# CONTRASTIVE PRINCIPAL COMPONENT REGRESSION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Positive pairs reveal reproducible variation, but the most reproducible directions need not be the most predictive. We introduce contrastive principal component regression (C-PCR), which combines response association and pair alignment to learn a low-dimensional representation, then fits ridge regression for prediction from a single measurement. Under a paired-factor model, we characterize when this joint selection preserves a predictive direction that selection based only on reproducibility would discard. We derive a computable prediction-risk limit as feature dimension and sample size grow proportionally. With one predictive factor and one retained component, the scalar-response formula accounts for using the same responses in component selection and coefficient fitting. The formula separates predictive-direction error from accumulated error elsewhere. Simulations support it and demonstrate finite-sample prediction gains from combining pairing with supervision. Across five naturally paired spectroscopy datasets, our method achieves the best predictive performance or remains competitive with the best alternative.

## 1 INTRODUCTION

Principal component regression (PCR) predicts a response from leading principal component scores, reducing the dimension of the coefficient fit (Pearson, 1901; Massy, 1965). Its ordering is determined by feature variation: a direction with modest variance can be discarded even when it is strongly associated with the response (Jolliffe, 1982; Green & Romanov, 2025). Repeated measurements offer another way to identify useful structure. Variation reproduced across two observations of the same unit is more likely to describe that unit than independent measurement noise. Yet reproducibility and predictive relevance are different properties. A strongly shared factor can be irrelevant to the prediction task and displace a more weakly shared but strongly predictive factor.

Supervised component methods use responses to guide dimension reduction (de Jong, 1993; Bair et al., 2006), whereas contrastive learning methods use positive pairs to learn structure shared across observations (Chen et al., 2020; Tian et al., 2020a). A natural idea is to first learn a low-dimensional linear representation from positive pairs without using response labels and then regress the response on its projected features. We call this baseline PCA++-PCR, combining PCA++-based component selection with regression. Because direction selection does not use responses, it may discard a weakly shared but strongly predictive direction that subsequent regression cannot recover. Increasing the retained dimension can avoid this truncation error, but also changes the representation and its estimation cost. This motivates selecting a small predictive space using responses and matching information together.

**Contributions.** Methodologically, we introduce C-PCR, which combines response association and pair alignment to select components, then fits regression coefficients for single-measurement prediction. Under a paired-factor model, we characterize when joint selection preserves a predictive direction discarded by pair-only selection and quantify the risk reduction. Theoretically, we derive a computable high-dimensional risk limit for a scalar response, one predictive factor, and one retained component. It accounts for using the same responses in selection and coefficient fitting, quantifies the effects of pair weighting and the two ridge penalties, and separates predictive-direction error from accumulated error elsewhere. A multiple-response extension appears in the supplement. Numerically, simulations support the formula and demonstrate prediction gains with complementary signals, including comparisons that expand the initial pair-only space to all features before supervised

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

compression. Across five naturally paired spectroscopy datasets, C-PCR achieves the best predictive performance or remains competitive with the best alternative.

### 1.1 Related work

We discuss here the work most closely related to ours; additional related work appears in the supplement.

Supervised dimension reduction addresses the omission of responses from PCR. Partial least squares (PLS) uses feature–response covariance (de Jong, 1993); supervised PCA selects features or projections through response associations (Bair et al., 2006; Barshan et al., 2011); and principal covariates regression (PCovR) balances feature reconstruction with prediction (de Jong & Kiers, 1992). Reduced-rank ridge regression regularizes both the rank and size of the prediction operator (Mukherjee & Zhu, 2011). These are relevant predictive controls. Our additional information is the correspondence between two views of each training unit.

Contrastive learning uses positive pairs to encourage representations that capture shared information (van den Oord et al., 2018; Tian et al., 2020a; Chen et al., 2020; Radford et al., 2021). These methods motivate learning from paired observations, but generally do not address response-guided component selection for regression. PCA++ studies shared-structure recovery under identity-covariance normalization and provides generalization theory for linear contrastive factor models (Wu et al., 2025). The baseline PCA++-PCR, defined in Section 3.2, adapts pair-only selection to prediction by first extracting PCA++ components and then regressing the response on their scores. Our question is whether response information should also guide the initial dimension reduction: alignment suppresses irrelevant background, while response association distinguishes a predictor from a shared distractor.

