{G}}(P)\rangle + \langle \mathbf{S}_n, \mathcal{C}_{q,\mathcal{G}}(P)\rangle\right], \tag{E.3}$$

$$\mathcal{C}_{q,\mathcal{G}}(P) = D_P - D_P^2 + q\,\mathcal{O}_{\mathcal{G}}\{P - D_P P - P D_P\}. \tag{E.4}$$

Proof of Proposition E.1. Expand

$$\|R(x - PWx)\|_2^2 = x^{\mathsf{T}} R x - 2x^{\mathsf{T}} RPWx + x^{\mathsf{T}} WPRPWx. \tag{E.5}$$

For groups g, h, $\mathbb{E}\{(1 - w_g)w_h\}$ is zero when $g = h$ and equals mq otherwise. Hence

$$\mathbb{E}(RPW) = mq\,\mathcal{O}_{\mathcal{G}}(P). \tag{E.6}$$

For the last term, its (g, g) block is $mq(P_{gg} - P_{gg}^2)$ by $P^2 = P$. When $g \neq h$, only intermediate groups different from both g and h survive, giving

$$\mathbb{E}(WPRPW)_{gh} = mq^2\{P - D_P P - P D_P\}_{gh}.$$

Combining the diagonal and off-block cases yields $\mathbb{E}(WPRPW) = mq\mathcal{C}_{q,\mathcal{G}}(P)$. Averaging (E.5) over the sample proves (E.3). □

Unlike the full-output loss, the result contains nonlinear functions of P and is not generally a spectral estimator. Nevertheless, it can be compared locally with diagonal attenuation by considering the same retained–omitted direction pairs used in Section 3.1.

E.2 Masked-target random slopes at the PCA target

Fix a retained population direction u_j, $j \le d$, and an omitted direction u_k, $k > d$. The rotation that replaces u_j by a direction in their span is

$$u_{jk}(s) = \frac{u_j + s u_k}{\sqrt{1+s^2}}, \qquad P_{jk}(s) = P_* - u_j u_j^\top + u_{jk}(s) u_{jk}(s)^\top. \tag{E.7}$$

Here $s = \tan(\theta)$ parameterizes the rotation angle. Its exact clean reconstruction cost is

$$\mathcal{E}\{P_{jk}(s)\} = \frac{(\lambda_j - \lambda_k) s^2}{1+s^2}. \tag{E.8}$$

We specialize to coordinate masks with $0 < q < 1$, so $mq > 0$, and remove this positive factor and the constant term in (E.3). Thus write

$$\Phi_q(P; S) = -2 \langle S, \mathrm{off}(P) \rangle + \langle S, \mathcal{C}_q(P) \rangle, \tag{E.9}$$

where $\mathcal{C}_q$ is (E.4) with singleton coordinate groups. For a rank-d projector P and a symmetric matrix S, define the symmetric matrix $\Gamma_q(P; S)$ entrywise by

$$[\Gamma_q(P;S)]_{ab} = \begin{cases} \{-2 + q(1 - P_{aa} - P_{bb})\} S_{ab}, & a \neq b, \\ (1 - 2P_{aa}) S_{aa} - 2q \sum_{c \neq a} S_{ac} P_{ac}, & a = b. \end{cases} \tag{E.10}$$

Proposition E.2 (Masked-target slopes at the PCA target)**.** *Let P_* be the leading rank-d projector of $\mathbf{\Sigma}$, and let $P_{jk}(s)$ be the rotation path in (E.7). For every $j \le d < k$ and every symmetric S,*

$$\frac{d}{ds} \Phi_q\{P_{jk}(s); S\}\Big|_{s=0} = 2 u_k^\top \Gamma_q(P_*; S) u_j. \tag{E.11}$$

Consequently, P_* is stationary for the population masked-target objective if and only if

$$P_*^\perp \Gamma_q(P_*; \mathbf{\Sigma}) P_* = 0. \tag{E.12}$$

Suppose $u_k^\top \Gamma_q(P_*; \mathbf{\Sigma}) u_j = 0$ for a particular pair (j,k) and define the population profile curvature and centered single-observation slope by

$$\kappa_{jk,q} = \frac{d^2}{ds^2} \Phi_q\{P_{jk}(s); \mathbf{\Sigma}\}\Big|_{s=0}, \tag{E.13}$$

$$\psi_{jk,q}(x) = 2 u_k^\top \{\Gamma_q(P_*; x x^\top) - \Gamma_q(P_*; \mathbf{\Sigma})\} u_j. \tag{E.14}$$

