Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Supervised-Residual Dimension Reduction for High-Dimensional Prediction

Ruijuan Zhong
City University of Hong Kong
Jinfeng Xu
City University of Hong Kong
Xin Chen
chenx8@sustech.edu.cn

Southern University of Science and Technology

Research Article

Keywords:

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11299033/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Supervised–Residual Dimension Reduction for High-Dimensional Prediction

Ruijuan Zhong[1], Jinfeng Xu[1*], Xin Chen[2*]

[1]Department of Biostatistics, City University of Hong Kong, Hong Kong, China.
[2]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China.

*Corresponding author(s). E-mail(s): jinfenxu@cityu.edu.hk; chenx8@sustech.edu.cn;

Abstract

Supervised–Residual Dimension Reduction (SRDR) is proposed for high-dimensional regression and classification, where response-guided and variance-based predictor information may provide complementary predictive information. SRDR first constructs a response-guided block Krylov subspace and then estimates principal directions from the predictor variation remaining after projecting out the entire supervised subspace, rather than applying PCA to the original predictors. The resulting supervised and residual loading spaces are orthogonal, and cross-validation selects the Krylov order and residual dimension, allowing supervised-only, residual-only, or mixed representations. An equivalent formulation based on the $\boldsymbol{n} \times \boldsymbol{n}$ Gram matrix avoids forming a $\boldsymbol{p} \times \boldsymbol{p}$ covariance matrix. We establish exact finite-sample orthogonality, an exact primal–dual equivalence, and deterministic perturbation bounds for the resulting loading space. Simulation studies in regression and classification show that the selected allocation between supervised and residual directions depends on the underlying signal structure, with predictive gains when useful predictor variation remains outside the fitted supervised space. Applications with up to 32,321 predictors demonstrate computational feasibility for dense and sparse data and illustrate the selection of supervised-only and mixed representations.

1 Introduction

High-dimensional data are common in genomics, neuroimaging, and other biomedical applications, where the number of predictors p can far exceed the sample size n (Vogelstein et al. 2021; Shen et al. 2016). When $p > n$, the sample covariance matrix is singular, and methods that require its inverse need modification (Bickel and Levina 2004). Regularization methods, such as the lasso and elastic net, address high-dimensional estimation through coefficient shrinkage and variable selection (Tibshirani 1996; Zou and Hastie 2005). Dimension reduction instead replaces the original predictors with a smaller set of linear combinations, which can be useful when predictive information is distributed across correlated variables.

Principal component analysis (PCA) provides one such reduction by selecting orthogonal eigendirections that successively maximize predictor variance (Pearson 1901; Hotelling 1933). Principal component regression uses a selected number of these components for prediction (Massy 1965). Because their ordering is determined by the predictor covariance alone, the leading principal directions need not coincide with directions associated with the response; components with small

eigenvalues can also be important in regression (Jolliffe 1982). Sparse PCA adapts this variance-based construction by imposing sparsity on the loading directions (Zou et al. 2006).

Sufficient dimension reduction (SDR) provides a complementary response-oriented approach by seeking a low-dimensional subspace that preserves the conditional distribution of the response given the predictors (Li 1991; Cook 2007). Classical methods include sliced inverse regression (SIR) and sliced average variance estimation (SAVE), while subsequent work has developed high-dimensional, sparse, and prediction-oriented extensions of SDR (Zhang and Mai 2019; Zeng et al. 2024). Recent developments have also expanded the methodological scope of SDR through slicing-free approaches and other modern frameworks (Li et al. 2024). The present work does not seek to estimate a sufficient dimension-reduction central subspace. Instead, we target a predictive loading space in which response-guided directions can be complemented by residual predictor variation.

Related response-guided dimension reduction methods have been developed with a more direct emphasis on prediction, among which partial least squares (PLS) is particularly prominent. PLS constructs response-guided components from predictor–response association and is commonly used with strongly correlated predictors (Wold et al. 1984). For a univariate response, the construction is closely related to a Krylov subspace generated by the predictor covariance and the predictor–response cross-covariance (Wold et al. 1984; Helland 1988). SIMPLS provides a formulation for multivariate responses (de Jong 1993), and high-dimensional PLS has been studied when both n and p increase (Cook and Forzani 2019). PLS is also connected to predictor envelopes, defined in linear regression as the smallest reducing subspace of the predictor covariance that contains the regression coefficient space (Cook et al. 2013). Envelope methods extend beyond linear regression (Cook and Zhang 2015), and their connections with principal component regression have led to non-iterative procedures for high-dimensional problems (Zhang et al. 2023).

A working supervised space initialized only by linear predictor–response cross-covariance can omit predictive directions. For example, classes may have equal predictor means but different covariance matrices (Vogelstein et al. 2021). The population cross-covariance between the predictors and centered class contrasts is then zero, as is the Krylov sequence generated from it, although the covariance differences may still permit discrimination. This example motivates retaining predictor variation outside the cross-covariance-based supervised space considered here.

Several methods combine predictor variation with response information. Continuum regression connects principal component regression, PLS, and ordinary least squares (Stone and Brooks 1990); principal covariates regression accounts for predictor variation and response fit (de Jong and Kiers 1992); and linear optimal low-rank projection combines class-mean differences with directions obtained from class-centered covariance estimates (Vogelstein et al. 2021). These methods provide different compromises between response association and predictor covariance structure. A separate issue arises when variance-based directions are added to a fitted supervised space: principal components of the original predictors may overlap with directions already represented by that space. Orthogonalizing those components after estimation does not generally recover the leading eigendirections of the covariance restricted to the orthogonal complement. To retain the leading predictor variation remaining outside that space, the additional directions are therefore defined conditional on the entire fitted supervised space.

Supervised–Residual Dimension Reduction (SRDR) addresses this problem by first estimating a response-guided block Krylov subspace and then applying PCA after projecting the predictors onto its orthogonal complement. The resulting estimator is referred to as SRDR throughout. Categorical responses are represented by centered class contrasts, whereas continuous responses can be used directly or represented by specified response features. Let X_c denote the centered predictor matrix and $\widehat{P}_{\mathrm{sup}}(a)$ the orthogonal projector onto the fitted supervised space. The residual predictor matrix is

$$X_{\mathrm{res}}(a) = X_c\{I_p - \widehat{P}_{\mathrm{sup}}(a)\},$$

and its retained leading right singular vectors associated with positive singular values define the

residual principal subspace. The resulting loading space is

$$\widehat{\mathcal{S}}(a,b) = \widehat{\mathcal{S}}_{\mathrm{sup}}(a) \oplus \widehat{\mathcal{S}}_{\mathrm{res}}(a,b),$$

where a is the Krylov order and b is the maximum number of residual principal directions. Cross-validation selects (a,b) and thereby allows supervised-only, residual-only, and mixed representations. SRDR targets a predictive loading space rather than a sufficient dimension-reduction central subspace.

The analysis establishes exact finite-sample orthogonality of the supervised and residual loading spaces, an exact primal–dual equivalence for the resulting projectors, and deterministic perturbation bounds for loading-space estimation at fixed tuning parameters. The dual formulation uses the $n \times n$ centered Gram matrix and avoids forming or diagonalizing a $p \times p$ covariance matrix. The perturbation analysis traces estimation error in the predictor covariance and predictor–response cross-covariance through the sequential supervised–residual construction.

Simulation studies in regression and classification examine how predictive behavior and the selected supervised–residual allocation change with the signal structure. Applications to dense and sparse datasets with up to 32,321 predictors assess held-out prediction, computational feasibility, and the selection of supervised-only and mixed representations.

Algorithm 1 SRDR embedding for fixed (a,b)

Require: Predictor matrix X, response-feature matrix H_0, Krylov order a, residual budget b, and relative tolerance τ
Ensure: Loading matrix $\widehat{Q}(a,b)$, scores Z, and effective dimension r_{eff}
Center X and H_0 to obtain X_c and H; remove dependent response-feature columns and set $q \leftarrow \mathrm{rank}(H)$
Set the requested nominal dimension $r \leftarrow aq + b$
$\widehat{\Sigma} \leftarrow X_c^\top X_c/n$ and $\widehat{C} \leftarrow X_c^\top H/n$
if $a = 0$ **then**
    Set $\widehat{Q}_{\mathrm{sup}} \leftarrow [\,]$
else
    Form $\widehat{K}_a \leftarrow [\widehat{C}, \widehat{\Sigma}\widehat{C}, \ldots, \widehat{\Sigma}^{a-1}\widehat{C}]$
    Retain its left singular vectors with $s_j > \tau s_1$ as $\widehat{Q}_{\mathrm{sup}}$; then $d_{\mathrm{sup}}(a) \le aq$
end if
$X_{\mathrm{res}} \leftarrow X_c(I_p - \widehat{Q}_{\mathrm{sup}}\widehat{Q}_{\mathrm{sup}}^\top)$
if $b = 0$ **then**
    Set $\widehat{Q}_{\mathrm{res}} \leftarrow [\,]$
else
    Retain at most b leading right singular vectors of X_{res} satisfying $s_j^2(X_{\mathrm{res}})/n > \tau\lambda_1(\widehat{\Sigma})$ as $\widehat{Q}_{\mathrm{res}}$
end if
Re-orthogonalize $\widehat{Q}_{\mathrm{res}}$ against $\widehat{Q}_{\mathrm{sup}}$ for floating-point stability
$\widehat{Q}(a,b) \leftarrow [\widehat{Q}_{\mathrm{sup}}, \widehat{Q}_{\mathrm{res}}]$ and $Z \leftarrow X_c\widehat{Q}(a,b)$
$r_{\mathrm{eff}} \leftarrow \mathrm{ncol}\{\widehat{Q}(a,b)\} \le r$
return $\widehat{Q}(a,b)$, Z, and r_{eff}

2 Methods

2.1 Problem setup

Let $X \in \mathbb{R}^{n\times p}$ contain n observations on p predictors, and let $Y_1, \ldots, Y_n$ be either categorical or continuous responses. Write $\mathbf{1}_n$ for the n-vector of ones and

$$M_n = I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top, \qquad X_c = M_nX.$$

The response is represented by a feature map $h : \mathcal{Y} \to \mathbb{R}^q$. If $H_0 = [h(Y_1), \ldots, h(Y_n)]^\top$, its centered sample matrix is

$$H = M_nH_0 \in \mathbb{R}^{n\times q}.$$

Only the column space of H is relevant. Linearly dependent response features are therefore removed, and the optional normalization $H \leftarrow H(H^\top H/n)^{-1/2}$ can be used on the retained columns. This whitening changes the coordinates of the response representation but not the estimated predictor subspace.

For a K-class response, let $G_Y \in \mathbb{R}^{n\times K}$ be the class indicator matrix. A full-rank contrast representation has the form $H = M_nG_YB$, where $B \in \mathbb{R}^{K\times(K-1)}$ spans the contrast space. This formulation includes baseline coding, effect coding, and orthonormal contrasts. For a scalar continuous response, the default choice is $H = M_n(Y_1, \ldots, Y_n)^\top$, while a prespecified collection of basis functions may be used when nonlinear response features are scientifically motivated.

The objective is to estimate a low-dimensional loading space $\widehat{\mathcal{S}} \subseteq \mathbb{R}^p$ and scores $Z = X_c\widehat{Q}$, where $\widehat{Q}$ is an orthonormal basis of $\widehat{\mathcal{S}}$. The space is constructed to retain response-associated variation together with leading residual predictor variation. It is a predictive loading space; no conditional-independence claim of the form $Y \perp X \mid \widehat{Q}^\top X$ is imposed.

2.2 Overview of the proposed embedding

The estimator has two mutually orthogonal components,

$$\widehat{\mathcal{S}}(a,b) = \widehat{\mathcal{S}}_{\mathrm{sup}}(a) \oplus \widehat{\mathcal{S}}_{\mathrm{res}}(a,b).$$

The supervised component is a block Krylov space generated by the sample predictor covariance and the predictor–response cross-covariance. Its first block contains the empirical mean-contrast directions for categorical responses or the predictor–response covariance directions for continuous responses. Subsequent blocks propagate these initial directions through the predictor covariance, providing partial least squares (PLS)-type extensions. The initial directions and their covariance-propagated extensions jointly span the supervised component. The residual component consists of leading principal directions after the entire supervised component has been removed.

The tuning parameter $a \geq 0$ is the Krylov order, and $b \geq 0$ is the maximum number of residual principal directions. With q retained response features, the requested nominal dimension is $r(a,b) = aq + b$. The convention $a = 0$ gives an empty supervised space, whereas $b = 0$ omits the residual block. Numerical ranks determine the actual fitted dimension $r_{\mathrm{eff}}(a,b) \leq r(a,b)$. Algorithm 1 gives the primal construction; its exact $p \gg n$ dual implementation appears in Section 3.3.

2.3 Supervised block Krylov subspace

Define the sample covariance and predictor–response cross-covariance matrices by

$$\widehat{\Sigma} = \frac{1}{n} X_c^\top X_c \in \mathbb{R}^{p\times p}, \qquad \widehat{C} = \frac{1}{n} X_c^\top H \in \mathbb{R}^{p\times q}.$$

For an integer $a \geq 1$, the supervised block Krylov matrix is

$$\widehat{K}_a = \left[\widehat{C}, \widehat{\Sigma}\widehat{C}, \ldots, \widehat{\Sigma}^{a-1}\widehat{C}\right] \in \mathbb{R}^{p\times aq}.$$

Its column space,

$$\widehat{\mathcal{S}}_{\mathrm{sup}}(a) = \mathrm{col}(\widehat{K}_a),$$

is the supervised loading space of order a. Let $d_{\mathrm{sup}}(a) = \mathrm{rank}_\tau(\widehat{K}_a)$ denote its numerical rank under a fixed relative singular-value tolerance τ, and let $\widehat{Q}_{\mathrm{sup}}(a) \in \mathbb{R}^{p\times d_{\mathrm{sup}}(a)}$ contain the retained left singular vectors. The associated orthogonal projector is

$$\widehat{P}_{\mathrm{sup}}(a) = \widehat{Q}_{\mathrm{sup}}(a)\widehat{Q}_{\mathrm{sup}}(a)^\top.$$

For $a = 0$, both $\widehat{K}_0$ and $\widehat{Q}_{\mathrm{sup}}(0)$ are empty and $\widehat{P}_{\mathrm{sup}}(0) = 0$.

The first block $\widehat{C}$ contains the empirical response-covariance directions. Under categorical contrast coding, these directions span the weighted class-mean contrasts; for a continuous response, they are the usual predictor–response covariance directions. The subsequent blocks apply powers of $\widehat{\Sigma}$ and therefore recover the sequence of supervised directions used by covariance-based latent-component methods. The single Krylov space thus unifies first-order class contrast and higher-order PLS information without introducing two potentially overlapping supervised blocks.

2.4 Residual structural variance subspace

Residual variation is defined after removing the full supervised Krylov space:

$$X_{\mathrm{res}}(a) = X_c\{I_p - \widehat{P}_{\mathrm{sup}}(a)\},$$

with residual covariance

$$\begin{aligned}\widehat{\Sigma}_{\mathrm{res}}(a) &= \frac{1}{n} X_{\mathrm{res}}(a)^\top X_{\mathrm{res}}(a) \\ &= \{I_p - \widehat{P}_{\mathrm{sup}}(a)\}\widehat{\Sigma}\{I_p - \widehat{P}_{\mathrm{sup}}(a)\}.\end{aligned}$$

Let $\widehat{Q}_{\mathrm{res}}(a,b)$ contain the eigenvectors associated with the largest, at most b, residual eigenvalues satisfying

$$\lambda_j\{\widehat{\Sigma}_{\mathrm{res}}(a)\} > \tau\lambda_1(\widehat{\Sigma}).$$

The reference is the largest eigenvalue of the original predictor covariance within the current training fit. If fewer than b eigenvalues pass this

threshold, all qualifying directions are used. Accordingly,

$$d_{\mathrm{res}}(a,b) = \mathrm{rank}\{\widehat{Q}_{\mathrm{res}}(a,b)\} \le b,$$
$$\widehat{P}_{\mathrm{res}}(a,b) = \widehat{Q}_{\mathrm{res}}(a,b)\widehat{Q}_{\mathrm{res}}(a,b)^{\top}.$$

Every retained residual loading belongs to the range of $I_p - \widehat{P}_{\mathrm{sup}}(a)$. Hence the residual block is exactly orthogonal to the entire supervised block, not only to its last Krylov direction.

2.5 Final embedded representation

The final loading matrix and projector are

$$\widehat{Q}(a,b) = \big[\widehat{Q}_{\mathrm{sup}}(a), \widehat{Q}_{\mathrm{res}}(a,b)\big],$$
$$\widehat{P}(a,b) = \widehat{Q}(a,b)\widehat{Q}(a,b)^{\top}.$$

Because the two blocks are constructed in orthogonal loading spaces, $\widehat{Q}(a,b)$ already has orthonormal columns and no post hoc orthogonalization is required. Its effective dimension is

$$\begin{aligned} r_{\mathrm{eff}}(a,b) &= d_{\mathrm{sup}}(a) + d_{\mathrm{res}}(a,b) \\ &\le r(a,b) = aq + b. \end{aligned}$$

The inequality can be strict after numerical-rank truncation. The training scores, with $r_{\mathrm{eff}}(a,b)$ columns, are

$$Z = X_c\widehat{Q}(a,b) = \big[Z_{\mathrm{sup}}, Z_{\mathrm{res}}\big].$$

For a new observation x, the same training mean and loading matrix are used to form $(x - \bar{x}_{\mathrm{tr}})^{\top}\widehat{Q}(a,b)$. Orthogonality of the loading blocks is a geometric property in predictor space; it does not require the two score blocks to be empirically uncorrelated.