**Notation.** We use $I_d$ for the identity matrix, $\mathrm{tr}$ for trace, $\mathrm{span}(U)$ for column space, and $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ for Frobenius and operator norms. For matrices of the same dimensions, $\langle A, B\rangle = \mathrm{tr}(A^{\mathrm{T}}B)$ denotes their Frobenius inner product. The notation $A \succeq 0$ means that $A$ is positive semidefinite. The representation dimension is $q$.

## 2 Problem setup

Suppose we observe independent training units $\mathcal{D}_n = \{(x_i, x_i^+, y_i)\}_{i=1}^n$, where $x_i, x_i^+ \in \mathbb{R}^p$ are one measurement pair in the same feature coordinates and $y_i \in \mathbb{R}$ is their shared scalar response. The two measurements within a unit need not be independent. We assume the variables are mean zero, while in practice we perform empirical centering.

**A predictive factor model.** To distinguish variation, alignment, and predictive relevance, we use the paired-factor model

$$x_i = U^*\Gamma^{1/2}w_i + W^*\Delta^{1/2}h_i + \epsilon_i,\ x_i^+ = U^*\Gamma^{1/2}w_i + W^*\Delta^{1/2}h_i^+ + \epsilon_i^+,\ y_i = \theta^{*\mathrm{T}}w_i + \xi_i, \qquad (1)$$

where $U^* = (u_1^*, \ldots, u_k^*) \in \mathbb{R}^{p\times k}$ and $W^* = (v_1^*, \ldots, v_m^*) \in \mathbb{R}^{p\times m}$ contain the unknown true shared and view-specific loading directions, respectively. The matrices $\Gamma = \mathrm{diag}(a_1, \ldots, a_k)$ and $\Delta = \mathrm{diag}(b_1, \ldots, b_m)$ contain their strengths, and $\theta^* \in \mathbb{R}^k$ contains the unknown true response coefficients. We allow $m = 0$, in which case the background is absent. We also consider a multiple-response model in Section D.1 of the supplement.

**Assumption 2.1 (Paired-factor structure)** The columns of $(U^*, W^*)$ are orthonormal, and all factor strengths $a_j$ and $b_\ell$ are positive. Training units are independent and identically distributed. Within each unit, all scalar coordinates of $w_i \in \mathbb{R}^k$, $h_i, h_i^+ \in \mathbb{R}^m$, and $\epsilon_i, \epsilon_i^+ \in \mathbb{R}^p$ are mutually independent, centered, and have variance one. The observation noise $\xi_i$ is independent of these coordinates, with mean zero and variance $\sigma^2$. The fourth moments of all these scalar variables are bounded by a common finite constant, uniformly over units, coordinates, and dimensions.

The shared or invariant signal vector $w_i$ appears in both views. The independent backgrounds $h_i, h_i^+$ and noises contribute variation but no population pair alignment. Orthonormal loadings

isolate these roles. The moment conditions suffice for controlling empirical fluctuations (Hachem et al., 2013; Knowles & Yin, 2017; Bai & Silverstein, 2010; Rubio & Mestre, 2011). The population moments below use only the second-moment implications of this assumption; the fourth-moment bound is used in the high-dimensional risk analysis.

**Prediction target.** At test time, we predict from one view of a new unit $(x_{\mathrm{te}}, y_{\mathrm{te}})$, distributed as $(x_i, y_i)$ in Equation (1) and independent of $\mathcal{D}_n$. For $\beta \in \mathbb{R}^p$, the predictor $x \mapsto x^{\mathrm{T}}\beta$ has conditional test prediction risk

$$\mathcal{R}(\beta) = \mathbb{E}\big[(y_{\mathrm{te}} - x_{\mathrm{te}}^{\mathrm{T}}\beta)^2 \mid \mathcal{D}_n\big], \tag{2}$$

where $\beta$ may be fitted using $\mathcal{D}_n$ and is held fixed in the conditional expectation.

**Moments.** Define the total covariance $\Sigma_{\mathrm{tot}} = \mathbb{E}[x_i x_i^{\mathrm{T}}]$, the contrastive covariance $\Sigma_{\mathrm{ct}} = \mathbb{E}[x_i x_i^{+\mathrm{T}}]$, and the feature–response cross-covariance vector $b_{\mathrm{reg}} = \mathbb{E}[x_i y_i]$. Under Assumption 2.1, the model gives

$$\Sigma_{\mathrm{tot}} = I_p + U^*\Gamma(U^*)^{\mathrm{T}} + W^*\Delta(W^*)^{\mathrm{T}}, \quad \Sigma_{\mathrm{ct}} = U^*\Gamma(U^*)^{\mathrm{T}}, \quad b_{\mathrm{reg}} = U^*\Gamma^{1/2}\theta^*. \tag{3}$$