If $\kappa_{jk,q} > 0$ and $\widehat{s}^{\mathrm{MT}}_{jk,q}$ is the local minimizer of the empirical profile converging to zero, then, under Assumption 2.1,

$$\widehat{s}^{\mathrm{MT}}_{jk,q} = -\frac{n^{-1} \sum_{i=1}^n \psi_{jk,q}(\mathbf{x}_i)}{\kappa_{jk,q}} + O_{\mathbb{P}}(n^{-1}), \tag{E.15}$$

$$\sqrt{n}\, \widehat{s}^{\mathrm{MT}}_{jk,q} \implies \mathcal{N}\left(0, \frac{\mathrm{var}\{\psi_{jk,q}(\mathbf{x})\}}{\kappa^2_{jk,q}}\right), \tag{E.16}$$

$$n \mathcal{E}\{P_{jk}(\widehat{s}^{\mathrm{MT}}_{jk,q})\} \implies (\lambda_j - \lambda_k) \frac{\mathrm{var}\{\psi_{jk,q}(\mathbf{x})\}}{\kappa^2_{jk,q}} \chi^2_1. \tag{E.17}$$

Proof of Proposition E.2. Put $\Pi = \mathrm{diag}(P)$ and $\dot{\Pi} = \mathrm{diag}(H)$. Differentiating (E.9) in an arbitrary symmetric direction H gives

$$\begin{aligned} D\Phi_q(P;S)[H] = -2 \langle S, \mathrm{off}(H) \rangle + \left\langle S, \dot{\Pi} - \dot{\Pi}\Pi - \Pi\dot{\Pi} \right\rangle \\ + q \left\langle \mathrm{off}(S), H - \dot{\Pi}P - \Pi H - H\Pi - P\dot{\Pi} \right\rangle. \end{aligned}$$

Collecting the coefficient of each entry of H yields $D\Phi_q(P;S)[H] = \langle \Gamma_q(P;S), H\rangle$ with Γ_q as in (E.10). Because $P'_{jk}(0) = u_j u_k^\top + u_k u_j^\top$, (E.11) follows. Such matrices span the tangent space of rank-d projectors at P_*, proving (E.12).

Under stationarity, the empirical derivative at zero is $n^{-1}\sum_i \psi_{jk,q}(\mathbf{x}_i)$. A second-order Taylor expansion of the profile score, Gaussian moment bounds, and $\kappa_{jk,q} > 0$ give (E.15). The central limit theorem and Slutsky's theorem then give (E.16). Finally, the exact identity (E.8) and the continuous mapping theorem give (E.17). □

This result exposes the distinction between the two objectives. Under balanced marginal variances, diagonal attenuation automatically has zero population leakage and retains the PCA curvature. The masked-target objective still requires the separate stationarity condition (E.12) and positive profile curvature. For the joint rank-d estimator, the scalar curvatures in (E.17) are replaced by the Hessian of Φ_q on the full tangent space; cross-direction Hessian terms need not vanish.

A closed-form example shows that stationarity alone is insufficient. Let $p = 2$, $d = 1$, $u_1 = u = (1,1)^\top/\sqrt{2}$, $u_2 = v = (1,-1)^\top/\sqrt{2}$, and $\mathbf{\Sigma} = \lambda_1 uu^\top + \lambda_2 vv^\top$. Along $P_{12}(s)$, the single-observation slope is identically zero. Substituting $[P_{12}(s)]_{12} = (1-s^2)/[2(1+s^2)]$ into (E.9) gives

$$\kappa_{12,q} = 2(\lambda_1 - 3\lambda_2). \tag{E.18}$$

Equivalently, the population loss is a convex quadratic in $t = P_{12}$ with unconstrained minimizer $t = (\lambda_1 - \lambda_2)/(\lambda_1 + \lambda_2)$ and feasible range $[-1/2, 1/2]$. Thus the population PCA direction is the masked-target solution when $\lambda_1/\lambda_2 \geq 3$ and is not even a local minimum when $1 < \lambda_1/\lambda_2 < 3$. The familiar threshold is therefore a curvature condition. In dimensions above two, the population slope need not vanish. When the stationarity and positive-curvature conditions hold, its empirical fluctuation produces the usual $n^{-1/2}$ local displacement.

E.3 Linear reconstruction beyond the projector constraint

The main estimator constrains the reconstruction map to be an orthogonal rank-d projector because the target is PCA. This section removes that constraint. It gives an exact characterization for a symmetric-map (weight-tied) linear autoencoder and for a fully asymmetric encoder–decoder pair. The latter is the coordinate-mask specialization of the linear full-output model analyzed by Bisulco et al. (2025). The calculation also makes precise the direction-specific gain shrinkage studied for additive-noise autoencoders by Pretorius et al. (2018).

Because these maps need not be projectors, define their clean reconstruction risk by

$$\mathcal{R}_{\mathrm{map}}(A) = \mathbb{E}\|\mathbf{x} - A\mathbf{x}\|_2^2 = \mathrm{tr}\{(I_p - A)^\top (I_p - A)\mathbf{\Sigma}\}. \tag{E.19}$$

For an orthogonal projector P, this agrees with $\mathcal{R}(P)$ in (2.2).