2.6 Prediction in the embedded space

The embedding can be followed by any predictor suitable for the response type. For a training–test split, let $\widehat{Q}_{\mathrm{tr}}(a,b)$ be estimated using the training observations only. The centered training and test matrices are projected as

$$Z_{\mathrm{tr}} = X_{c,\mathrm{tr}}\widehat{Q}_{\mathrm{tr}}(a,b), \qquad Z_{\mathrm{te}} = X_{c,\mathrm{te}}\widehat{Q}_{\mathrm{tr}}(a,b).$$

A prediction rule g is fitted to $(Z_{\mathrm{tr}}, Y_{\mathrm{tr}})$ and evaluated on Z_{te}:

$$\widehat{Y}_{\mathrm{te}} = g\{Z_{\mathrm{te}}; Z_{\mathrm{tr}}, Y_{\mathrm{tr}}\}.$$

Both score matrices have the $r_{\mathrm{eff}}(a,b)$ columns retained by the training fit. The downstream learner uses these columns, with any prescribed feature expansion applied to the retained scores. Linear or regularized discriminant analysis is a natural choice for a categorical response. Linear regression, ridge regression, or another low-dimensional regression rule may be used for a continuous response. The subspace estimator itself is not tied to a particular downstream model.

2.7 Cross-validation and dimension selection

The embedding depends on the Krylov order a and residual budget b. Candidate models may be indexed by triplets (a,b,r) subject to the nominal identity $r = aq + b$. This requested dimension indexes the tuning grid; the effective fitted dimension is $r_{\mathrm{eff}}(a,b) = d_{\mathrm{sup}}(a) + d_{\mathrm{res}}(a,b) \le r$. Numerical-rank truncation is applied separately within every fit, so the same candidate can have different effective dimensions across folds.

Let

$$\{(\mathcal{R}_j, \mathcal{T}_j)\}_{j=1}^{k}$$

denote the k cross-validation splits. For each fold j, centering, response encoding, the embedding, and the downstream predictor are estimated using only $\mathcal{R}_j$. They are then applied to the held-out set $\mathcal{T}_j$.

For classification, a possible fold-wise loss is the misclassification error

$$\widehat{L}_j^{\mathrm{cls}}(a,b) = 1 - \frac{1}{|\mathcal{T}_j|}\sum_{i\in\mathcal{T}_j} I(\widehat{Y}_i = Y_i),$$

whereas for a continuous response the squared-error loss is

$$\widehat{L}_j^{\mathrm{reg}}(a,b) = \frac{1}{|\mathcal{T}_j|}\sum_{i\in\mathcal{T}_j} (Y_i - \widehat{Y}_i)^2.$$

For a task-specific loss $\widehat{L}_j$, define

$$\mathrm{CV}(a,b) = k^{-1}\sum_{j=1}^{k}\widehat{L}_j(a,b).$$

Only candidates successfully evaluated in every fold are eligible. For SRDR, ties are resolved by the smaller mean fold-specific effective dimension, then the smaller a, and then the smaller b. The final embedding and predictor are refitted on the full training sample, with the effective dimension recomputed in that fit. Stability or oracle properties of this selection step are not asserted in the present theoretical development.

3 Theoretical and Computational Properties

3.1 Population targets and sample estimators

Population quantities for the two-block loading space are expressed for a centered predictor vector $X \in \mathbb{R}^p$ and a centered response feature $h(Y) \in \mathbb{R}^q$ satisfying $\mathbb{E}(X) = 0$ and $\mathbb{E}\{h(Y)\} = 0$. Define

$$\Sigma = \mathbb{E}(XX^\top), \qquad C = \mathbb{E}\{Xh(Y)^\top\}.$$

Motivated by the Krylov representation of PLS (Helland 1988), for $a \geq 1$ define the population block-Krylov matrix and supervised space by

$$K_a = [C, \Sigma C, \ldots, \Sigma^{a-1}C], \qquad \mathcal{S}_{\mathrm{sup}}(a) = \mathrm{col}(K_a).$$

Let $d^0_{\mathrm{sup}}(a) = \mathrm{rank}(K_a)$, $Q_{\mathrm{sup}}(a)$ be any orthonormal basis of this space, and

$$P_{\mathrm{sup}}(a) = Q_{\mathrm{sup}}(a)Q_{\mathrm{sup}}(a)^\top.$$

For $a = 0$, the space is $\{0\}$ and $P_{\mathrm{sup}}(0) = 0$.

The population residual covariance is

$$\Sigma_{\mathrm{res}}(a) = \{I_p - P_{\mathrm{sup}}(a)\}\Sigma\{I_p - P_{\mathrm{sup}}(a)\}.$$

For $b \geq 1$, let $Q_{\mathrm{res}}(a,b)$ contain orthonormal eigenvectors associated with the b largest positive eigenvalues of $\Sigma_{\mathrm{res}}(a)$, or all positive-eigenvalue eigenvectors if fewer than b exist. When truncation is required, the results below assume a positive eigengap at the truncation boundary. Write

$$P_{\mathrm{res}}(a,b) = Q_{\mathrm{res}}(a,b)Q_{\mathrm{res}}(a,b)^\top,$$
$$d^0_{\mathrm{res}}(a,b) = \mathrm{rank}\{P_{\mathrm{res}}(a,b)\}.$$

The population target and its projector are the orthogonal direct sum

$$\mathcal{S}(a,b) = \mathcal{S}_{\mathrm{sup}}(a) \oplus \mathcal{S}_{\mathrm{res}}(a,b),$$
$$P(a,b) = P_{\mathrm{sup}}(a) + P_{\mathrm{res}}(a,b).$$

When $b = 0$, the residual space and projector are defined to be zero.

The sample estimator replaces (Σ, C) by $(\widehat{\Sigma}, \widehat{C})$ and applies the same construction. To make the finite-sample rank convention explicit, suppose a matrix A has singular values $s_1(A) \geq s_2(A) \geq \cdots$. For a fixed relative tolerance $\tau \in (0,1)$, define

$$\mathrm{rank}_\tau(A) = \#\{j : s_j(A) > \tau s_1(A)\},$$

with rank zero when $A = 0$. The columns of $\widehat{Q}_{\mathrm{sup}}(a)$ are the left singular vectors retained by this rule. Up to b residual eigenvectors are retained using

$$\lambda_j\{\widehat{\Sigma}_{\mathrm{res}}(a)\} > \tau\lambda_1(\widehat{\Sigma}).$$

These conventions determine the effective fitted dimension $r_{\mathrm{eff}}(a,b) \leq r(a,b)$, with $r(a,b) = aq + b$ the nominal budget.

The projectors, rather than particular basis matrices, are the statistical estimands. A basis can be rotated within its column space without changing scores up to an invertible orthogonal coordinate transformation. At a residual eigenvalue tied across the truncation boundary, a b-dimensional leading eigenspace is not unique. The results below therefore assume a positive boundary eigengap whenever truncation is required.

Remark 3.1 (Scope of the target). The space $\mathcal{S}(a,b)$ is a predictive supervised–residual loading space. It is not asserted to be a sufficient dimension-reduction central subspace, because the construction does not require $Y \perp X \mid P(a,b)X$. Moreover, Euclidean loading orthogonality gives $Q_{\mathrm{sup}}^\top Q_{\mathrm{res}} = 0$, but generally $\mathrm{Cov}(Q_{\mathrm{sup}}^\top X, Q_{\mathrm{res}}^\top X) = Q_{\mathrm{sup}}^\top \Sigma Q_{\mathrm{res}} \neq 0$.

3.2 Exact geometry

The first main result establishes orthogonality between the supervised and residual loading spaces in both the population and sample constructions. We present the sample version below, using hats to denote sample quantities. The orthogonality holds exactly in finite samples.

Proposition 3.2 (Strict supervised–residual orthogonality)**.** *For every $a, b \geq 0$, the retained loading matrices satisfy*

$$\widehat{Q}_{\mathrm{sup}}(a)^\top \widehat{Q}_{\mathrm{res}}(a,b) = 0,$$
$$\widehat{Q}(a,b)^\top \widehat{Q}(a,b) = I_{r_{\mathrm{eff}}(a,b)}.$$

Consequently,

$$\widehat{P}(a,b) = \widehat{P}_{\mathrm{sup}}(a) + \widehat{P}_{\mathrm{res}}(a,b)$$

is an orthogonal projector and $\widehat{\mathcal{S}}(a,b) = \widehat{\mathcal{S}}_{\mathrm{sup}}(a) \oplus \widehat{\mathcal{S}}_{\mathrm{res}}(a,b)$.

This orthogonality follows from the definition of the residual covariance, rather than from orthogonalizing two separately constructed sets of directions. The residual covariance is defined on the orthogonal complement of the *entire* supervised Krylov space, so the retained residual directions capture predictor variation outside that space. The budget b therefore has a conditional interpretation: it limits the number of additional directions retained after the supervised block has been removed. The proof is given in Appendix C, Section C.1.

Further geometric properties are collected in Appendix C.2, together with their precise statements, qualifications, and proofs. Under the conditions stated there, the projector is invariant under nonsingular changes of response coordinates (Proposition C.1 and Corollary C.2), and its effective dimension is bounded by the sample row-space rank (Proposition C.3). The supervised spaces are nested in a, while the residual spaces are nested in b for fixed a (Proposition C.6). The final spaces need not be nested in a, however, because changing a also changes the residual covariance.

3.3 Primal–dual equivalence and computation

Directly forming a $p \times p$ covariance matrix is unnecessary. An algebraically equivalent representation uses matrices of order n and is preferable when $p \gg n$. Define the centered Gram matrix and the dual Krylov matrix by

$$G = \frac{1}{n} X_c X_c^\top \in \mathbb{R}^{n \times n},$$
$$U_a = [H, GH, \ldots, G^{a-1}H] \in \mathbb{R}^{n \times aq}.$$

The dual Krylov identities

$$\widehat{\Sigma}^j \widehat{C} = \frac{1}{n} X_c^\top G^j H \qquad (j \geq 0),$$

and hence

$$\widehat{K}_a = \frac{1}{n} X_c^\top U_a, \qquad \widehat{K}_a^\top \widehat{K}_a = \frac{1}{n} U_a^\top G U_a.$$

These identities are established in Lemma C.8 of Appendix C.3.

Let the positive retained eigendecomposition of the small matrix in the preceding display be

$$\frac{1}{n} U_a^\top G U_a = V D V^\top,$$
$$D = \mathrm{diag}(d_1, \ldots, d_{\mathrm{sup}}), \qquad d_j > 0.$$

Then a primal orthonormal basis and its training-score matrix are recovered as

$$\widehat{Q}_{\mathrm{sup}} = \widehat{K}_a V D^{-1/2} = \frac{1}{n} X_c^\top U_a V D^{-1/2},$$
$$T_{\mathrm{sup}} = X_c \widehat{Q}_{\mathrm{sup}} = G U_a V D^{-1/2}.$$

These formulas require only an $aq \times aq$ eigendecomposition after the Gram matrix has been formed.

For the residual block, put $A = I_p - \widehat{P}_{\mathrm{sup}}(a)$ and recall that $X_{\mathrm{res}} = X_c A$. Its observation-space Gram matrix is

$$G_{\mathrm{res}} = \frac{1}{n} X_{\mathrm{res}} X_{\mathrm{res}}^\top = G - \frac{1}{n} T_{\mathrm{sup}} T_{\mathrm{sup}}^\top.$$

If $G_{\mathrm{res}} u_j = \lambda_j u_j$ with $\lambda_j > 0$ and $\|u_j\|_2 = 1$, the corresponding residual loading and score vector

are

$$q_{\mathrm{res},j} = \frac{X_{\mathrm{res}}^{\top} u_j}{(n\lambda_j)^{1/2}}, \qquad X_c q_{\mathrm{res},j} = (n\lambda_j)^{1/2} u_j.$$

The second equality uses $q_{\mathrm{res},j} \in \mathrm{col}(A)$, so that $X_c q_{\mathrm{res},j} = X_{\mathrm{res}} q_{\mathrm{res},j}$.

Theorem 3.3 (Exact primal–dual equivalence)**.** *Assume that X_c and H are centered, that the same positive singular-value and eigenvalue thresholds are used in both implementations, and that residual directions are retained only for positive eigenvalues. Then the primal construction based on $(\widehat{\Sigma}, \widehat{C})$ and the dual construction based on (G, U_a) yield identical supervised, residual, and final projectors. Their training-score spaces are also identical. Bases may differ only by signs or by orthogonal rotations within repeated singular-value or eigenvalue subspaces.*

This equivalence is exact rather than an approximation introduced for scalability. It means that the observation-space implementation changes the size of the matrices being diagonalized without changing the statistical estimand. In particular, any difference between the two routes can arise only from numerical thresholding or the arbitrary orientation of a tied eigenspace, not from a different definition of the loading space. The proof is given in Appendix C.3.

Remark 3.4 (Computational cost). Forming G costs $O(n^2 p)$ operations. Generating U_a by recurrence costs $O(an^2 q)$, forming its small Gram matrix costs at most $O\{n(aq)^2\}$ after the required products are available, and its eigendecomposition costs $O\{(aq)^3\}$. A full eigendecomposition of G_{res} costs $O(n^3)$, while recovering r_{eff} primal loadings costs $O(npr_{\mathrm{eff}})$. The leading memory requirement is

$$O\{n^2 + naq + pr_{\mathrm{eff}}\}.$$

The dual route avoids storing or diagonalizing a $p \times p$ covariance matrix. Projection of external observations still requires either the recovered p-dimensional loadings or test–training cross-products with X_c.

3.4 Perturbation of the estimated loading space

The preceding results are exact. Errors in the covariance and cross-covariance operators propagate to the estimated loading space according to the deterministic bounds below. The argument combines a direct bound for the block-Krylov matrix with classical singular and eigenspace perturbation theory (Davis and Kahan 1970; Wedin 1972; Yu et al. 2015). Our result tracks these errors through the sequential supervised–residual construction, while separating its geometry from distribution-specific concentration arguments.

For fixed $a \geq 1$ and $b \geq 1$, define

$$\varepsilon_\Sigma = \|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}, \qquad \varepsilon_C = \|\widehat{C} - C\|_{\mathrm{op}},$$

and suppose

$$\max\{\|\Sigma\|_{\mathrm{op}}, \|\widehat{\Sigma}\|_{\mathrm{op}}\} \leq L.$$

Let $d^0_{\mathrm{sup}} = \mathrm{rank}(K_a)$ and define the supervised separation

$$\delta_{\mathrm{sup}} = s_{d^0_{\mathrm{sup}}}(K_a) > 0,$$

the smallest nonzero singular value of K_a. Let $\theta_1 \geq \cdots \geq \theta_p \geq 0$ be the eigenvalues of $\Sigma_{\mathrm{res}}(a)$, let $d^0_{\mathrm{res}} = d^0_{\mathrm{res}}(a, b)$ be the retained residual rank, and define

$$\delta_{\mathrm{res}} = \theta_{d^0_{\mathrm{res}}} - \theta_{d^0_{\mathrm{res}}+1} > 0,$$

where $\theta_{p+1} = 0$. This condition excludes a tie across the retained residual boundary. We assume that the sample rank rule retains the same supervised and residual ranks as the population construction. The assumption is necessary because a target projector of fixed dimension cannot be continuously identified at a zero singular value or a zero eigengap. The bounds below assume $d^0_{\mathrm{sup}}, d^0_{\mathrm{res}} \geq 1$; if a block is empty, its projector error is zero and the corresponding part of the argument is omitted.