The total covariance mixes shared and view-specific directions, whereas contrastive covariance keeps only the shared block $U^*\Gamma(U^*)^{\mathrm{T}}$. Response association is supported on the shared directions $U^*$, with weights $\Gamma^{1/2}\theta^*$. For a direction $u \in \mathbb{R}^p$, the three population quantities $u^{\mathrm{T}}\Sigma_{\mathrm{tot}}u$, $u^{\mathrm{T}}\Sigma_{\mathrm{ct}}u$, and $(u^{\mathrm{T}}b_{\mathrm{reg}})^2$ measure variation, pair alignment, and response association, respectively; they need not order directions the same way.

We estimate these quantities from the observed pairs. Let $X, X^+ \in \mathbb{R}^{n\times p}$ collect the features by row and let $y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ collect the responses. Their empirical counterparts are

$$\widehat{\Sigma}_{\mathrm{tot}} = \frac{X^{\mathrm{T}}X + X^{+\mathrm{T}}X^+}{2n}, \qquad \widehat{\Sigma}_{\mathrm{ct}} = \frac{X^{\mathrm{T}}X^+ + X^{+\mathrm{T}}X}{2n}, \qquad \widehat{b}_{\mathrm{reg}} = \frac{(X + X^+)^{\mathrm{T}}y}{2n}. \tag{4}$$

They satisfy $\mathbb{E}[\widehat{\Sigma}_{\mathrm{tot}}] = \Sigma_{\mathrm{tot}}$, $\mathbb{E}[\widehat{\Sigma}_{\mathrm{ct}}] = \Sigma_{\mathrm{ct}}$, and $\mathbb{E}[\widehat{b}_{\mathrm{reg}}] = b_{\mathrm{reg}}$.

## 3 Contrastive Principal Component Regression

### 3.1 Our procedure

Fix a representation dimension $q$, pair-alignment coefficient or pair weight $\lambda \geq 0$, selection ridge $\tau > 0$, and coefficient ridge $\rho \geq 0$. Let $U \in \mathbb{R}^{p\times q}$ contain the component directions, with component scores $U^{\mathrm{T}}x$, and define $\widehat{\Sigma}_\tau = \widehat{\Sigma}_{\mathrm{tot}} + \tau I_p$. C-PCR proceeds in two steps:

(i) *Component selection.* Using the empirical moments from Section 2, we select $q$ directions by

$$\widehat{U} \in \arg\max_{U^{\mathrm{T}}\widehat{\Sigma}_\tau U = I_q} \left\langle \widehat{b}_{\mathrm{reg}}\widehat{b}_{\mathrm{reg}}^{\mathrm{T}} + \lambda\widehat{\Sigma}_{\mathrm{ct}},\, UU^{\mathrm{T}} \right\rangle. \tag{5}$$

(ii) *Coefficient fitting and prediction.* Given $\widehat{U}$, we fit a coefficient vector $\theta \in \mathbb{R}^q$ to predict $y$ from the computed component scores $\widehat{U}^{\mathrm{T}}x$. With $\widehat{\Sigma}_\rho = \widehat{\Sigma}_{\mathrm{tot}} + \rho I_p$, the fitted coefficients and the resulting prediction function $\widehat{y}(x)$ are

$$\widehat{\theta} = (\widehat{U}^{\mathrm{T}}\widehat{\Sigma}_\rho\widehat{U})^{-1}\widehat{U}^{\mathrm{T}}\widehat{b}_{\mathrm{reg}}, \qquad \widehat{\beta} = \widehat{U}\widehat{\theta}, \qquad \widehat{y}(x) = x^{\mathrm{T}}\widehat{\beta}, \tag{6}$$

where at $\rho = 0$ we use the Moore–Penrose inverse if necessary.

The fitted $\widehat{\theta}$ acts on the learned scores $\widehat{U}^{\mathrm{T}}x$, whereas the ground-truth coefficient $\theta^*$ in Equation (1) acts on the latent shared factors $w$. In Equation (5), the response term $\|U^{\mathrm{T}}\widehat{b}_{\mathrm{reg}}\|_2^2$ rewards association between component scores and the scalar response, while $\mathrm{tr}(U^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{ct}}U)$ rewards pair alignment. The constraint fixes the scale of the representation and uses a ridge-regularized version of the hard uniformity constraint in Wu et al. (2025). Thus responses and pair alignment jointly determine which $q$ directions to retain. The columns of $U$ parameterize the predictive subspace, and the fitted $\widehat{U}$ estimates a predictive subspace of $\mathrm{col}(U^*)$ from data.