Write $D_n = \mathrm{diag}(\mathbf{S}_n)$ and, for $0 < q < 1$, define

$$V_{n,q} = q\mathbf{S}_n + (1-q)D_n. \tag{E.20}$$

If every sample coordinate variance is positive, then $V_{n,q} \succ 0$ even when $\mathbf{S}_n$ is singular. An asymmetric rank-d encoder–decoder has end-to-end map $A = LR^\top$, where $L, R \in \mathbb{R}^{p\times d}$; conversely, every matrix with rank at most d admits such a factorization. Weight tying sets $L = R = B$ and hence restricts A to $BB^\top \succeq 0$. For either class, consider the same full-output masked-input loss as in the main text,

$$\mathcal{L}_q(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_W \|\mathbf{x}_i - AW\mathbf{x}_i\|_2^2. \tag{E.21}$$

Theorem E.3 (Full-output linear autoencoders)**.** *Assume* $0 < q < 1$ *and* $V_{n,q} \succ 0$.

(i) Asymmetric map. *Put* $C_{n,q} = \mathbf{S}_n V_{n,q}^{-1/2}$. *Every global minimizer over* $\mathrm{rank}(A) \leq d$ *has the form*

$$A_{\mathrm{asym}} = C_d V_{n,q}^{-1/2}, \qquad C_d \in \operatorname*{arg\,min}_{\mathrm{rank}(C')\leq d} \|C' - C_{n,q}\|_{\mathrm{F}}. \tag{E.22}$$

If the d-th and $(d+1)$-th singular values of $C_{n,q}$ are distinct, this map is unique and its column space is the leading eigenspace of

$$\mathbf{S}_n V_{n,q}^{-1} \mathbf{S}_n. \tag{E.23}$$

(ii) Symmetric map. *Every tied map $A = BB^\mathsf{T} \succeq 0$ with rank at most d can be written as $A = U \operatorname{diag}(a_1, \dots, a_d) U^\mathsf{T}$, where $U^\mathsf{T} U = I_d$ and $a_r \geq 0$, allowing $a_r = 0$ when the rank is below d. Here a_r is the reconstruction multiplier along u_r: $Au_r = a_r u_r$. For fixed $U = (u_1, \dots, u_d)$, minimizing the loss over these multipliers gives*

$$a_r^*(U) = \frac{u_r^\mathsf{T} \mathbf{S}_n u_r}{u_r^\mathsf{T} V_{n,q} u_r}. \tag{E.24}$$

Consequently, the global tied directions solve

$$\max_{U^\mathsf{T} U = I_d} \sum_{r=1}^{d} \frac{(u_r^\mathsf{T} \mathbf{S}_n u_r)^2}{u_r^\mathsf{T} V_{n,q} u_r}. \tag{E.25}$$

In general, these solutions differ from both ordinary PCA and the attenuated spectral estimator $P_{\mathbf{S}_n - \alpha D_n}$.

Balanced population specialization. Suppose $\operatorname{diag}(\mathbf{\Sigma}) = \tau I_p$ and $\mathbf{\Sigma} = \sum_j \lambda_j u_j u_j^\mathsf{T}$, with $\lambda_d > \lambda_{d+1}$. The asymmetric population map is

$$A_{q,\mathrm{asym}}^{\mathrm{pop}} = \sum_{j=1}^{d} \frac{\lambda_j}{q\lambda_j + (1-q)\tau} u_j u_j^\mathsf{T}. \tag{E.26}$$

Its range is the PCA subspace, but its clean reconstruction risk relative to the PCA projector is

$$\mathcal{R}_{\mathrm{map}}(A_{q,\mathrm{asym}}^{\mathrm{pop}}) - \mathcal{R}(P_*) = \sum_{j=1}^{d} \lambda_j \left\{ \frac{(1-q)(\tau - \lambda_j)}{q\lambda_j + (1-q)\tau} \right\}^2. \tag{E.27}$$

For $q < 1$, this excess is strictly positive whenever some retained eigenvalue differs from τ. Thus the asymmetric autoencoder learns the correct subspace, while its raw reconstruction map is generally not the projector P_*.

Proof of Theorem E.3 and the balanced population specialization. For an arbitrary A, expansion of (E.21) gives

$$\begin{aligned} \mathcal{L}_q(A) &= \operatorname{tr}(\mathbf{S}_n) - 2q \operatorname{tr}(A\mathbf{S}_n) + q \operatorname{tr}(AV_{n,q}A^\mathsf{T}) \\ &= \operatorname{tr}(\mathbf{S}_n) - q \operatorname{tr}(\mathbf{S}_n V_{n,q}^{-1} \mathbf{S}_n) + q \|AV_{n,q}^{1/2} - \mathbf{S}_n V_{n,q}^{-1/2}\|_{\mathrm{F}}^2. \end{aligned} \tag{E.28}$$

The first equality uses $\mathbb{E}(W\mathbf{S}_n W) = q^2 \mathbf{S}_n + q(1-q) D_n = qV_{n,q}$. For the second, expand the squared Frobenius norm and use cyclicity of trace: its cross term is $-2q \operatorname{tr}\{AV_{n,q}^{1/2} V_{n,q}^{-1/2} \mathbf{S}_n\} = -2q \operatorname{tr}(A\mathbf{S}_n)$.