For $a \geq 1$, define

$$A_a(L) = \left(\sum_{j=0}^{a-1} L^{2j} \right)^{1/2},$$

$$B_a(L, C) = \|C\|_{\mathrm{op}} \left(\sum_{j=1}^{a-1} j^2 L^{2j-2} \right)^{1/2},$$

where the second sum is zero when $a = 1$. Then Lemma C.9 shows that the Krylov-matrix error is bounded by

$$\varepsilon_K = A_a(L)\varepsilon_C + B_a(L, C)\varepsilon_\Sigma, \qquad \eta_{\mathrm{sup}} = \frac{2\varepsilon_K}{\delta_{\mathrm{sup}}}.$$

Theorem 3.5 (Sample loading-space error)**.** *Suppose $\varepsilon_K < \delta_{\mathrm{sup}}/2$ and the supervised rank is stable. Then*

$$\|\widehat{P}_{\mathrm{sup}}(a) - P_{\mathrm{sup}}(a)\|_{\mathrm{op}} \leq \eta_{\mathrm{sup}}.$$

Furthermore, define

$$\varepsilon_{\mathrm{res}} = \varepsilon_\Sigma + 2\|\Sigma\|_{\mathrm{op}}\eta_{\mathrm{sup}}.$$

If $\varepsilon_{\mathrm{res}} < \delta_{\mathrm{res}}/2$ and the residual rank is stable, then

$$\|\widehat{P}_{\mathrm{res}}(a, b) - P_{\mathrm{res}}(a, b)\|_{\mathrm{op}} \leq \frac{2\varepsilon_{\mathrm{res}}}{\delta_{\mathrm{res}}},$$

and the final projector obeys

$$\|\widehat{P}(a, b) - P(a, b)\|_{\mathrm{op}} \leq \eta_{\mathrm{sup}} + \frac{2\varepsilon_{\mathrm{res}}}{\delta_{\mathrm{res}}}.$$

The projector norm in Theorem 3.5 is the largest sine of the principal angles between the estimated and population loading spaces. If a Frobenius loss is preferred and both final projectors have rank r_0, then

$$\|\widehat{P} - P\|_{\mathrm{F}} \leq \sqrt{2r_0}\|\widehat{P} - P\|_{\mathrm{op}}.$$

The bound exposes the sequential nature of the estimator. Error in $\widehat{C}$ and $\widehat{\Sigma}$ first perturbs the full supervised Krylov space through $\varepsilon_K/\delta_{\mathrm{sup}}$. Because the residual covariance is defined conditionally on that estimated space, the same error then enters the residual block through $2\|\Sigma\|_{\mathrm{op}}\eta_{\mathrm{sup}}$, in addition to the direct covariance error. The two separation parameters therefore have distinct roles: δ_{sup} identifies the response-guided space, whereas δ_{res} identifies which remaining structural directions are retained. The technical Krylov lemma and the proof of the theorem are given in Appendix C.4.

Remark 3.6 (What the bound does and does not imply when $p \gg n$). Theorem 3.5 is deterministic. Consistency follows only under a probabilistic model for which ε_C and ε_Σ vanish relative to the two separation parameters. For an unrestricted dense covariance matrix, operator-norm covariance error need not vanish when p is much larger than n; the displayed result alone does not claim otherwise. Under a Gaussian model, for example, effective-rank conditions yield dimension-adaptive operator-norm bounds for the sample covariance (Koltchinskii and Lounici 2017). Spiked, sparse, or regularized covariance and cross-covariance variants require their own structural assumptions, and the resulting rate depends on those assumptions.

Appendix C.5 gives a regression prediction-risk bound in terms of the projector error. The result assumes a linear model with Gaussian predictors and independent samples for estimating the loading space and fitting the regression coefficients. Because it requires stronger assumptions than the three main results above, we present it as a supplementary result. We do not derive a corresponding classification-risk bound, which would require additional assumptions on class margins and the fitted classifier.

4 Simulation Studies

The simulations examine when residual principal directions add predictive information beyond a fitted response-associated subspace and how the selected supervised–residual allocation changes with the data-generating mechanism. The classification experiments include binary and three-class responses with mixture, class-covariance, and latent suppressor structure. The continuous-response experiments vary the proportion of signal represented by a truncated response subspace and include a partially nonlinear mechanism.

4.1 Simulation design and evaluation protocol

All experiments use $p = 1000$ predictors, $n = 100$ training observations, 100 Monte Carlo replications, and an independently generated test sample of size 1000. Tuning is based on 10-fold cross-validation performed entirely within the training sample. Each selected model is refitted on all training observations and evaluated once on the independent test sample. PCA, PLS, and SRDR are compared for both response types; LOL is also included in the classification experiments. Candidate nominal dimensions are 1–50 for classification and 1–20 for continuous responses. SRDR jointly selects the supervised Krylov order a and the residual budget b, subject to the relevant nominal-dimension constraint. In classification, $r = a(K - 1)+b$. Appendix A.1 gives the full candidate grids, tie-breaking rules, numerical-rank convention, and fold-specific fitting procedure.

4.2 Classification experiments

Four scenarios were fixed before the reported Monte Carlo runs. The two QDA designs have equal class means. One distinguishes the classes through XOR mixture geometry, whereas the other reverses a diagonal covariance spectrum between the classes. Centered indicator coding therefore gives $C = \mathrm{Cov}(X, H) = 0$, and the population PLS/Krylov sequence has no first-order starting direction. In contrast, a variance-based subspace may retain directions along which the mixture or covariance structure differs. The remaining scenarios are three-class and binary latent suppressor designs analyzed by LDA. Their correlated supervised and complementary coordinates assess whether response-associated directions and residual predictor variation can both contribute under a common covariance classifier. Every embedding method uses the same downstream classifier within a scenario. For SRDR, the requested nominal dimension is $r = a(K - 1) + b$. Appendix A.1 gives the complete data-generating specifications.

All methods use the same stratified 10-fold splits. Each independent test sample contained $n_{\mathrm{te}} = 1000$ observations generated separately from its corresponding training sample. Centering, response coding, embedding estimation, numerical-rank decisions, and classifier fitting are repeated within each fold. The selected configuration is then refitted on all training observations. The test labels were never used for dimension selection, embedding estimation, classifier fitting, or refitting. A separate population-oracle QDA calculation examines complementary information in known supervised and residual subspaces under hybrid mean–covariance populations; Appendix A.2 gives its design and results.

The independent-test comparisons depend on the classification mechanism (Table 1). In the two equal-mean QDA designs, SRDR has the lowest mean error, improving on PCA by approximately 0.042 for the XOR mixture and 0.025 for reversed diagonal covariance. PLS remains near the 0.5 chance benchmark because the linear response-feature map provides no population cross-covariance seed. This is a limitation of the response representation used here; nonlinear response or predictor features would define a different supervised subspace. In the two latent suppressor designs, PLS has the lowest mean error, although the three-class difference between PLS and SRDR is small at the reported Monte Carlo resolution. Paired comparisons are given in Appendix A.2.

The dimension-wise CV profiles show the same mechanism-dependent pattern (Figure 1). For each requested nominal dimension r, the SRDR value is first profiled over admissible (a, b) pairs within a replication and then averaged across replications. Its separation from the baseline profiles is most pronounced in the two QDA designs and is smaller in the LDA suppressor designs.

SRDR compares several (a, b) allocations at each nominal dimension r; these allocations may have different effective dimensions. The resulting CV profiles are tuning diagnostics, not unbiased estimates of test risk. We therefore use the independent-test results for predictive comparisons. Selected allocations and paired Monte Carlo differences are reported in Appendix A.2.

Table 1 Independent-test classification error over 100 Monte Carlo replications. Entries are means with Monte Carlo standard errors in parentheses; lower values are better. The smallest mean within each scenario is shown in bold. All methods use the classifier listed for that scenario, and each test sample contains 1000 observations.

Scenario	Classifier	PCA	PLS	LOL	SRDR
XOR mixture	QDA	0.09672 (0.00347)	0.50374 (0.00153)	0.40734 (0.00465)	**0.05486 (0.00179)**
Equal means, reversed diagonal covariance	QDA	0.11206 (0.00305)	0.49856 (0.00149)	0.31665 (0.00422)	**0.08708 (0.00211)**
Three-class mixed suppressor	LDA	0.51344 (0.00267)	**0.49454 (0.00380)**	0.50723 (0.00211)	0.49803 (0.00260)
Binary mixed suppressor	LDA	0.35327 (0.00238)	**0.33215 (0.00222)**	0.35217 (0.00196)	0.34347 (0.00212)

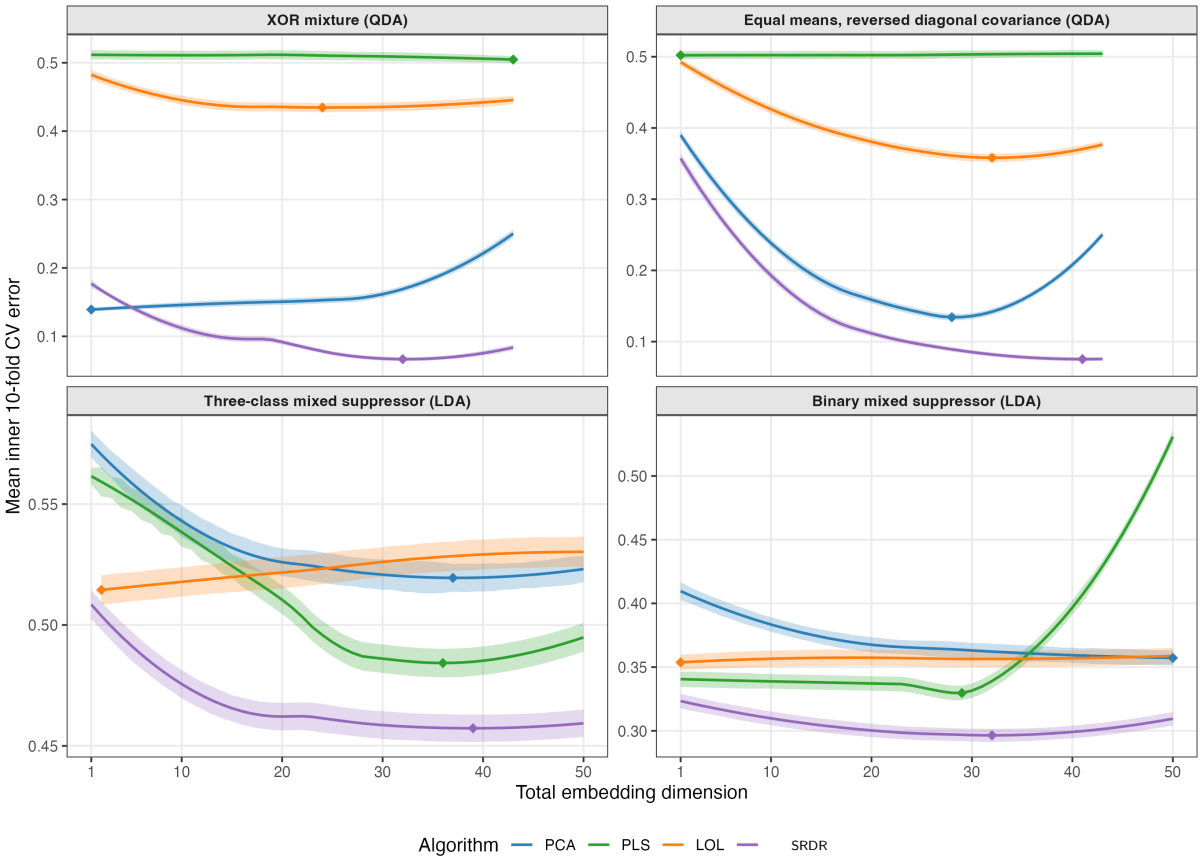


Figure 1 Aggregated inner 10-fold cross-validation error profiles across 100 Monte Carlo replications. Curves are quadratic LOESS summaries with `span`=0.9; ribbons show one Monte Carlo standard error around the smoothed means, and diamonds mark the minima of the displayed smooths. Smoothing is graphical only and does not change the discrete candidates selected within a replication. The horizontal axis denotes the requested nominal dimension r. In the QDA panels, all four methods use the common conservative cap $r \leq 43$, although the axis shows the range 1–50; covariance fitting uses the actual retained scores. In the three-class LDA panel, LOL is unavailable at $r = 1$ because its intrinsic supervised rank is two.

4.3 Continuous-response experiments

The four scenarios use a 24-factor predictor model, a 20-dimensional response, and a six-dimensional signal-bearing response subspace. The response features are obtained by principal response reduction (PRR) (Yoo and Cook 2008)(Yoo 2018). In practical terms, PRR replaces the 20-dimensional response vector by a small number of linear combinations that capture its dominant sample variation. The primary analysis retains $h = 2$ fold-estimated response components for both PLS and SRDR, thereby omitting four signal-bearing response directions. The three linear scenarios set $\eta \in \{0.4, 0.6, 0.8\}$ to vary the fraction of response signal associated with the retained two-dimensional subspace. The fourth scenario, with

$\eta = 0.4$, combines two linear components with four nonlinear functions of higher-variance predictor factors. These settings examine the contribution of residual predictor directions when the working supervised subspace is incomplete or, in the nonlinear case, mismatched to the signal. Appendix A.1 gives the complete factor model, spectra, and signal functions.

PCA, PLS, and SRDR use the same 10-fold splits; LOL is omitted because the available LOL–LDA pipeline is classification-specific. With $h = 2$, PCA and SRDR are evaluated at nominal dimensions 1–20, whereas block PLS is available at the even nominal dimensions. Within every training fold, we eigendecomposed the fold-specific sample response covariance matrix and retained the leading $h = 2$ eigenvectors; the corresponding scores were whitened to form H_2. Both PLS and SRDR used this same H_2, so this is not a comparison against PLS supplied with the full six-dimensional signal-bearing response subspace. The retained response-reduction dimension is $h = 2$, and the requested nominal SRDR dimension is $r = 2a + b$. Ordinary least squares is used in the three linear scenarios. The linear–nonlinear scenario uses multivariate ridge regression on standardized scores, squared scores, and pairwise score interactions; its penalty is selected jointly with the embedding. Response reduction, preprocessing, embedding estimation, and regression tuning are repeated within each training fold. A paired $h = 6$ analysis provides an oracle-dimension sensitivity check while retaining the same fold-specific estimation and training-only tuning protocol. Full tuning details are provided in Appendix A.1.

In the three linear scenarios, the training-CV profiles generally decrease toward the upper end of the prespecified nominal-dimension range (Figure 2). The profiles therefore do not locate an interior optimum for every method. At a given nominal dimension r, SRDR is profiled over admissible (a, b) allocations, and the linear–nonlinear scenario also profiles over the ridge penalty. PLS is evaluated only at its feasible even nominal dimensions.

SRDR has the lowest mean training-CV RMSE across the three linear panels, but its separation from PLS decreases as η increases and more response signal is assigned to the retained two-dimensional subspace. In the linear–nonlinear panel, the low-RMSE region occurs at lower nominal dimensions for PCA and SRDR, whereas the PLS curve continues toward the upper boundary. These curves describe training-stage tuning behavior only. Table 2 reports the selected CV minima, independent-test RMSEs, and SRDR allocations; paired comparisons are reported in Appendix A.3.

The $h = 2$ results are conditional on a deliberately incomplete PRR representation. Under this restriction, SRDR has lower mean test RMSE than PLS in all four scenarios. Its improvement over PCA in the three linear settings is modest, ranging from approximately 0.002 to 0.006, and PCA has slightly lower mean RMSE in the linear–nonlinear setting. The difference between SRDR and PLS also decreases as η increases, consistent with the larger fraction of signal assigned to the retained response directions.

Retaining all six signal-bearing response components changes the comparison. In the three linear scenarios, the paired $h = 6$ analysis reduces PLS test RMSE by approximately 0.020–0.058 and narrows or reverses its difference from SRDR. SRDR also assigns fewer dimensions to the residual subspace as the supervised response representation becomes more complete. Thus the $h = 2$ results describe compensation for omitted response-associated directions, not evidence that retaining fewer response directions is generally beneficial. The linear–nonlinear results change little at $h = 6$ because the four additional response components have zero linear cross-covariance with X. Complete paired results and allocation changes are reported in Appendix A.4.

Residual principal directions are useful in the equal-mean QDA settings, where the population linear response-associated starting block is zero, and when PRR is restricted to two response components ($h = 2$). The gains are smaller or absent in the LDA suppressor settings and in the linear PRR settings when all six signal-bearing response components are retained ($h = 6$). The selected residual budget also decreases as the supervised representation becomes more complete. These results suggest that the benefit of residual variation depends on the setting. SRDR does not perform best across all settings. Tables 1 and 2 report the independent-test results. Appendix A gives the paired Monte Carlo comparisons.