The dimension $q$ controls the size of this representation, and $\lambda$ balances pair alignment against response association. The two ridge parameters act at different steps: $\tau$ regularizes direction selection, while $\rho$ regularizes coefficient fitting within the selected space.

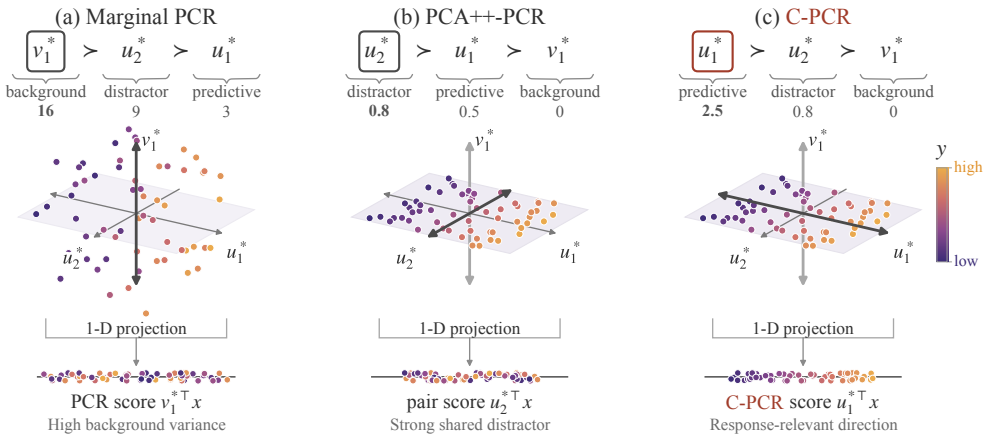


**Figure 1:** Points are schematic; color shows the response signal, which varies only along $u_1^*$, from low to high, not prediction quality. With one retained component, PCR can favor background variance, while PCA++-PCR suppresses independent background but retains the shared distractor. By incorporating response association, C-PCR can retain the predictor. The lower panels show how the resulting scores preserve or mix response information. Headers rank directions by the population criterion ($a_1 = 2$, $a_2 = 8$, $b_1 = 15$, $\vartheta = 2$, $\tau = \lambda = 1$); boxed directions are retained (dark thick arrows). Pair alignment projects points onto the shaded plane $\mathrm{span}(u_1^*, u_2^*)$ in (b)–(c).

### 3.2 Why supervision matters in component selection

To isolate the role of supervision in component selection, we define the two-stage baseline PCA++-PCR. It first selects components from paired features without using responses, using a ridge-regularized PCA++ criterion:

$$\widehat{U}_{\mathrm{pair}} \in \arg \max_{U^\top \widehat{\Sigma}_\tau U = I_q} \left\langle \widehat{\Sigma}_{\mathrm{ct}}, UU^\top \right\rangle. \tag{7}$$

It then applies the same coefficient fit Equation (6). Thus the difference from C-PCR is whether responses guide component selection.

The two endpoints of C-PCR make the role of $\lambda$ explicit:

(i) At $\lambda = 0$ ("the no-pair endpoint"), component selection uses only response association under the $\widehat{\Sigma}_\tau$ normalization.

(ii) As $\lambda \to \infty$ ("the pair-only endpoint"), the selected space converges to that of PCA++-PCR for fixed $\tau$, whenever the limiting space is unique.

Section B.4 derives both endpoints. Figure 1 illustrates the distinction. Under Assumption 2.1, index the single predictive shared factor first, so $y = \vartheta w_1 + \xi$ with $\vartheta \neq 0$ and $\theta^* = \vartheta e_1$, where $e_1$ is the first coordinate vector. Its loading $u_1^*$ is the predictive direction. Each remaining shared loading $u_j^*$, $j = 2, \ldots, k$, is a shared distractor direction: it carries variation common to both views, but its component score $(u_j^*)^\top x$ is uncorrelated with $y$. Writing $\eta_j = a_j/(1 + a_j + \tau)$, the population values of the component-selection criteria are:

| Direction | PCR | Pair-only | C-PCR |
|---|---|---|---|
| Predictive direction $u_1^*$ | $1 + a_1$ | $\eta_1$ | $(\lambda + \vartheta^2)\eta_1$ |
| Shared distractor $u_j^*$, $j = 2, \ldots, k$ | $1 + a_j$ | $\eta_j$ | $\lambda\eta_j$ |
| Background $v_\ell^*$ | $1 + b_\ell$ | 0 | 0 |