Right multiplication by the invertible $V_{n,q}^{1/2}$ preserves rank. Thus minimizing (E.28) over $\operatorname{rank}(A) \leq d$ is exactly the best rank-at-most-d approximation of $C_{n,q} = \mathbf{S}_n V_{n,q}^{-1/2}$. The Eckart–Young theorem gives (E.22). The left singular vectors of $C_{n,q}$ are the eigenvectors of $C_{n,q} C_{n,q}^\mathsf{T} = \mathbf{S}_n V_{n,q}^{-1} \mathbf{S}_n$, proving (E.23).

With a singular-value gap, (E.22) yields

$$A_{\mathrm{asym}} = P_{\mathbf{S}_n V_{n,q}^{-1} \mathbf{S}_n} \mathbf{S}_n V_{n,q}^{-1}. \tag{E.29}$$

For a tied map, substitute $A = U \operatorname{diag}(a_1, \dots, a_d) U^\mathsf{T}$ into the first line of (E.28). The part depending on A, divided by q, is

$$-2 \sum_{r=1}^{d} a_r u_r^\mathsf{T} \mathbf{S}_n u_r + \sum_{r=1}^{d} a_r^2 u_r^\mathsf{T} V_{n,q} u_r. \tag{E.30}$$

There are no cross terms in the final sum: the trace selects only the diagonal entries after the two diagonal coefficient matrices are applied. Each scalar quadratic in (E.30) is strictly convex and has minimizer (E.24). Substitution subtracts $(u_r^\top \mathbf{S}_n u_r)^2/(u_r^\top V_{n,q} u_r)$ from the loss, yielding (E.25).

Under population balance, $V_q = q\mathbf{\Sigma} + (1-q)\tau I_p$ shares every eigenvector with $\mathbf{\Sigma}$. For an eigenvalue $\lambda > 0$, the squared singular value that orders the asymmetric solution is

$$f_q(\lambda) = \frac{\lambda^2}{q\lambda + (1-q)\tau}. \tag{E.31}$$

Its derivative is

$$f_q'(\lambda) = \frac{\lambda\{q\lambda + 2(1-q)\tau\}}{\{q\lambda + (1-q)\tau\}^2} > 0,$$

so the asymmetric solution selects $u_1, \ldots, u_d$. Equation (E.22) gives its multiplier $\lambda_j/\{q\lambda_j + (1-q)\tau\}$, proving (E.26). Finally,

$$\mathcal{R}_{\text{map}}(A) - \mathcal{R}(P_*) = \sum_{j=1}^{d} \lambda_j (1 - a_j)^2$$

for any map diagonal in the population eigenbasis that acts as zero on the orthogonal complement of the leading subspace. Substituting the displayed multiplier proves (E.27) and its strictness statement.

□

E.4 Additional related context

PCA and covariance regularization. Classical and modern PCA theory describes eigenspace fluctuations, perturbation bounds, and reconstruction risk (Anderson, 1963; Koltchinskii & Lounici, 2017; Reiß & Wahl, 2020; El Hanchi et al., 2025). Linear shrinkage toward a scalar identity improves covariance conditioning but preserves the sample eigenvectors (Ledoit & Wolf, 2004); shrinkage toward a diagonal target also modifies sample cross-covariances (Schäfer & Strimmer, 2005). Diagonal attenuation takes a different direction: it retains the observed cross-covariances and weakens only the sample diagonal. Diagonal deletion is established for noisy or incomplete low-rank matrices, and hollowed PCA can improve spectral recovery under heteroskedastic noise (Cai et al., 2021; Abbe et al., 2022). HeteroPCA further alternates diagonal deletion and imputation in a latent low-rank measurement model (Zhang et al., 2022). Our observations are instead complete draws from the covariance whose PCA subspace is the target. We study the continuous path from no deletion to complete deletion and ask when a partial correction lowers the clean reconstruction risk of ordinary PCA.

Masking and linear reconstruction objectives. Denoising autoencoders learn from corrupted inputs (Vincent et al., 2008; 2010), and marginalized denoising integrates over the corruption to obtain a deterministic moment objective (Chen et al., 2012; 2014). An unmasked undercomplete linear autoencoder recovers the PCA subspace under squared loss (Baldi & Hornik, 1989). The objective and feasible-set distinctions are summarized in Remark E.1 and derived above.