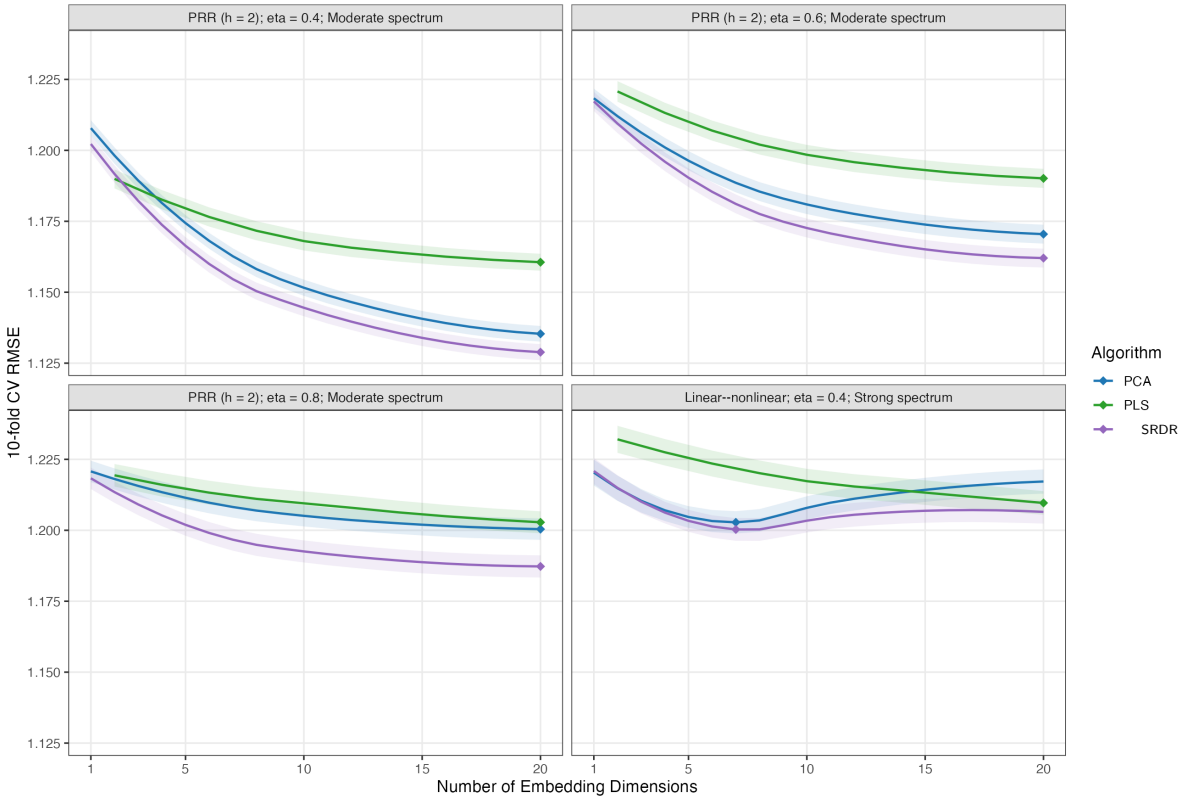


Figure 2 Training-only 10-fold cross-validation RMSE profiles for the four continuous-response scenarios. At each requested nominal dimension r and within each replication, the best eligible candidate is retained before averaging across 100 Monte Carlo replications. Lines are LOESS smooths of the dimension-wise means, ribbons show plus or minus one Monte Carlo standard error, and diamonds mark smoothed minima. At each nominal dimension, the SRDR profile minimizes over admissible (a, b) allocations satisfying $r = 2a + b$; PLS is evaluated only at even nominal dimensions. Effective fitted dimensions may be smaller than r. Here PLS denotes block PLS constructed from the same fold-estimated PRR representation with $h = 2$ used by SRDR. The common vertical scale permits comparison across panels. These curves are tuning diagnostics rather than independent-test risk estimates.

Table 2 Continuous-response performance and selected SRDR dimensions over 100 Monte Carlo replications. Panel A reports the mean replication-specific minimum training-CV RMSE and the mean independent-test RMSE, with Monte Carlo standard errors in parentheses for the test results. CV minima are tuning scores and are not unbiased estimates of test risk. Panel B reports mean selected SRDR dimensions: $r = 2a + b$ is the requested nominal dimension and r_{eff} is the fitted dimension after numerical-rank truncation. Lower RMSE is better.

	Selected training-CV RMSE			Independent-test RMSE (SE)		
Scenario	PCA	PLS	SRDR	PCA	PLS	SRDR
PRR, $h = 2$, $\eta = 0.4$	1.13463	1.15909	**1.12838**	1.12905 (0.00101)	1.15479 (0.00114)	**1.12665 (0.00103)**
PRR, $h = 2$, $\eta = 0.6$	1.16914	1.18765	**1.16128**	1.16692 (0.00108)	1.18774 (0.00121)	**1.16352 (0.00109)**
PRR, $h = 2$, $\eta = 0.8$	1.19703	1.19939	**1.18603**	1.20330 (0.00123)	1.20556 (0.00126)	**1.19733 (0.00127)**
Linear–nonlinear, $\eta = 0.4$	1.19528	1.20849	**1.19356**	**1.18606 (0.00201)**	1.20158 (0.00203)	1.18826 (0.00209)
Panel B: selected SRDR allocation						
Scenario	$\bar{a}$	$\bar{b}$	$\bar{r}$	$\bar{r}_{\text{eff}}$		
PRR, $h = 2$, $\eta = 0.4$	1.37	16.93	19.67	19.64		
PRR, $h = 2$, $\eta = 0.6$	1.16	16.95	19.27	19.19		
PRR, $h = 2$, $\eta = 0.8$	1.98	13.77	17.73	16.85		
Linear–nonlinear, $\eta = 0.4$	1.06	5.12	7.24	6.83		

Appendix B reports six additional high-dimensional binary-classification stress tests. These supplementary simulations examine computational feasibility and allocation behavior across distinct mean and covariance mechanisms; they are not used as primary evidence of predictive superiority.

5 Real Data Applications

Two high-dimensional binary-classification applications were used to examine computational feasibility and held-out predictive performance. GSE80796 provides a dense gene-expression example with moderately imbalanced classes, whereas DEXTER is a sparse text- classification benchmark with balanced official training and validation sets. These analyses illustrate the behavior of the dense and sparse implementations when $p \gg n$; they are not intended to establish broad predictive superiority.

5.1 Datasets

GSE80796.

GSE80796 contains gene-expression profiles from nasal epithelial brushings collected from current and former smokers undergoing evaluation for pulmonary lesions Perez-Rogers et al. (2017). The processed GEO Series Matrix contains 505 subjects measured on 32,321 Affymetrix Human Gene 1.0 ST features. The binary outcome distinguishes 309 subjects diagnosed with lung cancer from 196 controls with benign disease. The downloaded processed expression matrix was used without outcome-guided feature screening. According to the GEO record, the expression values had already been jointly normalized by robust multi-array averaging and corrected for array-batch effects using ComBat. Consequently, the analysis below evaluates the supplied processed data rather than reproducing the upstream microarray preprocessing.

DEXTER.

DEXTER is a sparse bag-of-words benchmark for classifying Reuters articles as being about corporate acquisitions or not Guyon et al. (2004). It contains 20,000 predictors: 9,947 word-stem frequency features and 10,053 randomly generated probe variables. The official training set contains 300 articles (150 positive and 150 negative), and the official labelled validation set contains another 300 articles with the same class balance. The unlabelled 2,000-article test set was not used.

5.2 Analysis and evaluation protocol

Downstream centering, embedding estimation, tuning, and classifier fitting used only the training observations, with these steps repeated within each cross-validation training fold. GSE80796 was analyzed as a dense matrix, whereas DEXTER retained its sparse bag-of-words representation and used the dual computations in Section 3.3. The comparison included PCA, supervised Krylov PLS, LOL, and SRDR, followed in each case by linear discriminant analysis on the embedded scores.

All methods used a common stratified ten-fold partition and the same candidate search ranges. The primary selection criterion was ordinary error for both GSE80796 and DEXTER; balanced error, macro-F1, sensitivity, and specificity were retained as secondary metrics. Parameter selection minimized the unsmoothed mean training-CV error. Details of the fixed DEXTER scaling, the discriminant-rule stabilization, numerical-rank convention, candidate grids, and random-number seeds are provided in Appendix A.5.

GSE80796 has no external validation cohort, so a fixed class-stratified 70/30 split assigned 353 subjects (137 controls and 216 cancer cases) to training and 152 subjects (59 controls and 93 cancer cases) to the test set. All tuning, downstream feature learning, centering, and fitting excluded the held-out observations. DEXTER retained the official split: model and tuning selection used only the 300 training articles, and the officially provided 300-article validation set was not used during model selection. For both datasets, the selected configurations were refitted on the complete training partition and evaluated once after parameter selection and refitting.

The GSE80796 split provides internal rather than external validation. It is not prospectively untouched because the dataset had already been examined in earlier full-sample and hold-out analyses, and GEO reports joint upstream RMA/ComBat preprocessing. The lock therefore concerns the downstream training/test partition

and refitting procedure, not an independently collected or independently preprocessed cohort.

5.3 Results and analysis

Figure 3 summarizes the training-stage tuning diagnostics. The GSE80796 curves use only the locked 353-subject training partition, whereas the DEXTER curves use only the official 300-article training set; no held-out labels enter either set of curves. The profiles summarize how the mean fold error varies with the requested nominal embedding dimension and are not estimates of held-out prediction error.

After tuning and refitting, each method was evaluated once on the locked GSE80796 test set or the official DEXTER validation set (Table 3). Unlike the training-CV curves in Figure 3, these results use held-out observations.

On GSE80796, SRDR achieved the best overall predictive performance among the four methods. Its ordinary error was 0.362 (55 of 152 subjects misclassified), compared with 0.375 for PCA, 0.428 for LOL, and 0.454 for PLS. SRDR also attained the lowest balanced error (0.432), the highest macro-F1 (0.551), and the highest sensitivity (0.882). It correctly identified 82 of the 93 cancer cases, compared with 79 for PCA, 70 for LOL, and 69 for PLS. Thus, the combined metrics favor SRDR, with a gain in the detection of cancer cases.

On DEXTER, SRDR and PLS achieved the joint-best performance, with an ordinary error of 0.067 (20 of 300 articles misclassified), outperforming LOL (0.077; 23 misclassifications) and PCA (0.103; 31 misclassifications). SRDR matched PLS in balanced error (0.067), macro-F1 (0.933), sensitivity (0.973), and specificity (0.893), and retained the strongest aggregate performance among the evaluated methods. Across the two applications, SRDR performed competitively with the strongest comparators, attaining the lowest error on GSE80796 and matching PLS on DEXTER. These results provide evidence of computational feasibility and competitive predictive performance.

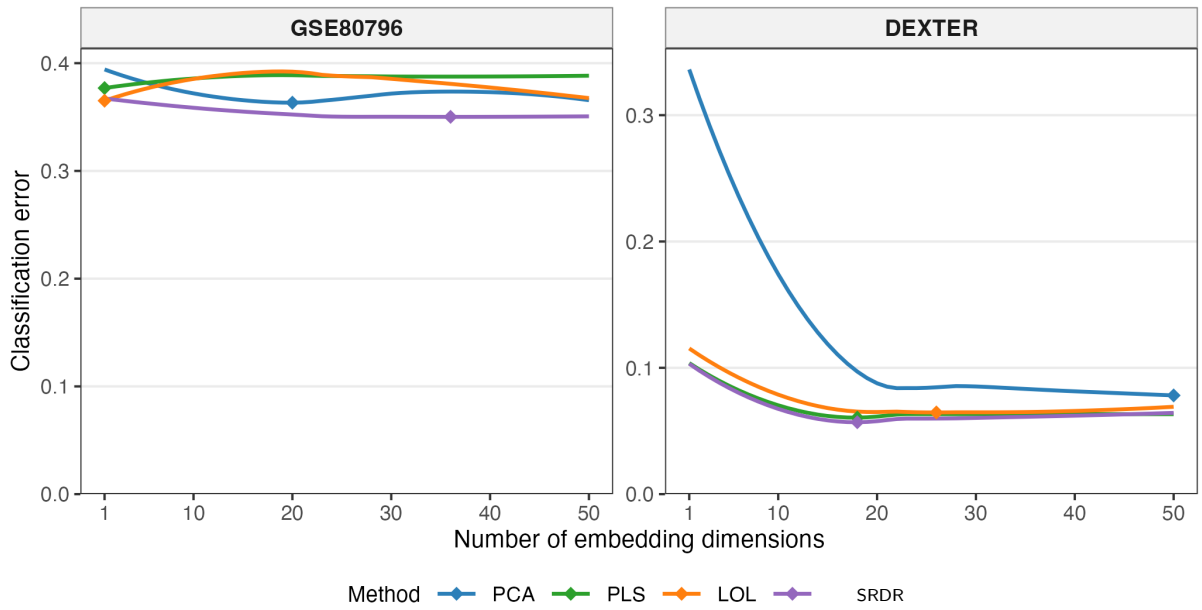


Figure 3 Ten-fold cross-validation error curves for GSE80796 and DEXTER. The horizontal axis shows the requested nominal embedding dimension r. Lines are degree-two LOESS summaries (span 0.9) of the dimension-specific mean fold errors. Diamonds mark the minima of the displayed LOESS summaries and are descriptive; actual parameter selection minimized the unsmoothed mean ten-fold CV error. The vertical scales differ between panels, and both begin at zero.

Table 3 Held-out classification performance in the two real-data applications after training-only ten-fold cross-validation selection by ordinary error for both GSE80796 and DEXTER, followed by refitting on the complete training partition. GSE80796 uses the locked internal test set ($n_{\text{test}} = 152$); DEXTER uses the official labelled validation set ($n_{\text{val}} = 300$). Cancer and articles about corporate acquisitions are the positive classes.

Method	Error	Balanced error	Macro-F1	Sensitivity	Specificity	TN	FP	FN	TP
GSE80796: Locked internal test									
PCA	0.375	0.440	0.547	0.849	0.271	16	43	14	79
PLS	0.454	0.510	0.478	0.742	0.237	14	45	24	69
LOL	0.428	0.480	0.513	0.753	0.288	17	42	23	70
SRDR	0.362	0.432	0.551	0.882	0.254	15	44	11	82
DEXTER: Official validation									
PCA	0.103	0.103	0.896	0.973	0.820	123	27	4	146
PLS	0.067	0.067	0.933	0.973	0.893	134	16	4	146
LOL	0.077	0.077	0.923	0.973	0.873	131	19	4	146
SRDR	0.067	0.067	0.933	0.973	0.893	134	16	4	146

6 Discussion

Principal components of the original predictors can overlap with variation already represented by a supervised embedding. Orthogonalizing the two estimated bases afterwards does not generally yield the leading directions of the residual covariance. SRDR instead projects out the entire supervised Krylov space before estimating residual principal directions. Its residual budget therefore limits the number of directions added beyond the response-guided space. Cross-validation can select a supervised-only, residual-only, or mixed representation.

The supervised and residual loading spaces form an exact finite-sample orthogonal direct sum. Primal–dual equivalence allows the same target to be computed in observation space when $p \gg n$,

without forming a $p \times p$ covariance matrix. The perturbation bound tracks covariance and cross-covariance errors through the supervised Krylov projector and then the residual covariance. It distinguishes two identification conditions: the supervised singular-value separation and the residual eigengap.

The selected allocation varies across the mean, covariance, and nonlinear simulation mechanisms. The clearest gains occur when useful predictor structure remains outside the fitted supervised space. With $h = 6$, PLS receives all six signal-bearing response directions, the SRDR advantage narrows or disappears, and the selected residual budget decreases. The residual block can therefore compensate for information omitted by a restricted response representation; this does not justify withholding available response information from a supervised method. The real-data analyses establish feasibility in dense and sparse $p \gg n$ settings, but the held-out comparisons do not establish a general predictive advantage.

The GSE80796 holdout adds an internal evaluation after training-only selection, but it comprises one locked split with $n_{\text{test}} = 152$. PCA and SRDR had nearly equal balanced error, but different sensitivity–specificity trade-offs. Joint upstream preprocessing, the prior full-sample analysis, and the later harmonization of the selection criterion further limit its independence. It is internal rather than external validation; reuse after further method development means that these observations are not a fresh prospective test set.

DEXTER provides an official labelled validation set that was excluded from training and model selection. SRDR matched the lowest validation error obtained by PLS and had lower error than LOL and PCA on this split. However, this is a single benchmark split and does not establish a general predictive advantage.

The supervised component depends on a prespecified response feature map and linear cross-covariance, so it may miss nonlinear or distributional response associations. The theory treats (a, b) as fixed and does not quantify uncertainty from cross-validation, numerical-rank selection, or the downstream learner. In an unrestricted $p \gg n$ model, operator-norm covariance consistency requires additional structure. Sparse, low-effective-rank, or regularized covariance and cross-covariance operators are possible extensions, alongside nonlinear response features and selection-aware risk theory. Broader independent validation is needed to assess when conditional residual structure yields practically important gains.

Appendix A

This appendix provides paired comparisons, allocation details, sensitivity interpretation, and real-data implementation details accompanying the core numerical results in the main text. Training cross-validation values are tuning diagnostics, whereas independently generated simulation test samples and the labelled DEXTER validation set evaluate fitted selections on observations not used for model selection.

A.1 Simulation data-generating mechanisms and implementation details

Common numerical protocol.

The classification candidate nominal dimensions were $\mathcal{R}_{\text{class}} = \{1, 2, \ldots, 50\}$, with $a_{\max} = b_{\max} = 35$ for SRDR. For a K-class response, the retained contrast rank is $K - 1$ and the requested nominal dimension is $r = a(K - 1) + b$. The continuous-response candidate nominal dimensions were $\mathcal{R}_{\text{cont}} = \{1, 2, \ldots, 20\}$. All numerical-rank decisions used the fixed relative tolerance $\tau = \sqrt{\epsilon_{\text{machine}}} \approx 1.49 \times 10^{-8}$. This tolerance was applied to Krylov-matrix singular values for the supervised block and to residual covariance eigenvalues relative to the largest original predictor covariance eigenvalue for the residual block.

Classification mechanisms.