Thus, *alignment in C-PCR removes the background directions $v_\ell^*$ from the ranking, and supervision distinguishes the predictive direction $u_1^*$ from the shared distractor directions $u_j^*$, $j = 2, \ldots, k$.* Without loss of generality, relabel the remaining shared factors so that $\eta_2 = \max_{2 \le j \le k} \eta_j$. The following proposition formalizes this mechanism with one retained component, $q = 1$.

**Proposition 3.1 (Predictive selection before truncation)** Under Assumption 2.1, let $\theta^* = \vartheta e_1$, $k \geq 2$, $\vartheta \neq 0$, and $q = 1$. Suppose $\eta_2 > \eta_1$, then the shared distractor direction $u_2^*$ maximizes the population PCA++-PCR criterion, which has zero response association. Population C-PCR selects the predictive direction $u_1^*$ whenever

$$0 \leq \lambda < \lambda_c := \frac{\vartheta^2 \eta_1}{\eta_2 - \eta_1}. \tag{8}$$

After the ridge fit with $\rho \geq 0$ in the second step, let $\beta_{\text{PCA++-PCR}}$ and $\beta_{\text{C-PCR}}$ denote their population prediction coefficients. Then their prediction risks satisfy

$$\mathcal{R}(\beta_{\text{PCA++-PCR}}) - \mathcal{R}(\beta_{\text{C-PCR}}) = \frac{a_1 \vartheta^2 (1 + a_1 + 2\rho)}{(1 + a_1 + \rho)^2} > 0. \tag{9}$$

At $\lambda = \lambda_c$, $u_1^*$ and $u_2^*$ tie; for $\lambda > \lambda_c$, $u_2^*$ maximizes the criterion and outranks $u_1^*$.

Larger response strength $\vartheta^2$ increases the threshold $\lambda_c$, while a larger pair-alignment gap $\eta_2 - \eta_1$ favors the distractor and lowers it. Section C.1 derives the fitted coefficients and risk gap.

This comparison fixes $q = 1$. A wider PCA++-PCR control retains $q_0 \geq q$ pair directions before supervised compression to $q$ components; if its initial space contains $u_1^*$, population refitting with the same $\rho$ removes the truncation gap. Section 5.1 compares these procedures with finite data.

## 4 High-dimensional prediction theory

The population calculation identifies the directions worth retaining. We now quantify the prediction error from estimating the predictive direction and fitting its coefficient from the same data.

We focus on one predictive coordinate, $\theta^* = \vartheta e_1$, and one retained component, $q = 1$. Writing $a = a_1$, the best linear predictor $x \mapsto x^{\mathrm{T}} \beta^*$ has coefficient and risk

$$\beta^* = \frac{\sqrt{a}\,\vartheta}{1 + a} u_1^*, \qquad \mathcal{R}_{\text{lin}}^* = \mathcal{R}(\beta^*) = \sigma^2 + \frac{\vartheta^2}{1 + a}. \tag{10}$$

For any fitted coefficient $\widehat{\beta}$, the excess risk is

$$\mathcal{R}(\widehat{\beta}) - \mathcal{R}_{\text{lin}}^* = \|\widehat{\beta} - \beta^*\|_{\Sigma_{\text{tot}}}^2, \tag{11}$$

where $\|u\|_{\Sigma_{\text{tot}}}^2 = u^{\mathrm{T}} \Sigma_{\text{tot}} u$ weights coefficient error by the test covariance. Section C.1 gives the general benchmark and verifies these identities.

**Assumption 4.1 (Single-coordinate prediction regime)** The following conditions hold as $p, n \to \infty$.

(i) In the paired-factor model Equation (1), let $q = 1$ and $\theta^* = \vartheta e_1$, where $e_1$ is the first coordinate vector and $\vartheta > 0$. Write $a = a_1$ for the predictive factor strength and $s = \vartheta^2 + \sigma^2$ for the response variance.

(ii) As $p, n \to \infty$, assume $p/(2n) \to \gamma \in (0, \infty)$, and the factor loadings may vary with $p$, preserving orthonormality. The numbers of factors $k, m$, their strengths $a_j, b_\ell$, and the response parameters $\vartheta, \sigma^2$ are fixed. The ridge parameters $\tau > 0$ and $\rho \geq 0$ are fixed.