Closest spectral analyses of masking. Ji et al. (2023) connect complementary random masks in a contrastive objective to diagonal deletion and feature recovery under a spiked additive noise model. Our full-output projector objective instead produces a continuous partial-attenuation path and is evaluated against the ordinary PCA subspace of the clean covariance. Bisulco et al. (2025) study full-output block masking with an untied linear encoder and decoder. For a fixed data matrix, they derive the global minimizers through a generalized eigenvalue problem and characterize all critical points. We instead constrain the reconstruction map to be an orthogonal projector and compare its expected clean PCA risk under random sampling. Zurich et al. (2026) analyze coordinatewise masked self-supervised ridge regression in proportional asymptotics; their estimator is an asymmetric full-rank predictor trained to infer hidden coordinates. Our estimator is a fixed-rank projector, and our risk theory is fixed dimensional and centered on the finite-sample PCA target. More broadly, Lin et al. (2024) show that linear data augmentation induces data-dependent spectral regularization. Here we identify the particular cross-subspace fluctuations, attenuated gaps, and population movement that determine the clean-risk comparison with PCA.

Other corruption and missing-data settings. Input corruptions can induce coordinate-dependent regularization (Wager et al., 2013). For linear denoising autoencoders, Pretorius et al. (2018) study direction-specific reconstruction gains under additive isotropic noise, while Kong et al. (2023) relate the mask ratio and patch size of a hidden-target objective to identifiable levels of a hierarchical latent model. Missing-data PCA instead observes physically reduced vectors and estimates an unobserved signal covariance (Dobriban et al., 2016). These settings motivate nearby uses of corruption or diagonal operators, but they differ from the complete-vector PCA target and clean reconstruction risk studied here. The exact objective and feasible-set boundaries needed for that distinction are given in Sections E.2 and E.3.

F Additional numerical and real-data studies

The main text reports Gaussian experiments for preserved and moving population targets. This section adds a non-Gaussian stress test and gives the complete protocol and cellwise results for the fixed fitting-sample-budget real-data study.

F.1 Robustness beyond Gaussian observations

The non-Gaussian audit uses the same balanced $p = 16, d = 3$ covariance as the general Gaussian experiment. Before multiplication by the covariance square root, the latent coordinates are either Gaussian, standardized Student t_{10}, or standardized centered exponential. All three constructions have covariance $\mathbf{\Sigma}$ and finite eighth moments. A separate pilot of 150,000 observations per distribution estimates the constants in (D.2); each sample size $n \in \{64, 256, 1024\}$ then uses 5,000 fresh common-sample replicates, with all attenuation strengths evaluated on the same observations. At $n = 1024$, the largest absolute difference between the simulated scaled risk and the moment prediction across all displayed distributions and attenuation levels is 1.02. The sign and curvature predicted by Theorem D.1 therefore persist under both heavy-tailed and skewed observations in this controlled design.

Figure 4 displays this agreement between the simulated curves and the fourth-moment prediction across all three latent distributions.

F.2 Fixed-budget real-data comparisons

Implementation details. The data, ranks, patch construction, block definitions, and fixed fitting-sample-budget protocol are given in Section G.3; here we record the additional details needed to reproduce the comparison. Every covariance matrix is centered within its fitting fold, and no coordinate rescaling is applied except within the correlation-PCA baseline. When a validation curve has multiple minimizers, we choose the smallest attenuation strength or mask rate. For Optdigits, each of the ten patch-location seeds contributes a mean over 100 training subsamples, and intervals use the standard error of these ten seed-level means. For Phoneme and HAR, intervals use the standard error across 100 training subsamples, conditional on the observed data and official split. Within every subsample, all methods use the same observations and validation folds.

HeteroPCA starts from the hollowed sample covariance. At each iteration it replaces the diagonal by that of the best rank-d truncated-SVD approximation while retaining the original sample off-diagonal; it stops when the relative change of the imputed diagonal is at most 10^{-7} or after 200 iterations (Zhang et al., 2022). Of its 2,400 fits, 133 reach the iteration cap. Each masked-target projector is approximately optimized on the Grassmann manifold from both the PCA solution and the corresponding full-output block-attenuation solution. Optimization stops when the relative tangent-gradient norm is at most 10^{-7} or after 800 iterations. Among the 1,163 coordinate-masked fits that do not select the PCA fallback, 86 reach this cap; none of the 175 nonfallback block-masked fits does. We therefore report masked-target training as a two-start optimization rather than a certified global optimum.