For the XOR mixture, let $\mathbf{0}$, $\mathbf{1}$, $\mathbf{m}_{10}$, and $\mathbf{m}_{01}$ denote the all-zero, all-one, and two alternating binary location vectors. Their common weighted center is $\mathbf{c} = \frac{1}{2}\mathbf{1}$, and the multiplier $\kappa = 0.5$ gives the four locations

$$\tfrac{1}{4}\mathbf{1}, \quad \tfrac{3}{4}\mathbf{1}, \quad \tfrac{1}{4}\mathbf{1} + \tfrac{1}{2}\mathbf{m}_{10}, \quad \tfrac{1}{4}\mathbf{1} + \tfrac{1}{2}\mathbf{m}_{01}.$$

Conditional on class, the two equal-weight mixtures are

$$\begin{aligned} X \mid Y = 1 &\sim \tfrac{1}{2} N_p\big(\tfrac{1}{4}\mathbf{1}, vI_p\big) \\ &\quad + \tfrac{1}{2} N_p\big(\tfrac{3}{4}\mathbf{1}, vI_p\big), \\ X \mid Y = 2 &\sim \tfrac{1}{2} N_p\big(\tfrac{1}{4}\mathbf{1} + \tfrac{1}{2}\mathbf{m}_{10}, vI_p\big) \\ &\quad + \tfrac{1}{2} N_p\big(\tfrac{1}{4}\mathbf{1} + \tfrac{1}{2}\mathbf{m}_{01}, vI_p\big), \\ v &= \sqrt{p/100}. \end{aligned}$$

Both classes have mean $\frac{1}{2}\mathbf{1}$, so discrimination depends on the mixture geometry rather than a first-order contrast.

In the reversed-covariance design, the two classes share

$$\mu_j = \frac{4}{\sqrt{2(j-1)+1}}, \qquad j = 1, \ldots, p,$$

but have covariance matrices

$$\begin{aligned} \Sigma_1 &= \mathrm{diag}(d_1, \ldots, d_p), \\ \Sigma_2 &= \mathrm{diag}(d_p, \ldots, d_1), \qquad d_j = \frac{100}{\sqrt{p-j+1}}. \end{aligned}$$

All class information is therefore second order.

For the three-class suppressor design, eleven random orthonormal directions in $\mathbb{R}^p$ were partitioned into four supervised, three complementary, and four nuisance directions. The class means form an equilateral two-contrast configuration in the supervised coordinates with $\delta = 2.2$. The common latent covariance uses supervised variances $\lambda = (0.5, 1, 2, 4)$, complementary variances $\nu = (6, 4.5, 3.5)$, and a rotated cross-correlation with $\rho = 0.5$. The complementary coordinates have zero marginal class-mean contrast but remain conditionally relevant through this correlation. The nuisance directions have variance 1.5 and the orthogonal complement has unit variance. The binary suppressor design uses the same covariance structure, two opposite class means, and $\delta = 2.0$.

Classification tuning.

All competing methods used the same stratified 10-fold split in each replication. For fold j, the misclassification error and its cross-fold average were

$$\widehat{L}_j = 1 - \frac{1}{|T_j|} \sum_{i \in T_j} I(\widehat{Y}_i = Y_i),$$

$$\mathrm{CV} = \frac{1}{10} \sum_{j=1}^{10} \widehat{L}_j.$$

PCA, PLS, and LOL selected the requested nominal dimension r by minimizing this criterion, with ties resolved in favor of the smaller nominal dimension. SRDR searched

$$\begin{aligned} \mathcal{A}_K = \{(a,b) : 0 \le a, b \le 35, \\ a(K-1) + b \in \mathcal{R}_{\text{class}}\}. \end{aligned}$$

and retained only configurations having finite loss in every fold. Ties were resolved sequentially by smaller mean fold-specific effective dimension, smaller a, and smaller b. The fitted rank was

$$r_{\text{eff}} = d_{\text{sup}}(a) + d_{\text{res}}(a,b) \le r = a(K-1) + b.$$

Within each training fold, the predictors were centered and the class-indicator response was centered, reduced to its nonzero contrast subspace, and whitened to obtain $H^\top H / n_{\text{tr}} = I_{K-1}$. Writing

$$\widehat{\Sigma}_X = X_c^\top X_c / n_{\text{tr}}, \qquad \widehat{C}_{XH} = X_c^\top H / n_{\text{tr}},$$

the supervised candidate matrix was

$$\mathcal{K}_a = [\widehat{C}_{XH}, \widehat{\Sigma}_X \widehat{C}_{XH}, \ldots, \widehat{\Sigma}_X^{a-1} \widehat{C}_{XH}].$$

A rank-revealing factorization produced $Q_{\text{sup}}(a)$. The residual matrix

$$X_{\text{res}}(a) = X_c \{I_p - Q_{\text{sup}}(a) Q_{\text{sup}}(a)^\top\}$$

was used to estimate at most b residual principal directions, followed by re-orthogonalization against the supervised block. Because $p > n$, these operations used the training-fold Gram matrix $G = X_c X_c^\top / n_{\text{tr}}$ and mapped retained directions back to the predictor space. Each downstream classifier used the actual retained score columns, whose number is denoted by r_{eff} for each method. In the QDA settings, all methods used the common conservative search cap $r \le 43$ on the requested nominal dimension. This implies $r_{\text{eff}} \le 43$, but covariance feasibility still depends on the retained scores within each class; candidates must

have finite loss in every fold. In the three-class LDA setting, the intrinsic LOL rank is two, so its $r = 1$ candidate is unavailable.

Continuous-response mechanisms.

All four settings use a 24-factor predictor model and a six-dimensional response signal. For observation i,

$$Z_i \sim N_{24}(0, I_{24}), \qquad E_i \sim N_p(0, I_p),$$
$$\varepsilon_i \sim N_q(0, \tfrac{1}{2} I_q),$$

independently, and

$$X_i = U\Lambda^{1/2} Z_i + E_i, \qquad Y_i = \sqrt{q}\, A D_\eta^{1/2} s_i + \varepsilon_i,$$

where $U \in \mathbb{R}^{p\times 24}$ and $A \in \mathbb{R}^{q\times 6}$ have orthonormal columns. Fresh U and A were generated in each replication and shared by its training and test samples. The response allocation matrix was

$$D_\eta = \operatorname{diag}\left(\frac{\eta}{2}, \frac{\eta}{2}, \frac{1-\eta}{4}, \frac{1-\eta}{4}, \frac{1-\eta}{4}, \frac{1-\eta}{4}\right).$$

Together with $\operatorname{Var}(\varepsilon_i) = I_q/2$, the $\sqrt{q}$ scaling gives target signal-to-noise ratio two after averaging over response coordinates. The remaining $q - 6 = 14$ response directions contain isotropic noise.

The three linear PRR settings use

$$s_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}, Z_{i5}, Z_{i6})^\top$$

and

$$\operatorname{diag}(\Lambda) = (1, 0.8, 8, 7, 6, 5, \lambda_7, \ldots, \lambda_{24}),$$

where $\lambda_7, \ldots, \lambda_{24}$ are equally spaced from 6 to 2. The values $\eta = 0.4$, 0.6, and 0.8 vary the signal assigned to the first two response directions. The linear–nonlinear setting uses $\eta = 0.4$,

$$s_i = \big(Z_{i1}, Z_{i2}, (Z_{i3}^2 - 1)/\sqrt{2},$$
$$Z_{i4}Z_{i5}, (Z_{i5}^2 - 1)/\sqrt{2}, Z_{i3}Z_{i6}\big)^\top,$$

and the stronger spectrum

$$\operatorname{diag}(\Lambda) = (1, 0.8, 16, 14, 12, 10, \lambda_7, \ldots, \lambda_{24}),$$

with the same nuisance sequence from 6 to 2.

Continuous-response tuning.

Within every training fold, the leading $h = 2$ eigenvectors of the sample response covariance were retained and the corresponding scores were whitened to form the common response feature matrix H_2 used by both PLS and SRDR. Here h is the retained response-feature rank in the general construction, so the requested nominal dimension satisfies $r = ha + b = 2a + b$. PCA used $(a, b) = (0, r)$ for $r \in \mathcal{R}_{\text{cont}}$, PLS used $(a, b) = (r/2, 0)$ for $r = 2, 4, \ldots, 20$, and SRDR searched

$$0 \le a \le 10, \qquad 0 \le b \le 20, \qquad 1 \le 2a + b \le 20.$$

Each regression was based on the actual r_{eff} retained score columns, with the prescribed feature expansion applied in the nonlinear setting. The nominal dimension r indexed the candidate grid. Candidate RMSE pooled squared prediction errors over held-out observations and all 20 response coordinates. Ordinary least squares was used for the linear settings. For the linear–nonlinear setting, standardized scores, squared scores, and pairwise interactions were entered into multivariate ridge regression, with penalty selected from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. The minimizing configuration was refitted on all 100 training observations and evaluated once on the independent test sample.

A.2 Classification simulation results

Table 1 in the main text reports the independent-test mean errors. This section gives the paired comparisons and selected-allocation details that qualify those means.

As a population mechanism check, two Gaussian populations divided predictive information between orthogonal supervised and residual blocks. Both used a four-dimensional supervised block with covariance $\operatorname{diag}(0.5, 1, 2, 4)$ and class-dependent means, together with a zero-mean residual block having class-dependent covariance. The binary population used $\delta = 2.0$ and residual covariances $\operatorname{diag}(12, 4)$ and $\operatorname{diag}(4, 12)$. The three-class population used $\delta = 2.2$ and cyclic residual covariances $\operatorname{diag}(12, 4, 4)$, $\operatorname{diag}(4, 12, 4)$, and $\operatorname{diag}(4, 4, 12)$. Six nuisance directions had variance 3 and the remaining ambient directions had unit variance.

QDA was applied to the known supervised block, known residual block, and their known union. Table 4 reports risks estimated from 10^6 independent observations per population. The combined representation improved on either block alone in both populations, confirming the intended population-level complementarity. These oracle risks do not measure finite-sample subspace estimation accuracy.

The two QDA rows isolate the information carried by the first-order predictor–response cross-covariance. The classes are balanced and have identical population means, so $C = \mathrm{Cov}(X, H)$ is zero under centered indicator coding. Consequently, the population sequence $C, \Sigma C, \ldots$ used by PLS is identically zero; its finite-sample directions arise from estimation noise rather than a stable class-mean contrast. PLS errors of 0.50374 and 0.49856 are therefore near the 0.5 chance benchmark, whereas PCA errors of 0.09672 and 0.11206 show that QDA can use informative covariance or mixture structure retained by an unsupervised projection. Increasing the requested PLS dimension only applies further covariance powers to the same weak starting block and does not introduce second-order response features. This comparison identifies a limitation of the linear cross-covariance construction used here, not of all possible nonlinear PLS extensions.

Paired Monte Carlo comparisons used the best baseline according to the scenario-level mean test error and defined the margin as best-baseline error minus SRDR error, so positive values favor SRDR. The mean margins were 0.04186 for XOR and 0.02498 for equal means with reversed diagonal covariance; their paired 95% Monte Carlo intervals were $[0.03404, 0.04968]$ and $[0.01880, 0.03116]$, respectively. SRDR beat the best baseline in 90 of 100 XOR replications and 78 of 100 reversed-covariance replications. These two QDA designs therefore provide clear evidence that the combined embedding can recover covariance and mixture information missed by purely supervised or purely variance-based reductions.

The LDA results were more qualified. Relative to PLS, the corresponding margins were -0.00349 in the three-class mixed suppressor design and -0.01132 in the binary mixed suppressor design, with intervals $[-0.01095, 0.00397]$ and $[-0.01472, -0.00792]$. Thus the three-class difference was unresolved at this Monte Carlo resolution, whereas PLS had lower test error in the binary design. SRDR nevertheless improved on PCA and LOL in both LDA scenarios. The contrast between the favorable inner-CV profiles and the independent-test table also illustrates why a more flexible profiled search must be evaluated on observations not used during tuning.

The selected SRDR allocations changed with the data-generating mechanism. The mean $(a, b, r, r_{\mathrm{eff}})$ values, with r the nominal dimension and r_{eff} the fitted dimension, were $(14.14, 4.20, 18.34, 5.59)$ for XOR, $(6.95, 25.29, 32.24, 26.53)$ for reversed diagonal covariance, $(8.20, 11.98, 28.38, 17.00)$ for the three-class suppressor, and $(13.13, 10.68, 23.81, 13.68)$ for the binary suppressor. In the reversed-covariance design, the mean residual budget $b = 25.29$ was much larger than the mean Krylov order $a = 6.95$. Together with the PCA and PLS errors of 0.11206 and 0.49856, respectively, this allocation is consistent with the conditional residual PCA block carrying a substantial share of the usable covariance structure after the weak supervised seed has been removed. The allocation alone does not identify a blockwise causal contribution, because no post-selection block-ablation experiment was performed. Both LDA suppressor designs retained material contributions from the two blocks. These patterns support tuning a and b separately rather than fixing their relative allocation.

XOR also illustrates why nominal and effective dimensions must be distinguished. Its mean requested nominal dimension was $\bar{r} = \bar{a} + \bar{b} = 18.34$, but its mean fitted dimension was only $\bar{r}_{\mathrm{eff}} = 5.59$. Because $q = 1$, a counts the number of Krylov powers attempted and is not the supervised dimension. The latter is $d_{\mathrm{sup}}(a) = \mathrm{rank}_\tau(K_a)$, and the zero population starting block C explains why increasing a need not create new stable supervised directions. Since $r_{\mathrm{eff}} = d_{\mathrm{sup}} + d_{\mathrm{res}}$ aggregates rank loss from both blocks, the observed gap cannot be assigned entirely to the Krylov block from these summaries alone. It nevertheless shows that reporting a without the effective rank can substantially overstate the fitted dimension.

The independent-test results qualify the dimension-wise CV patterns shown in the main text. SRDR was strongest in the two QDA mechanisms, whereas its comparison with PLS was

Table 4 Population-oracle QDA misclassification risks for the hybrid mean–covariance populations. The supervised, residual, and combined representations use the corresponding known population subspaces. Lower is better.

Population	Supervised only	Residual only	Combined	Chance benchmark
Binary	0.158487	0.333909	**0.130323**	0.500
Three-class	0.221725	0.459072	**0.184221**	0.667

mixed in the LDA mechanisms. These results support an adaptive combination under complementary structure, not universal classifier dominance.

A.3 Continuous-response simulation results

Table 2 separates the optimized training-CV scores from performance on independently generated test samples. The following paired summaries report the size and consistency of the independent-test differences and the selected SRDR allocations.

The paired independent-test comparisons clarify the size and consistency of these differences. Relative to PLS, the mean SRDR-minus-PLS RMSE differences were -0.02814, -0.02422, and -0.00823 for PRR at $h = 2$ and $\eta = 0.4$, 0.6, and 0.8, respectively, and -0.01332 in the linear–nonlinear scenario. The corresponding SRDR win rates were 100%, 100%, 84%, and 88%. Relative to PCA, SRDR reduced mean RMSE by 0.00239, 0.00340, and 0.00598 in the three linear PRR scenarios, but increased it by 0.00220 in the linear–nonlinear scenario. Paired Monte Carlo 95% intervals for these four SRDR-minus-PCA differences were $[-0.00323, -0.00156]$, $[-0.00443, -0.00237]$, $[-0.00752, -0.00443]$, and $[0.00096, 0.00345]$, respectively. These intervals summarize Monte Carlo variation and are not multiplicity-adjusted inferential intervals.

The selected SRDR allocations also reflected the intended mechanism. Both blocks were selected in 72%, 67%, and 65% of the linear PRR replications at $\eta = 0.4$, 0.6, and 0.8, respectively. Mean residual budgets remained large in these settings, whereas the linear–nonlinear scenario selected a mixed allocation in 34% of replications and a residual-only allocation in the remainder. Together with the boundary behavior of the CV profiles, these results support complementary recovery under an intentionally incomplete PRR representation while also identifying the prespecified dimension range and nonlinear learner as important qualifications.

A.4 Sensitivity to the response-reduction dimension

The continuous-response data-generating mechanism has six signal-bearing population response directions, whereas the primary analysis retains the first two fold-specific response principal components. The other 14 population response directions contain only isotropic noise. The $h = 2$ PRR setting is an intentional incomplete-representation stress test: it asks whether the residual predictor block can recover useful predictor structure when the fitted supervised block is constructed from a deliberately compressed response. It also prevents PLS from using the full six-dimensional signal-bearing response subspace. We therefore repeated the tuning and independent-test evaluation with six fold-specific response principal components. The number six is known from the data-generating mechanism and is therefore an oracle response-reduction dimension, but the response principal directions themselves were re-estimated using only the training portion of every fold; the true loading matrix A was never supplied to either method. The data, train–test splits, CV folds, downstream learners, penalty grid, and maximum nominal dimension were held fixed. The admissible allocation grids followed $r = ha + b$ separately for $h = 2$ and $h = 6$; the effective dimension was determined separately in each fit.

Table 5 summarizes the paired $h = 2$ and $h = 6$ results.

The three linear PRR settings show the anticipated sensitivity. Retaining all six estimated signal-bearing response principal components reduces PLS RMSE by 0.01999–0.05764 and narrows or removes the apparent $h = 2$ advantage of SRDR. SRDR improves as well because its supervised Krylov block uses the richer response representation. Its mean selected residual budgets change from 16.93, 16.95, and 13.77 at $h = 2$ to 2.91, 5.09, and 6.21 at $h = 6$ for $\eta = 0.4$, 0.6, and 0.8, respectively. The method therefore shifts away from the residual block as the supervised response representation becomes more complete.

Table 5 Sensitivity of independent-test RMSE to the retained response-reduction dimension h over 100 paired Monte Carlo replications. The difference is the $h = 6$ minus $h = 2$ RMSE; negative values favor retaining all six estimated signal-bearing response principal components. Parentheses give the Monte Carlo standard error of the paired difference, and the win rate is the proportion of replications with lower RMSE at $h = 6$.

Scenario	Method	$h = 2$	$h = 6$	Paired difference (SE)	$h = 6$ win rate
PRR, $\eta = 0.4$	PLS	1.15479	1.09715	−0.05764 (0.00085)	100%
	SRDR	1.12665	1.09715	−0.02950 (0.00054)	100%
PRR, $\eta = 0.6$	PLS	1.18774	1.14170	−0.04604 (0.00063)	100%
	SRDR	1.16352	1.14235	−0.02117 (0.00062)	97%
PRR, $\eta = 0.8$	PLS	1.20556	1.18557	−0.01999 (0.00049)	100%
	SRDR	1.19733	1.18811	−0.00922 (0.00073)	82%
Linear–nonlinear, $\eta = 0.4$	PLS	1.20158	1.21150	+0.00993 (0.00110)	11%
	SRDR	1.18826	1.18626	−0.00200 (0.00060)	26%

In the linear–nonlinear setting, $h = 6$ slightly worsens PLS and changes SRDR only marginally. The additional signal-bearing response components are nonlinear functions of Gaussian predictor factors and have zero population linear cross-covariance with X; retaining them cannot by itself make a linear supervised Krylov block recover the nonlinear predictor directions. PCA is an exact negative control: its fitted predictor space and test RMSE are identical at $h = 2$ and $h = 6$ in every replication because PCA does not use the response. The primary $h = 2$ experiment is therefore a targeted partial-supervision stress test, while the $h = 6$ analysis is an oracle-dimension benchmark showing that the benefit of the residual block contracts when the supervised response map contains the complete signal-bearing subspace.

A.5 Real-data implementation details

GSE80796 was analyzed as a dense matrix. DEXTER was retained in sparse bag-of-words form and processed using the dual and sparse-matrix operations in Section 3.3. The supplied nonzero DEXTER feature values were divided by a fixed divisor of 1,000 before analysis. This scaling was specified independently of the labels and applied to both official partitions; no validation-derived preprocessing was used.

Each embedding was followed by linear discriminant analysis on its actual r_{eff} retained score columns. If the ordinary LDA covariance calculation was numerically singular, the corresponding ridge-stabilized pooled-covariance discriminant rule was used. All methods shared the same stratified ten-fold partition, generated with seed 20260722. The GSE80796 training/test partition was generated with seed 20260918, and its training-only CV used seed 20260722.

The numerical-rank tolerance and spectral reference convention matched those used in the simulations: $\tau = \sqrt{\epsilon_{\text{machine}}} \approx 1.49 \times 10^{-8}$, applied to Krylov singular values and to residual covariance eigenvalues relative to the largest eigenvalue of the original predictor covariance. The same convention was used within every cross-validation fold and in each full-data refit.

For PCA, PLS, and LOL, the candidate nominal embedding dimension was $r \in \{1, \ldots, 50\}$. For SRDR, the candidate set included every admissible pair (a, b) satisfying $0 \le a \le 35$, $0 \le b \le 35$, and $1 \le r \le 50$. Because both outcomes are binary and the whitened response contrast has rank one, the requested nominal SRDR dimension is $r = a + b$. The effective dimension can be smaller when the supervised Krylov matrix or residual covariance has deficient numerical rank.

Appendix B Supplementary robustness simulations

Supplementary robustness simulations in this appendix comprise six high-dimensional binary-classification stress tests. Their purpose is robustness rather than primary evidence of superiority: they assess whether the current algorithm can be tuned and refitted reliably across distinct mean and covariance mechanisms, and whether its allocation between supervised and residual directions changes with the mechanism. The experiments therefore emphasize computational feasibility and adaptive allocation, with all comparisons evaluated on independent test samples.

B.1 Current algorithm and common protocol

For a fold-training sample, let X_c be the centered predictor matrix and let H contain the centered, whitened categorical response contrasts. With fold size n_f, the current SRDR implementation forms

$$\widehat{\Sigma} = \frac{X_c^\top X_c}{n_f},$$
$$\widehat{C} = \frac{X_c^\top H}{n_f},$$
$$\widehat{K}_a = \left[\widehat{C}, \widehat{\Sigma}\widehat{C}, \ldots, \widehat{\Sigma}^{a-1}\widehat{C}\right].$$

For $a = 0$, $\widehat{K}_0$ is the empty block. After numerical rank truncation, $Q_{\mathrm{sup}} = \mathrm{orth}(\widehat{K}_a)$ spans the entire supervised Krylov space of order a. The residual component is defined conditionally on this full space: with $P_\perp = I - Q_{\mathrm{sup}}Q_{\mathrm{sup}}^\top$, the algorithm retains at most b leading eigenvectors of

$$\widehat{\Sigma}_{\mathrm{res}} = P_\perp \widehat{\Sigma} P_\perp$$

whose eigenvalues exceed $\tau\lambda_1(\widehat{\Sigma})$, and orthogonalizes them against Q_{sup}. The fitted loading matrix is $Q(a,b) = [Q_{\mathrm{sup}}, Q_{\mathrm{res}}]$. Thus the residual block is not ordinary PCA appended after PLS; it is recomputed after conditioning on the complete supervised space selected by a.

If there are K_{cls} classes, the categorical response-feature rank is $h = K_{\mathrm{cls}} - 1$, and the requested nominal dimension is $r = ha + b$. Here $K_{\mathrm{cls}} = 2$, so $h = 1$ and $r = a + b$. Numerical dependence can reduce the realized dimensions to $d_{\mathrm{sup}} \le ha$ and $d_{\mathrm{res}} \le b$; the effective fitted dimension is therefore $r_{\mathrm{eff}} = d_{\mathrm{sup}} + d_{\mathrm{res}} \le r$. The automatic solver uses the dual eigensystem in these $p = 1000$, $p > n_f$ folds, while the same implementation also provides the primal form. No additional first-moment block is appended.

All tuning is nested within the training sample. PCA, PLS, and LOL search over the common nominal-dimension grid

$$\mathcal{R} = \{1, 2, \ldots, 16, 18, 19, 21, 23, 24, 26, 29, 31, 34, 36, 39, 43, 46, 50\}.$$

SRDR considers every admissible (a, b) with $0 \le a, b \le 35$ whose nominal dimension $r = a + b$ belongs to $\mathcal{R}$. Within each of five folds, the embedding and LDA classifier are fitted only on the fold-training observations; LDA uses the actual retained score columns. The mean validation error selects the candidate, after which the selected method is refitted on all 100 training observations and scored on 1000 newly generated test observations.

The inner-CV figures retain a one-dimensional summary without discarding the two-parameter search. For each replication and each nominal dimension r, the SRDR value is the minimum mean fold error over all admissible (a, b) pairs with that r. Curves then average these profiled errors over the 100 replications, and ribbons show Monte Carlo standard errors. The circled point minimizes the displayed mean profile; the actual (a, b) used for holdout evaluation is selected separately within each replication. The holdout summaries below report Monte Carlo means and standard errors rather than treating inner-CV minima as test-risk estimates.

B.2 Scenario design and parameter settings

Table 7 summarizes the purpose of each scenario. Detailed code-level settings are given immediately after the table in paragraph form to avoid overloading a single table with long parameter lists.

Scenario I: low-variance discriminative signal.

The code uses `low_variance_high_count`=60. Therefore, when $p = 1000$, the discriminative coordinate is coordinate 61 rather than one of the leading high-variance coordinates. The first 60 nuisance eigenvalues decrease from 60 to 3, the signal eigenvalue is 0.02, and the remaining tail eigenvalues are 0.01. The noise variance is $\sigma^2 = 1$, and the target Bayes error is 0.05.

Scenario II: rotated covariance and discriminant whitening.

The rotated-pair construction uses $\theta = \pi/4$ and `n_pairs` $= 12$. The high and low eigenvalues in each pair are 180 and 0.001, respectively. The pair decay is 0.98, the mean-weight decay is 0.99, the background eigenvalue is 0.01, and the nuisance count is 0. The target Bayes error is 0.05.

Table 6 Locked computational settings for the six supplementary robustness simulations.

Setting	Value
Number of Monte Carlo replications	100
Training sample size	$n_{\text{train}} = 100$
Independent test sample size	$n_{\text{test}} = 1000$
Ambient dimension	$p = 1000$
Class balance	$1 : 1$
Inner cross-validation folds	$K = 5$
Classifier after dimension reduction	Linear discriminant analysis
Nominal-dimension grid	$\mathcal{R}$ given above, with $\max r = 50$
SRDR caps and solver	$a_{\max} = b_{\max} = 35$; automatic primal/dual computation
Numerical-rank tolerance	$\tau = \sqrt{\epsilon_{\text{machine}}} \approx 1.49 \times 10^{-8}$; applied to Krylov singular values and to residual covariance eigenvalues relative to $\lambda_1(\widehat{\Sigma})$
Data seed in scenario s, replication i	$95000 + 10000s + i$
Train–test split seed; fold seed	$97000 + i$; $98000 + i$
Implementation identifier	`supplementary_robustness_current_ab_v1`

Table 7 Design purpose of the six supplementary robustness scenarios.

Scenario	Purpose of the design
I. Low-variance signal	Probes recovery when discriminative information is not aligned with leading predictor-variance directions.
II. Rotated covariance	Probes the behavior of residual covariance directions when the Bayes direction is induced by rotated covariance structure.
III. Multi-component signal	Tests whether the method can recover class information distributed across multiple supervised components rather than concentrated in one direction.
IV. Nuisance plus residual covariance	Tests robustness to high-variance nuisance variation when useful lower-rank residual covariance remains after supervised directions are removed.
V. Brain-region block correlation	Mimics block-correlated brain features, where nuisance regions may dominate the leading covariance structure but discriminative regions are more subtle.
VI. Weak residual network	Provides a harder stress test with weak sparse signal and a distributed brain-network-like residual signal.

Scenario III: multi-component supervised signal.

The signal coordinates are $\{25, 40, 60, 80\}$, with equal weights $(0.25, 0.25, 0.25, 0.25)$. The design includes 20 high-variance nuisance directions, with nuisance eigenvalues decreasing from 80 to 18. The signal eigenvalues start at 7 and decay by a factor of 0.94. The target Bayes error is 0.05.

Scenario IV: high-variance nuisance plus useful residual covariance.

The design uses $\theta = \pi/4$, nuisance eigenvalue 12, residual eigenvalue 8, direct signal eigenvalue 0.2, and noise variance $\sigma^2 = 1$. The target Bayes error is 0.05.

Scenario V: brain-region block correlation.

The features are partitioned into $G = 100$ regions. Since $p = 1000$, each region has block size 10. The simulation uses 30 signal regions and 45 nuisance regions. Background within-block correlations range from 0.30 to 0.60, signal correlations range from 0.35 to 0.55, and nuisance correlations range from 0.94 to 0.99. The signal covariance rotation is $55\pi/180$, with high eigenvalue 500 and low eigenvalue 10^{-5}. The target Bayes error is 0.05.

Scenario VI: weak distributed residual brain-network signal.

This scenario uses target Bayes error 0.20, making it intentionally harder than the first five scenarios. The sparse signal weight is 0.8 and the distributed network weight is 2.4. The number of network components is 6. For $p = 1000$, the sparse signal size is 30 and the network size is 250. The network eigenvalue is 4.0 with decay 0.95. The design also

includes 12 nuisance components, with nuisance eigenvalues decreasing from 35 to 10.

B.3 Profiled inner-CV tuning curves

Figure 4 shows the inner-CV tuning curves retained as a diagnostic of the tuning landscape. PCA, PLS, and LOL have one candidate at each displayed nominal dimension r. For SRDR, each replication-specific value is the profile minimum over all (a, b) pairs having the same $r = a + b$. Consequently, the SRDR curve represents the full two-dimensional search and is not a curve obtained by fixing either allocation parameter.

Lines are degree-two LOESS smooths with span 0.75. The ribbons use the Monte Carlo standard errors of the unsmoothed dimension-wise means, linearly interpolated only for display. Smoothing is graphical only: it is not used for replication-specific tuning or independent holdout evaluation.

B.4 Holdout test-error results

Figure 5 retains the full replication-level holdout distributions, while Table 8 reports their Monte Carlo means and standard errors. The scenario-specific vertical scales improve within- scenario readability; numerical comparisons across scenarios should use the table rather than the apparent box heights.

B.5 Paired comparisons and selected allocations

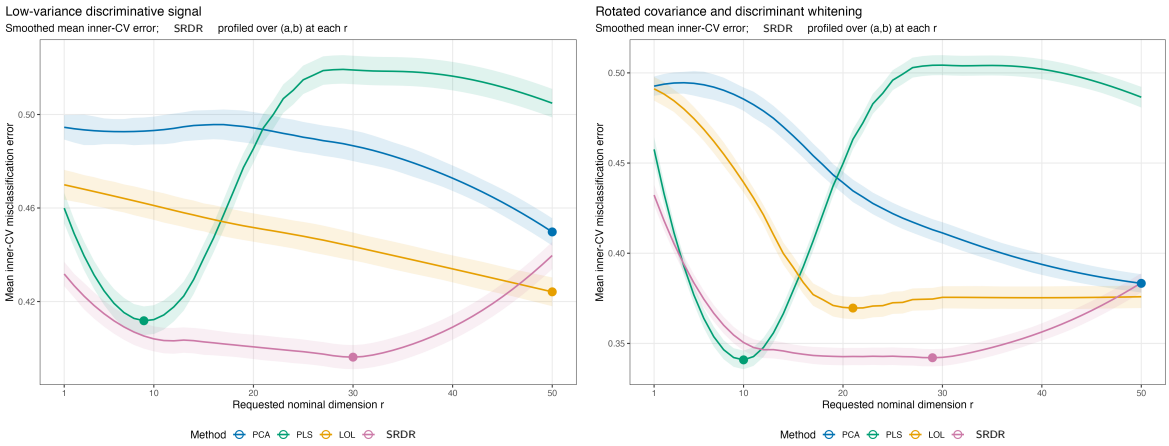
Table 9 reports paired differences because every method is evaluated on the same generated data in a replication. A negative SRDR-minus-baseline difference favors SRDR; the win rate is the proportion of replications in which its error is smaller. These summaries distinguish stable differences from small changes in the unpaired means.

The selected SRDR allocations in Table 10 provide the more direct robustness diagnostic. "Supervised only" means $a > 0, b = 0$, "residual only" means $a = 0, b > 0$, and "mixed" means $a > 0, b > 0$. These labels describe the requested block budgets. Candidates with zero nominal dimension are excluded by the common grid.

B.6 Robustness interpretation

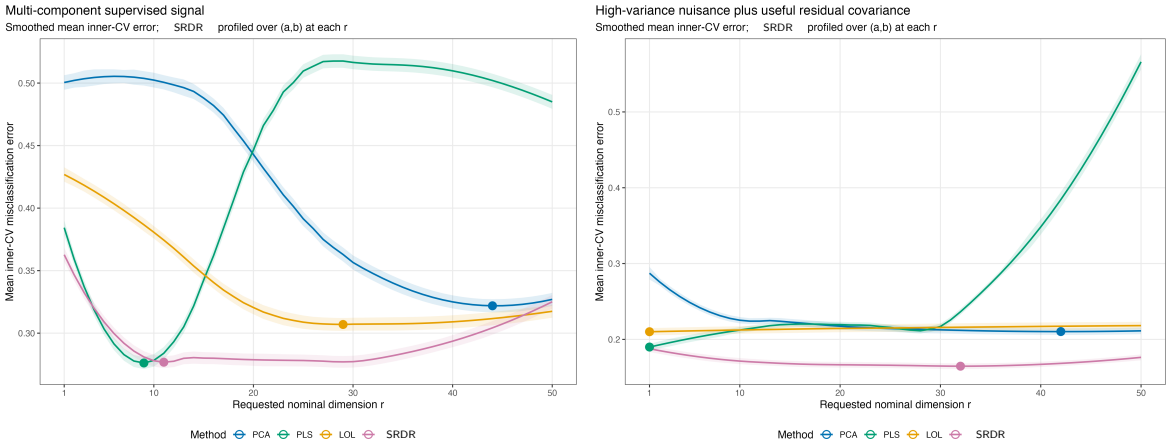
PLS has the smallest mean test error in all six stress tests. SRDR is lower than PCA in every scenario and lower than LOL in Scenarios II, III, IV, and VI. Its differences from PLS are small in Scenarios II, III, IV, and VI (0.00166–0.00634), but are materially larger in the low-variance and brain-block settings (0.03459 and 0.06620). The recomputed results therefore do not support a claim of uniform or scenario-specific dominance over PLS.