Table 1 gives the complete attenuation results. Both paths have positive mean gains with intervals above zero in all six cells. Table 2 reports the four comparison methods on all three data sets. Correlation PCA is competitive on Optdigits and HAR but does not improve Phoneme. Coordinate masked-target training is positive in three ordered-signal cells but not on the image patches, while

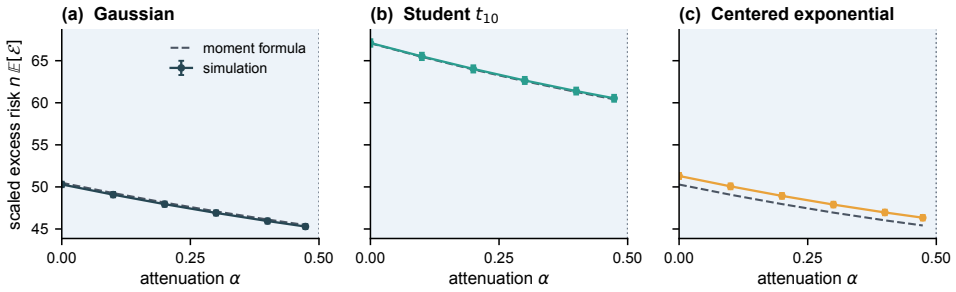

Figure 4: The moment risk formula beyond Gaussian observations. Solid lines and intervals are Monte Carlo estimates at $n = 1024$; dashed lines are the corresponding fourth-moment predictions. The covariance and target rank are common across panels, and shading marks the mask-feasible interval.

Table 1: Fixed fitting-sample-budget results for diagonal attenuation. Entries are the mean percentage of PCA test excess reconstruction risk removed, with 95% intervals in brackets. Optdigits intervals treat the ten patch-location seeds as independent clusters, with each seed-level mean averaging 100 training subsamples; the other intervals use 100 training subsamples.

Data	N	Mask-derived DA	Direct DA
Optdigits	64	4.32 [4.08, 4.57]	6.74 [6.30, 7.17]
Optdigits	128	4.18 [3.85, 4.51]	6.40 [5.80, 7.00]
Phoneme	64	0.84 [0.65, 1.03]	1.09 [0.78, 1.40]
Phoneme	128	0.72 [0.51, 0.93]	0.70 [0.36, 1.03]
HAR	64	0.65 [0.45, 0.85]	0.76 [0.48, 1.05]
HAR	128	0.93 [0.78, 1.07]	1.20 [0.97, 1.44]

Table 2: Comparison methods across the three real-data studies. Each entry is the mean percentage of PCA test excess reconstruction risk removed, with its 95% interval in brackets. Optdigits intervals cluster over ten patch-location seeds, each averaging 100 training subsamples; the remaining intervals use 100 training subsamples. MT denotes masked-target training.

Data	N	Correlation PCA	HeteroPCA	Coordinate MT	Block MT
Optdigits	64	5.84 [5.34, 6.33]	−25.79 [−27.60, −23.99]	−1.63 [−2.01, −1.25]	−3.58 [−4.55, −2.61]
Optdigits	128	5.41 [4.99, 5.82]	−37.36 [−41.93, −32.79]	−1.24 [−1.80, −0.68]	−1.27 [−1.98, −0.56]
Phoneme	64	−0.14 [−0.70, 0.42]	0.04 [−0.53, 0.60]	0.93 [0.57, 1.28]	0.00 [0.00, 0.00]
Phoneme	128	−0.52 [−1.06, 0.01]	0.53 [0.42, 0.64]	0.28 [−0.10, 0.66]	0.00 [0.00, 0.00]
HAR	64	0.36 [0.06, 0.66]	−13.94 [−15.46, −12.43]	0.56 [0.28, 0.83]	0.00 [0.00, 0.00]
HAR	128	1.01 [0.74, 1.28]	−4.87 [−6.23, −3.50]	0.93 [0.64, 1.23]	0.00 [0.00, 0.00]

Table 3: Mean cross-validation-selected strengths under fitting-sample budget N. DA denotes diagonal attenuation and MT denotes masked-target training. Parentheses give the percentage of fits selecting the explicit PCA fallback. DA columns report α; MT columns report the mask rate m.

Data	N	Mask DA α	Direct DA α	Coordinate MT m	Block MT m
Optdigits	64	0.474 (1.7%)	0.927 (1.7%)	0.432 (51.8%)	0.121 (86.6%)
Optdigits	128	0.479 (1.1%)	0.931 (1.1%)	0.323 (64.1%)	0.037 (95.9%)
Phoneme	64	0.432 (7.0%)	0.800 (7.0%)	0.589 (18.0%)	0.000 (100.0%)
Phoneme	128	0.434 (7.0%)	0.764 (7.0%)	0.479 (15.0%)	0.000 (100.0%)
HAR	64	0.351 (22.0%)	0.615 (18.0%)	0.370 (27.0%)	0.000 (100.0%)
HAR	128	0.418 (9.0%)	0.756 (8.0%)	0.563 (18.0%)	0.000 (100.0%)

the block-masked family usually selects its explicit PCA fallback. Thus none of these alternatives reproduces the uniform improvement of diagonal attenuation.