The allocation results nevertheless show that the joint search is active rather than collapsing to one fixed block. Mixed selections occur in 72%–95% of replications, while the mean residual budget ranges from 9.71 to 24.08 and the mean supervised order ranges from 3.23 to 9.15. For example, the brain-block scenario assigns a large mean residual budget of 24.08, whereas the nuisance-residual scenario uses the largest mean supervised order, 9.15. Alongside the completed nested-CV fits and independent holdout evaluations in all 600 replications, this variation demonstrates computational feasibility and adaptive behavior of the current algorithm. These supplementary simulations are thus best read as robustness evidence for adaptive allocation, not as the primary performance argument for SRDR.



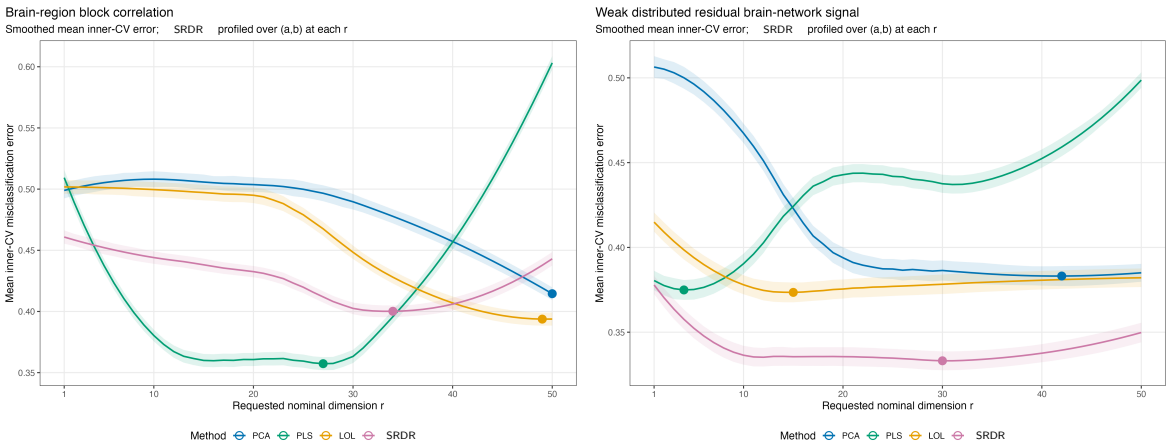
(a) Scenario I

(b) Scenario II



(c) Scenario III

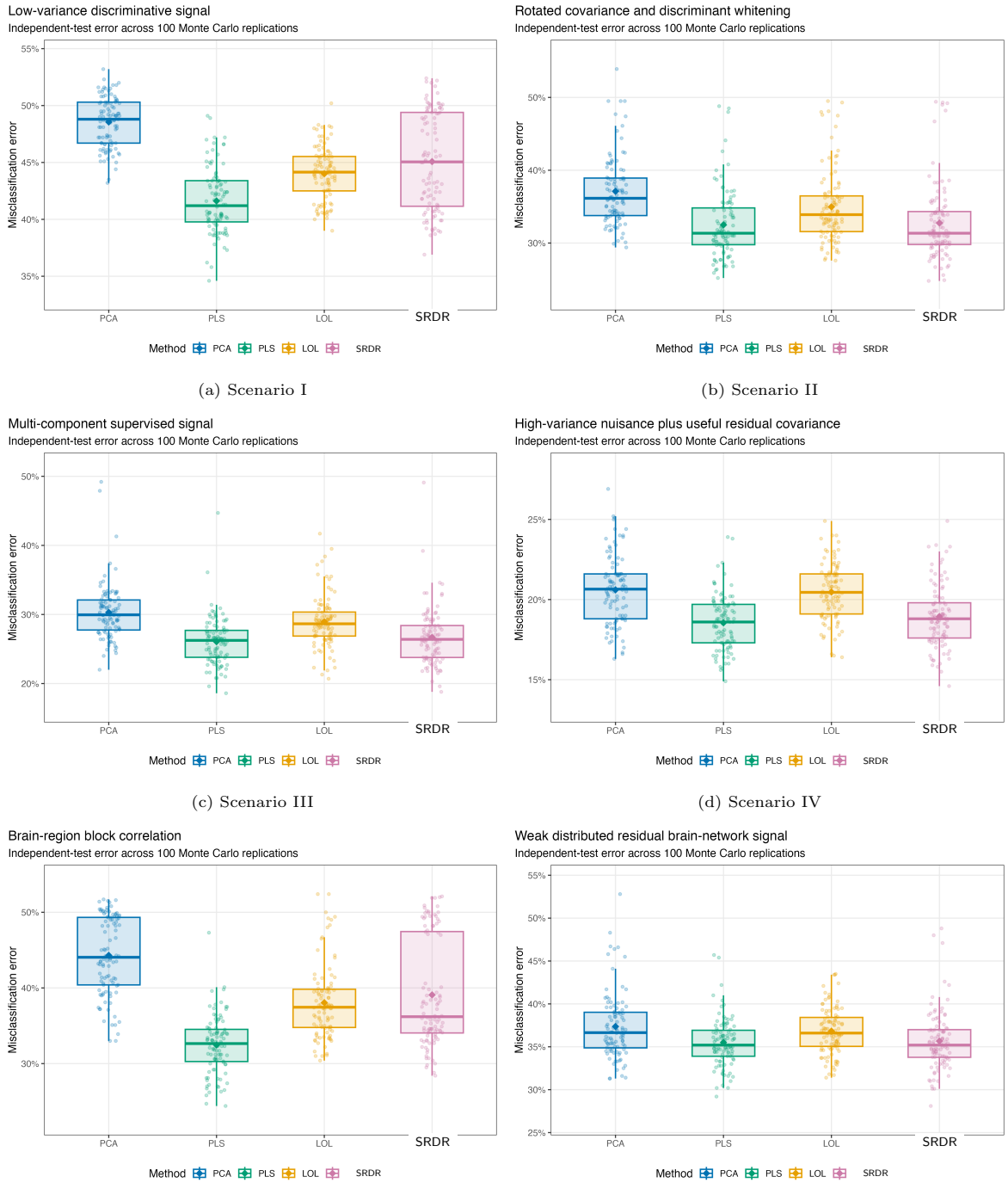
(d) Scenario IV



(e) Scenario V

(f) Scenario VI

Figure 4 Profiled inner-CV tuning curves for the six supplementary robustness simulations. At each nominal dimension r, the SRDR value is profiled over all admissible (a, b) pairs before averaging over 100 replications. Lines are degree-two LOESS smooths with span 0.75; ribbons give Monte Carlo standard errors from the unsmoothed dimension-wise means, interpolated only for display. Circled markers identify the minima of the displayed smooths. Smoothing is graphical only, whereas each holdout fit uses the candidate selected in its own replication.



(a) Scenario I

(b) Scenario II

(c) Scenario III

(d) Scenario IV

(e) Scenario V

(f) Scenario VI

Figure 5 Independent holdout test-error distributions for the six supplementary robustness simulations. Points show all 100 replications, boxplots summarize their distributions, and diamonds mark Monte Carlo means. The vertical scale is chosen separately for each scenario and should not be used for cross-scenario visual comparison.

Table 8 Independent holdout misclassification error over 100 replications. Entries are Monte Carlo means with standard errors in parentheses. Lower is better; the smallest mean in each row is bold.

Scenario	PCA	PLS	LOL	SRDR
I	0.48559 (0.00216)	**0.41617 (0.00286)**	0.44010 (0.00232)	0.45076 (0.00427)
II	0.37108 (0.00475)	**0.32498 (0.00473)**	0.34973 (0.00503)	0.32755 (0.00521)
III	0.30302 (0.00407)	**0.26061 (0.00348)**	0.28948 (0.00373)	0.26695 (0.00416)
IV	0.20606 (0.00217)	**0.18550 (0.00173)**	0.20471 (0.00174)	0.18938 (0.00194)
V	0.44301 (0.00515)	**0.32462 (0.00379)**	0.38054 (0.00487)	0.39082 (0.00714)
VI	0.37366 (0.00384)	**0.35501 (0.00282)**	0.36828 (0.00270)	0.35667 (0.00345)

Table 9 Paired independent-test comparisons over 100 replications. Entries give mean SRDR-minus-baseline error (Monte Carlo SE) and the SRDR win rate. Negative differences favor SRDR.

	PCA		PLS		LOL	
Scenario	Difference (SE)	Win	Difference (SE)	Win	Difference (SE)	Win
I	−0.03483 (0.00372)	82%	0.03459 (0.00420)	12%	0.01066 (0.00416)	44%
II	−0.04353 (0.00392)	94%	0.00257 (0.00494)	36%	−0.02218 (0.00556)	71%
III	−0.03607 (0.00329)	92%	0.00634 (0.00365)	41%	−0.02253 (0.00393)	83%
IV	−0.01668 (0.00205)	81%	0.00388 (0.00173)	37%	−0.01533 (0.00196)	80%
V	−0.05219 (0.00675)	78%	0.06620 (0.00670)	4%	0.01028 (0.00688)	59%
VI	−0.01699 (0.00298)	77%	0.00166 (0.00318)	48%	−0.01161 (0.00280)	71%

Table 10 SRDR selections over 100 replications under the current (a, b) parameterization.

Scenario	Mean selected a	Mean selected b	Supervised only	Residual only	Mixed
I	6.46	11.18	18%	10%	72%
II	3.88	12.30	4%	1%	95%
III	3.74	12.53	6%	1%	93%
IV	9.15	10.00	22%	3%	75%
V	3.23	24.08	3%	10%	87%
VI	6.54	9.71	9%	3%	88%

Appendix C Technical results and proofs

This appendix contains the proofs of the three main results and the complete statements of supporting invariance, rank, nesting, and prediction-risk results. The ordering follows the three theoretical and computational pillars in Section 3.

C.1 Proof of the exact geometry

Proof of Proposition 3.2 Let $A = I_p - \widehat{P}_{\text{sup}}(a)$. If q is an eigenvector of $A\widehat{\Sigma}A$ with eigenvalue $\lambda > 0$, then

$$q = \lambda^{-1} A\widehat{\Sigma}Aq \in \text{col}(A).$$

Thus $\widehat{P}_{\text{sup}}(a)q = 0$. This holds for every retained residual eigenvector, which proves block orthogonality. Each block is orthonormal by construction, so their concatenation is orthonormal. The stated projector identity and direct-sum representation follow immediately. □

C.2 Response invariance, rank, and nesting

The response representation is identifiable only through its column space. The next proposition makes this fact precise and removes dependence on an arbitrary categorical baseline.

Proposition C.1 (Invariance to response coordinates)**.** *Fix $a \geq 1$ and suppose $H^\star = HR$ for a nonsingular $R \in \mathbb{R}^{q\times q}$. In exact arithmetic, if the residual truncation eigenspace is unique, then*

$$\widehat{P}^\star_{\text{sup}}(a) = \widehat{P}_{\text{sup}}(a),$$
$$\widehat{P}^\star_{\text{res}}(a,b) = \widehat{P}_{\text{res}}(a,b),$$
$$\widehat{P}^\star(a,b) = \widehat{P}(a,b).$$

Therefore the estimator is invariant to any nonsingular change of coordinates within the centered response-feature space. Under numerical rank truncation, the same conclusion holds provided the threshold retains the common exact Krylov column space. In particular, whitening both response representations reduces their relationship to an orthogonal coordinate transformation and preserves all singular values used by a relative threshold.

Proof The transformed cross-covariance is $\widehat{C}^\star = \widehat{C}R$. Hence

$$\widehat{K}^\star_a = [\widehat{C}R, \widehat{\Sigma}\widehat{C}R, \ldots, \widehat{\Sigma}^{a-1}\widehat{C}R] = \widehat{K}_a(I_a \otimes R).$$

Because $I_a \otimes R$ is nonsingular, the two Krylov matrices have the same column space and therefore the same supervised projector. Their residual covariance matrices are consequently identical. Uniqueness of the retained residual eigenspace then gives equality of the residual and final projectors. If both versions of H are whitened, their whitened matrices have the same column space and orthonormal columns under the $n^{-1}H^\top H$ inner product; they are therefore related by an orthogonal matrix O. Right multiplication of $\widehat{K}_a$ by $I_a \otimes O$ preserves its singular values, which proves the numerical-threshold assertion. □

Corollary C.2 (Categorical coding and affine response invariance)**.** *For a categorical response, the exact projectors are unchanged by any two full-rank contrast codings with the same centered class-indicator column space. This includes a change of baseline class, a permutation of class labels, effect coding, and orthonormal contrast coding. The result also holds for numerical rank truncation after response whitening, or whenever the retained exact rank is stable. For a scalar continuous response, replacing Y by $Y^\star = \alpha + \beta Y$ with $\beta \neq 0$ leaves all projectors unchanged. If the centered response is standardized, its only change is multiplication by* $\text{sign}(\beta)$.

Proof Any two bases of the same centered categorical contrast space are related by a nonsingular right transformation. Label permutation merely changes that basis. For the scalar response, $M_n Y^\star = \beta M_n Y$, so Proposition C.1 applies with $R = \beta$. □

The same argument applies to a vector of continuous-response basis functions if $h(\alpha+\beta y) = Rh(y) + c$ for a nonsingular R and a constant vector c; centering removes c. It does not cover a change that alters the response feature column space, such as arbitrary class-specific reweighting. Such a change defines a different supervised target rather than a different coding of the same target.

Proposition C.3 (Finite-sample rank bounds)**.** *Let* $c = \operatorname{rank}(\widehat{C})$ *and* $r_X = \operatorname{rank}(X_c) \le \min(p, n-1)$*. For* $a \ge 1$*,*

$$c \le \min(p, n-1, q),$$
$$d_{\mathrm{sup}}(a) \le \min\{p, n-1, ac\}.$$

For every $b \ge 0$*,*

$$d_{\mathrm{res}}(a,b) \le \min\{b, p - d_{\mathrm{sup}}(a), n-1-d_{\mathrm{sup}}(a)\},$$

and therefore

$$r_{\mathrm{eff}}(a,b) \le \min(p, n-1).$$

The same conclusions hold at $a = 0$ *with* $d_{\mathrm{sup}}(0) = 0$*.*

Proof Because $\widehat{C} = n^{-1} X_c^\top H$, its column space is contained in $\mathrm{col}(X_c^\top)$, giving the first bound. Every block $\widehat{\Sigma}^j \widehat{C}$ has rank at most c, while all blocks remain in $\mathrm{col}(X_c^\top)$. Thus the concatenated Krylov matrix has rank at most $\min(ac, r_X)$.

Furthermore, $\widehat{\mathcal{S}}_{\mathrm{sup}}(a) \subseteq \mathrm{col}(X_c^\top)$. The row space available after right projection by $I_p - \widehat{P}_{\mathrm{sup}}(a)$ has dimension at most $r_X - d_{\mathrm{sup}}(a)$. Hence

$$\operatorname{rank}\{X_{\mathrm{res}}(a)\} \le r_X - d_{\mathrm{sup}}(a),$$

and the residual covariance has the same rank. Combining this fact with the budget b and with orthogonality to a $d_{\mathrm{sup}}(a)$-dimensional subspace proves the residual bound. Adding the two block ranks proves the final assertion. □

Remark C.4 (Scalar and binary responses). For a nondegenerate scalar response, and for a binary response represented by one contrast, $c \le 1$. If all a Krylov singular directions and at least b residual eigenvalues pass their respective numerical thresholds, then $d_{\mathrm{sup}}(a) = a$, $d_{\mathrm{res}}(a,b) = b$, and $r_{\mathrm{eff}}(a,b) = r(a,b) = a + b$. These equalities require full retention under the stated numerical-rank convention; they are not assumptions built into the estimator.

Remark C.5 (Downstream LDA). With K nonempty classes, the pooled within-class covariance of an r_{eff}-dimensional score vector has rank at most $n-K$. Classical unregularized LDA therefore requires $r_{\mathrm{eff}} \le n - K$ and a nonsingular within-class score covariance. This restriction belongs to the downstream classifier; regularized discriminant analysis does not require the same inverse to exist.

Proposition C.6 (Valid nesting relations)**.** *For every* $a \ge 0$*,*

$$\widehat{\mathcal{S}}_{\mathrm{sup}}(a) \subseteq \widehat{\mathcal{S}}_{\mathrm{sup}}(a+1).$$

For fixed a*, the leading residual eigenspaces can be chosen so that*

$$\widehat{\mathcal{S}}_{\mathrm{res}}(a,b) \subseteq \widehat{\mathcal{S}}_{\mathrm{res}}(a,b+1),$$

provided ties are retained as complete eigenspaces or a consistent eigenbasis is used.