Table 3 shows how the validation rule uses the path. Across the three data sets, mean selected strengths are 0.351–0.479 on the mask-derived path and 0.615–0.931 on the full direct path. The direct search selects strengths beyond $1/2$ on 63%–95% of fits, while retaining PCA as a candidate in every search.

G Additional details for the conference version

G.1 Balanced residual coupling and rotation

The leakage matrix $\overline{Z}_{n,\alpha}$ in (3.4) has (u_k, u_j) entry $\overline{a}_{jk} - \alpha\overline{b}_{jk}$. At $\alpha = 0$, PCA uses the unattenuated coupling $\overline{a}_{jk}$. Attenuation subtracts its sample-diagonal component $\alpha\overline{b}_{jk}$, leaving the residual coupling $\overline{a}_{jk} - \alpha\overline{b}_{jk}$. The next lemma shows how this residual coupling produces subspace rotation and clean reconstruction risk. To express the resulting rotation, divide each residual coupling by its population eigengap and define

$$X_{n,\alpha} := \sum_{j \le d < k} \frac{\overline{a}_{jk} - \alpha\overline{b}_{jk}}{\lambda_j - \lambda_k} u_k u_j^\mathsf{T}, \qquad x_{n,kj,\alpha} := u_k^\mathsf{T} X_{n,\alpha} u_j = \frac{\overline{a}_{jk} - \alpha\overline{b}_{jk}}{\lambda_j - \lambda_k}. \tag{G.1}$$

Lemma G.1 (Residual coupling, rotation, and clean risk)**.** *Under Assumptions 2.1, 2.2 and 3.1, for each fixed $\alpha \in [0,1]$ there is a remainder $R_{n,\alpha}$ such that, for every $j \le d < k$,*

$$\begin{aligned} P_*^\perp(\widehat{P}_\alpha - P_*)P_* &= X_{n,\alpha} + R_{n,\alpha}, \qquad \|R_{n,\alpha}\|_\mathrm{F} = O_\mathbb{P}(n^{-1}), \\ u_k^\mathsf{T}(\widehat{P}_\alpha - P_*)u_j &= x_{n,kj,\alpha} + u_k^\mathsf{T} R_{n,\alpha} u_j = x_{n,kj,\alpha} + O_\mathbb{P}(n^{-1}). \end{aligned} \tag{G.2}$$

Moreover,

$$\mathcal{E}(\widehat{P}_\alpha) = \sum_{j \le d < k} (\lambda_j - \lambda_k) x_{n,kj,\alpha}^2 + O_\mathbb{P}(n^{-3/2}) = \sum_{j \le d < k} \frac{(\overline{a}_{jk} - \alpha\overline{b}_{jk})^2}{\lambda_j - \lambda_k} + O_\mathbb{P}(n^{-3/2}). \tag{G.3}$$

The scalar $x_{n,kj,\alpha}$ is the first-order rotation from u_j toward u_k, so a larger eigengap weakens the effect of the same residual coupling. The risk formula weights its square by that eigengap, leaving one eigengap in the denominator. This is a full rank-d statement and does not require the pairwise couplings to be independent. The proof is given in Section C.4.1. Although the individual coordinates $x_{n,kj,\alpha}$ in (G.1) depend on the chosen eigenbasis when eigenvalues repeat within either block, their aggregate contribution in (G.3) does not. This basis invariance suggests a coordinate-free risk analysis for general covariance, which we develop in Section 3.2.

G.2 Gaussian experiment construction and evaluation

Settings. We begin with two target-preserving covariances. The balanced design in Figure 2(a) has $p = 16$, target rank $d = 3$, and eigenvectors given by normalized columns of a Hadamard matrix, so each loading has magnitude $p^{-1/2}$. Specifically, the covariance is $I_p + \sum_{r=1}^3 \theta_r u_r u_r^\mathsf{T}$ with $(\theta_1, \theta_2, \theta_3) = (8, 4, 2)$. Panel (b) uses target rank two and the unequal covariance

$$\mathbf{\Sigma} = \begin{pmatrix} 4 & 2 & 0.75 & 0.25 \\ 2 & 4 & 0.25 & 0.75 \\ 0.75 & 0.25 & 2.5 & 1.5 \\ 0.25 & 0.75 & 1.5 & 2.5 \end{pmatrix}. \tag{G.4}$$

Here $D = \mathrm{diag}(\mathbf{\Sigma})$ is not proportional to the identity, but it leaves the leading two-dimensional target invariant and preserves its spectral ordering along the attenuation path.

To move the population target continuously, panel (c) uses the bivariate family

$$\mathbf{\Sigma}_\phi = R_{\pi/4-\phi} \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix} R_{\pi/4-\phi}^\mathsf{T}, \qquad 0 \le \phi \le \frac{\pi}{8}, \qquad R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{G.5}$$

Let $D_\phi = \mathrm{diag}(\mathbf{\Sigma}_\phi)$ and let $P_{*,\phi}$ be the leading population projector of $\mathbf{\Sigma}_\phi$. The population coupling introduced by attenuation is measured by

$$\|P_{*,\phi}^\perp D_\phi P_{*,\phi}\|_\mathrm{F} = \sin(4\phi). \tag{G.6}$$

It increases from zero to one along the path. Thus panel (c) changes target movement while keeping the two population eigenvalues fixed; in contrast, panel (b) shows that unequal marginal variances alone need not move the target.