Proof The columns of $\widehat{K}_a$ form the first a block columns of $\widehat{K}_{a+1}$, proving the first inclusion. For fixed a, the residual covariance is fixed, and its ordered leading eigenvectors generate a nested sequence, which proves the second inclusion under the stated convention. □

Remark C.7 (Lack of joint nesting in a). Increasing a changes $\widehat{P}_{\mathrm{sup}}(a)$ and hence changes $\widehat{\Sigma}_{\mathrm{res}}(a)$. Although the supervised spaces are nested, the residual spaces at orders a and $a+1$ need not be nested. Consequently, the final spaces $\widehat{\mathcal{S}}(a,b)$ are not generally nested as a varies.

C.3 Dual identities and equivalence proof

Lemma C.8 (Dual representation of the supervised Krylov matrix)**.** *For every integer* $j \ge$

0,

$$\widehat{\Sigma}^j \widehat{C} = \frac{1}{n} X_c^\top G^j H.$$

Consequently,

$$\widehat{K}_a = \frac{1}{n} X_c^\top U_a, \qquad \widehat{K}_a^\top \widehat{K}_a = \frac{1}{n} U_a^\top G U_a.$$

Proof The identity at $j = 0$ is the definition of $\widehat{C}$. If it holds at j, then

$$\widehat{\Sigma}^{j+1} \widehat{C} = \frac{1}{n} X_c^\top X_c \left(\frac{1}{n} X_c^\top G^j H \right) = \frac{1}{n} X_c^\top G^{j+1} H,$$

because $G = n^{-1} X_c X_c^\top$. Induction proves the first identity, and horizontal concatenation gives the expression for $\widehat{K}_a$. Finally,

$$\widehat{K}_a^\top \widehat{K}_a = \frac{1}{n^2} U_a^\top X_c X_c^\top U_a = \frac{1}{n} U_a^\top G U_a.$$

□

Proof of Theorem 3.3 Lemma C.8 shows that the dual supervised basis equals a thin eigendecomposition-based orthonormalization of the primal matrix $\widehat{K}_a$. It therefore spans $\mathrm{col}(\widehat{K}_a)$ and gives the same supervised projector. The displayed formula for T_{sup} follows by left multiplication with X_c.

For a positive eigenpair (λ_j, u_j) of G_{res}, the recovered loading has unit norm because

$$\|q_{\mathrm{res},j}\|_2^2 = \frac{u_j^\top X_{\mathrm{res}} X_{\mathrm{res}}^\top u_j}{n \lambda_j} = 1.$$

Moreover,

$$\widehat{\Sigma}_{\mathrm{res}} q_{\mathrm{res},j} = \frac{X_{\mathrm{res}}^\top X_{\mathrm{res}}}{n} \frac{X_{\mathrm{res}}^\top u_j}{(n\lambda_j)^{1/2}} = \lambda_j q_{\mathrm{res},j}.$$

Thus the dual recovery produces exactly the positive-eigenvalue eigenvectors of the primal residual covariance. Matching truncation rules gives the same residual projector. Proposition 3.2 then gives equality of the final projectors. Nonuniqueness of eigenvectors within repeated eigenvalue subspaces affects only their orientation, not the projectors or score spaces. □

C.4 Perturbation details

Lemma C.9 (Perturbation of the block Krylov matrix)**.** *For $a \geq 1$, let $A_a(L)$ and $B_a(L, C)$ be defined as in Section 3.4. Then*

$$\|\widehat{K}_a - K_a\|_{\mathrm{op}} \leq A_a(L)\varepsilon_C + B_a(L, C)\varepsilon_\Sigma.$$

Proof For $j \geq 0$,

$$\widehat{\Sigma}^j \widehat{C} - \Sigma^j C = \widehat{\Sigma}^j (\widehat{C} - C) + (\widehat{\Sigma}^j - \Sigma^j) C.$$

For $j \geq 1$, the telescoping identity

$$\widehat{\Sigma}^j - \Sigma^j = \sum_{\ell=0}^{j-1} \widehat{\Sigma}^{j-1-\ell} (\widehat{\Sigma} - \Sigma) \Sigma^\ell$$

gives

$$\|\widehat{\Sigma}^j - \Sigma^j\|_{\mathrm{op}} \leq j L^{j-1} \varepsilon_\Sigma.$$

Therefore the norm of the jth block difference is bounded by

$$L^j \varepsilon_C + j L^{j-1} \|C\|_{\mathrm{op}} \varepsilon_\Sigma.$$

For a horizontal block matrix $[E_0, \ldots, E_{a-1}]$, $\|[E_0, \ldots, E_{a-1}]\|_{\mathrm{op}} \leq (\sum_j \|E_j\|_{\mathrm{op}}^2)^{1/2}$. Applying the triangle inequality in $\mathbb{R}^a$ to the preceding block bounds proves the result. □

Proof of Theorem 3.5 Write $\widehat{K}_a = K_a + E_K$. A Wedin-type singular-subspace perturbation inequality for two matrices with the same retained rank (Wedin 1972) gives

$$\|\widehat{P}_{\mathrm{sup}} - P_{\mathrm{sup}}\|_{\mathrm{op}} \leq \frac{\|E_K\|_{\mathrm{op}}}{\delta_{\mathrm{sup}} - \|E_K\|_{\mathrm{op}}}.$$

Lemma C.9 and $\varepsilon_K < \delta_{\mathrm{sup}}/2$ yield the first bound.

Let $A = I_p - P_{\mathrm{sup}}$ and $\widehat{A} = I_p - \widehat{P}_{\mathrm{sup}}$. Adding and subtracting $\widehat{A}\Sigma\widehat{A}$ gives

$$\begin{aligned}\widehat{\Sigma}_{\mathrm{res}} - \Sigma_{\mathrm{res}} = {} & \widehat{A}(\widehat{\Sigma} - \Sigma)\widehat{A} \\ & + (\widehat{A} - A)\Sigma\widehat{A} + A\Sigma(\widehat{A} - A).\end{aligned}$$

Since orthogonal projectors and their complements have operator norm one,

$$\|\widehat{\Sigma}_{\mathrm{res}} - \Sigma_{\mathrm{res}}\|_{\mathrm{op}} \leq \varepsilon_\Sigma + 2\|\Sigma\|_{\mathrm{op}} \|\widehat{P}_{\mathrm{sup}} - P_{\mathrm{sup}}\|_{\mathrm{op}} \leq \varepsilon_{\mathrm{res}}.$$

The corresponding Davis–Kahan symmetric eigenspace perturbation inequality (Davis and Kahan 1970); see also Yu et al. (2015), together with $\varepsilon_{\mathrm{res}} < \delta_{\mathrm{res}}/2$, gives the residual bound. Finally, $\widehat{P} - P = (\widehat{P}_{\mathrm{sup}} - P_{\mathrm{sup}}) + (\widehat{P}_{\mathrm{res}} - P_{\mathrm{res}})$, so the triangle inequality proves the final assertion. □

C.5 Supplementary prediction-risk result

The subspace bound also has a direct predictive interpretation for a continuous response. The following result uses independent samples for estimating the embedding and fitting the regression coefficients. Cross-fitting provides the same separation of the two estimation stages.

Corollary C.10 (Regression prediction risk)**.** *Suppose* $X \sim N_p(0, \Sigma)$ *and*

$$Y = X^\top \beta + \epsilon, \qquad \epsilon \perp X,$$
$$\mathbb{E}(\epsilon) = 0, \qquad \mathbb{E}(\epsilon^2) = \sigma^2,$$

and let $\widehat{P} = \widehat{P}(a, b)$ *be estimated independently of a fitting sample of size* n_{fit}*. Let*

$$\gamma^\star_{a,b} \in \arg\min_{\gamma \in \mathcal{S}(a,b)} (\beta - \gamma)^\top \Sigma (\beta - \gamma),$$

and define the population approximation error

$$\mathcal{A}_{a,b} = (\beta - \gamma^\star_{a,b})^\top \Sigma (\beta - \gamma^\star_{a,b}).$$

Assume that the covariance of the retained scores is nonsingular and $n_{\text{fit}} > r_{\text{eff}} + 1$*. If least squares without an intercept is fitted using the centered* $\widehat{P}$ *scores, then, conditional on* $\widehat{P}$*,*

$$\mathbb{E}_{\text{fit}}\{R(\widehat{f}_{a,b}) - \sigma^2 \mid \widehat{P}\} = \widehat{\mathcal{A}}_{a,b} + (\sigma^2 + \widehat{\mathcal{A}}_{a,b}) \frac{r_{\text{eff}}}{n_{\text{fit}} - r_{\text{eff}} - 1},$$

where $\widehat{\mathcal{A}}_{a,b} = \inf_{\gamma \in \widehat{\mathcal{S}}(a,b)} (\beta - \gamma)^\top \Sigma (\beta - \gamma)$*. Here* $R(f) = \mathbb{E}_{\text{new}}\{Y_{\text{new}} - f(X_{\text{new}})\}^2$*. If* $(r_{\text{eff}}+1)/n_{\text{fit}} \le \rho < 1$*, this identity implies*

$$\mathbb{E}_{\text{fit}}\{R(\widehat{f}_{a,b}) - \sigma^2 \mid \widehat{P}\} \lesssim \mathcal{A}_{a,b} + \frac{\sigma^2 r_{\text{eff}}}{n_{\text{fit}}} + \|\Sigma\|_{\text{op}} \|\gamma^\star_{a,b}\|_2^2 \|\widehat{P} - P\|_{\text{op}}^2.$$

If $\|\gamma^\star_{a,b}\|_2 \le C_\beta \|\beta\|_2$*, the final term may be replaced, up to a constant, by* $\|\Sigma\|_{\text{op}} \|\beta\|_2^2 \|\widehat{P} - P\|_{\text{op}}^2$*.*

Proof Because $P\gamma^\star_{a,b} = \gamma^\star_{a,b}$, the vector $\widetilde{\gamma} = \widehat{P}\gamma^\star_{a,b}$ belongs to the estimated space and satisfies

$$\beta - \widetilde{\gamma} = (\beta - \gamma^\star_{a,b}) + (P - \widehat{P})\gamma^\star_{a,b}.$$

For the seminorm $\|v\|_\Sigma^2 = v^\top \Sigma v$,

$$\|\beta - \widetilde{\gamma}\|_\Sigma^2 \le 2\mathcal{A}_{a,b} + 2\|\Sigma\|_{\text{op}} \|\gamma^\star_{a,b}\|_2^2 \|\widehat{P} - P\|_{\text{op}}^2.$$

Thus $\widehat{\mathcal{A}}_{a,b}$ is no larger than the right-hand side.

Conditional on the independently estimated space, the retained score vector is Gaussian. Its population linear-projection residual is independent of the scores and has variance $\sigma^2 + \widehat{\mathcal{A}}_{a,b}$. The Gaussian design cross-product has a Wishart distribution, and the inverse-Wishart first moment (Muirhead 1982, Chapter 3) gives the conditional coefficient-estimation contribution

$$(\sigma^2 + \widehat{\mathcal{A}}_{a,b}) \frac{r_{\text{eff}}}{n_{\text{fit}} - r_{\text{eff}} - 1}.$$

Adding approximation error proves the exact identity. Under $(r_{\text{eff}} + 1)/n_{\text{fit}} \le \rho < 1$, the denominator is of order n_{fit}; substituting the preceding bound for $\widehat{\mathcal{A}}_{a,b}$ proves the stated inequality. □

For fixed a, Proposition C.6 implies that $\mathcal{A}_{a,b}$ is nonincreasing in b. The estimation term, however, typically grows with r_{eff}. The risk bound therefore formalizes the approximation–estimation tradeoff controlled by the residual budget.

References

Bickel, P.J., Levina, E.: Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. Bernoulli **10**(6), 989–1010 (2004) https://doi.org/10.3150/bj/1106314847

Cook, R.D., Forzani, L.: Partial least squares prediction in high-dimensional regression. The Annals of Statistics **47**(2), 884–908 (2019) https://doi.org/10.1214/18-AOS1681

Cook, R.D., Helland, I.S., Su, Z.: Envelopes and partial least squares regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(5), 851–877 (2013) https://doi.org/10.1111/rssb.12018

Cook, R.D.: Fisher lecture: Dimension reduction in regression. Statistical Science **22**(1), 1–26 (2007) https://doi.org/10.1214/088342306000000682

Cook, R.D., Zhang, X.: Foundations for envelope models and methods. Journal of the American Statistical Association **110**(510), 599–611 (2015) https://doi.org/10.1080/01621459.2014.983235

Jong, S.: SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems **18**(3), 251–263 (1993) https://doi.org/10.1016/0169-7439(93)85002-X

Jong, S., Kiers, H.A.L.: Principal covariates regression: Part i. theory. Chemometrics and Intelligent Laboratory Systems **14**(1–3), 155–164 (1992) https://doi.org/10.1016/0169-7439(92)80100-I

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

Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Dexter. UCI Machine Learning Repository (2004). https://doi.org/10.24432/C5P898

Helland, I.S.: On the structure of partial least squares regression. Communications in Statistics—Simulation and Computation **17**(2), 581–607 (1988) https://doi.org/10.1080/03610918808812681

Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology **24**(6), 417–441 (1933) https://doi.org/10.1037/h0071325

Jolliffe, I.T.: A note on the use of principal components in regression. Journal of the Royal Statistical Society: Series C (Applied Statistics) **31**(3), 300–303 (1982) https://doi.org/10.2307/2348005

Koltchinskii, V., Lounici, K.: Concentration inequalities and moment bounds for sample covariance operators. Bernoulli **23**(1), 110–133 (2017) https://doi.org/10.3150/15-BEJ730

Li, K.-C.: Sliced inverse regression for dimension reduction. Journal of the American Statistical Association **86**(414), 316–327 (1991) https://doi.org/10.1080/01621459.1991.10475035

Li, L., Shao, X., Yu, Z.: A slicing-free perspective to sufficient dimension reduction: Selective review and recent developments. International Statistical Review **92**(3), 355–382 (2024) https://doi.org/10.1111/insr.12565

Massy, W.F.: Principal components regression in exploratory statistical research. Journal of the American Statistical Association **60**(309), 234–256 (1965) https://doi.org/10.1080/01621459.1965.10480787

Muirhead, R.J.: Aspects of Multivariate Statistical Theory. John Wiley & Sons, New York (1982). https://doi.org/10.1002/9780470316559

Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901) https://doi.org/10.1080/14786440109462720

Perez-Rogers, J.F., Gerrein, J., Anderlind, C., Liu, G., Zhang, S., Alekseyev, Y., Porta Smith, K., Whitney, D., Johnson, W.E., Elashoff, D.A., Dubinett, S.M., Brody, J., Spira, A., Lenburg, M.E.: Shared gene expression alterations in nasal and bronchial epithelium for lung cancer detection. Journal of the National Cancer Institute **109**(7), 327 (2017) https://doi.org/10.1093/jnci/djw327

Stone, M., Brooks, R.J.: Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. Journal of the Royal Statistical Society: Series B (Methodological) **52**(2), 237–258 (1990) https://doi.org/10.1111/j.2517-6161.1990.tb01786.x

Shen, D., Shen, H., Zhu, H., Marron, J.S.: The statistics and mathematics of high dimension low sample size asymptotics. Statistica Sinica **26**(4), 1747–1770 (2016) https://doi.org/10.5705/ss.202015.0088

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996) https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vogelstein, J.T., Bridgeford, E.W., Tang, M., Zheng, D., Douville, C., Burns, R., Maggioni, M.: Supervised dimensionality reduction for big data. Nature Communications **12**, 2872 (2021) https://doi.org/10.1038/s41467-021-23102-2

Wedin, P.-Å.: Perturbation bounds in connection with singular value decomposition. BIT

12(1), 99–111 (1972) https://doi.org/10.1007/BF01932678

Wold, S., Ruhe, A., Wold, H., Dunn, I. W. J.: The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing **5**(3), 735–743 (1984) https://doi.org/10.1137/0905052

Yoo, J.K., Cook, R.D.: Response dimension reduction for the conditional mean in multivariate regression. Computational Statistics & Data Analysis **53**(2), 334–343 (2008) https://doi.org/10.1016/j.csda.2008.07.029

Yoo, J.K.: Response dimension reduction: Model-based approach. Statistics **52**(2), 409–425 (2018) https://doi.org/10.1080/02331888.2017.1410152

Yu, Y., Wang, T., Samworth, R.J.: A useful variant of the Davis–Kahan theorem for statisticians. Biometrika **102**(2), 315–323 (2015) https://doi.org/10.1093/biomet/asv008

Zhang, X., Deng, K., Mai, Q.: Envelopes and principal component regression. Electronic Journal of Statistics **17**(2), 2447–2484 (2023) https://doi.org/10.1214/23-EJS2154

Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2), 301–320 (2005) https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics **15**(2), 265–286 (2006) https://doi.org/10.1198/106186006X113430

Zhang, X., Mai, Q.: Efficient integration of sufficient dimension reduction and prediction in discriminant analysis. Technometrics **61**(2), 259–272 (2019) https://doi.org/10.1080/00401706.2018.1512901

Zeng, J., Mai, Q., Zhang, X.: Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction. Journal of the American Statistical Association **119**(545), 343–355 (2024) https://doi.org/10.1080/01621459.2022.2118601