Evaluation. For panels (a) and (b), every point averages 3,000 Gaussian samples at $n = 256$ and reports $n\mathbb{E}\mathcal{E}(\widehat{P}_\alpha)$. The theoretical curve is $\mathcal{K}(\alpha)$ from (3.17), with no fitted parameters. For panel (c), we use 13 values of ϕ from 0° to 22.5°, spaced by 1° through 5° and by 2.5° thereafter. At each value, 3,000 independent samples of size $n = 64$ estimate the mean clean risk over $\{0, 0.025, \ldots, 1\} \cup \{0.49\}$. We compare its minimizing strength with the minimizer of $\mathsf{B}_{\mathrm{pop}}(\alpha) + \mathcal{K}_{\mathrm{move}}(\alpha)/n$ on the same grid.

G.3 Real-data construction and selection protocol

Settings. A fitting-sample budget N gives every method the same N observations for fitting and five-fold reconstruction cross-validation. The selected method is then refit on all N observations and evaluated on the official test set. We use $N \in \{64, 128\}$ throughout.

The Optdigits data comprise 3,823 training and 1,797 test grayscale images of size 8×8 (Alpaydin & Kaynak, 1998). Each image contributes one 3×3 patch: before examining its pixel values, we sample the patch's upper-left corner uniformly from the 36 possible locations and flatten the patch into a vector in $\mathbb{R}^9$, so $p = 9$. Thus the nine coordinates record relative positions within a local patch, while the sampled patches range over the full image. We set $d = 2$. Ten independent patch-location seeds define ten versions of the training and test data; for each seed and budget, we draw 100 training subsamples. Within each subsample, all methods receive the same training patches and the same five validation folds.

The Phoneme data contain 256 ordered log-periodogram ordinates for each speech frame (Hastie et al., 2009). We retain the official speaker-disjoint split of 3,340 training and 1,169 test frames, set $d = 8$, and use all $p = 256$ coordinates. For the structured masked-target baseline, we group every 16 adjacent frequency bins. The UCI Human Activity Recognition data contain three-axis total acceleration over 128 time points in each 2.56-second window (Reyes-Ortiz et al., 2013). We use the magnitude series, retain the official split of 7,352 windows from 21 training subjects and 2,947 windows from 9 test subjects, use all $p = 128$ coordinates, set $d = 16$, and group every eight adjacent time points for the structured masked-target baseline. Each ordered-signal cell uses 100 training subsamples. All methods again share the observations and validation folds within each subsample, with all observations from the same speaker or subject kept in one fold. Digit, speech, and activity labels are never used.

Methods and evaluation. Every data set compares ordinary PCA with six alternatives under the same fitting budget. Mask-derived and direct diagonal attenuation retain the original coordinate scales and select α from $\{0, 0.05, \ldots, 1\} \cup \{0.49\}$, restricted to $[0, 0.49]$ for the mask-derived path. Correlation PCA instead standardizes coordinates before estimating the subspace, while HeteroPCA iteratively imputes the sample diagonal. Coordinate and block masked-target projectors select a mask rate from $\{0, 0.25, 0.50, 0.75, 0.90\}$, which includes PCA at zero. The structured coordinate groups are the three rows of an Optdigits patch, consecutive sets of 16 frequency bins for Phoneme, and consecutive sets of eight time points for HAR. Definitions and optimization details are given in Section F.2.

G.4 Numerical comparison across fitting budgets

Table 4: Comparison under the same fitting-sample budgets. Entries are the mean percentage of PCA test excess reconstruction risk removed, so PCA is zero and larger is better. DA denotes diagonal attenuation and MT denotes masked-target training. The best method in each data–budget cell is bold.

	Optdigits patches		Phoneme		HAR	
Method	$N = 64$	$N = 128$	$N = 64$	$N = 128$	$N = 64$	$N = 128$
PCA	0.00	0.00	0.00	0.00	0.00	0.00
Mask-derived DA	4.32	4.18	0.84	**0.72**	0.65	0.93
Direct DA	**6.74**	**6.40**	**1.09**	0.70	**0.76**	**1.20**
Correlation PCA	5.84	5.41	−0.14	−0.52	0.36	1.01
HeteroPCA	−25.79	−37.36	0.04	0.53	−13.94	−4.87
Coordinate MT	−1.63	−1.24	0.93	0.28	0.56	0.93
Block MT	−3.58	−1.27	0.00	0.00	0.00	0.00