(iii) The pair weight is fixed and satisfies

$$0 \leq \lambda < \lambda_{\text{ex}}, \qquad \lambda_{\text{ex}} = \frac{2\gamma s + 2a\vartheta^2}{1 + \tau - 2\gamma + \sqrt{(1 + \tau - 2\gamma)^2 + 8\gamma\tau}}. \tag{12}$$

Condition (iii) keeps $\lambda$ in a range such that, asymptotically, the response term raises the leading component's criterion value above the range attainable by pair alignment alone. Section C.2 derives the threshold $\lambda_{\text{ex}}$. Let $\widehat{\beta} \in \mathbb{R}^p$ denote the fitted coefficient from Equations (5) and (6) with $q = 1$.

**Theorem 4.2 (Single-coordinate prediction risk)** Under Assumptions 2.1 and 4.1, there are deterministic quantities $b_*, v_* > 0$, computed below, such that the following limits hold in probability:

$$(u_1^*)^\top \widehat{\beta} \to b_*, \quad \|\widehat{\beta}\|_2^2 \to v_*, \quad (u_j^*)^\top \widehat{\beta} \to 0, \ \ 2 \le j \le k, \quad (v_\ell^*)^\top \widehat{\beta} \to 0, \ \ 1 \le \ell \le m.$$

Thus $b_*$ is the limiting coefficient along the predictive direction, and $v_*$ is the limiting squared norm of the fitted coefficient. The excess prediction risk converges in probability as follows:

$$\mathcal{R}(\widehat{\beta}) - \mathcal{R}_{\text{lin}}^* = \|\widehat{\beta} - \beta^*\|_{\Sigma_{\text{tot}}}^2 \longrightarrow \underbrace{(1+a)\left(b_* - \frac{\sqrt{a}\,\vartheta}{1+a}\right)^2}_{\text{error along } u_1^*} + \underbrace{v_* - b_*^2}_{\text{energy in other directions}}. \tag{13}$$

Adding the best linear prediction risk $\mathcal{R}_{\text{lin}}^* = \sigma^2 + \vartheta^2/(1+a)$ gives the limiting conditional test risk of the fitted predictor, denoted by $\mathcal{R}_\infty(\lambda, \tau, \rho)$:

$$\mathcal{R}(\widehat{\beta}) \longrightarrow \mathcal{R}_\infty(\lambda, \tau, \rho) = s - 2\sqrt{a}\,\vartheta b_* + v_* + a b_*^2, \qquad \text{in probability.} \tag{14}$$

To interpret the two limiting terms, write $\widehat{b} = (u_1^*)^\top \widehat{\beta}$ for the fitted coefficient along the predictive direction and $\widehat{\beta}_\perp = \widehat{\beta} - \widehat{b}\,u_1^*$ for its orthogonal remainder. The exact finite-sample excess risk is

$$\mathcal{R}(\widehat{\beta}) - \mathcal{R}_{\text{lin}}^* = (1+a)\left(\widehat{b} - \frac{\sqrt{a}\,\vartheta}{1+a}\right)^2 + \|\widehat{\beta}_\perp\|_2^2 + \sum_{j=2}^{k} a_j \{(u_j^*)^\top \widehat{\beta}\}^2 + \sum_{\ell=1}^{m} b_\ell \{(v_\ell^*)^\top \widehat{\beta}\}^2. \tag{15}$$

The first term is the prediction error from giving the true predictive signal too much or too little weight: it compares the fitted weight $\widehat{b}$ with the optimal linear weight $\sqrt{a}\,\vartheta/(1+a)$. Their squared difference is multiplied by $1+a$, the variance of that component on test data. In the theorem, $\widehat{b}$ is replaced by its limit $b_*$; this term is zero when the two weights agree. The second term sums squared fitted coefficients over all orthogonal directions; the last two sums add the extra test variance along shared distractors and background factors. The theorem makes these last sums vanish because their numbers and strengths are fixed, while

$$\|\widehat{\beta}_\perp\|_2^2 = \|\widehat{\beta}\|_2^2 - \widehat{b}^{\,2} \longrightarrow v_* - b_*^2.$$

This squared norm can remain positive even when individual coefficients are small: for example, $d$ orthogonal coefficients of size $d^{-1/2}$ have total squared norm one as $d \to \infty$. Thus $v_* - b_*^2$ measures the aggregate prediction error from fitted coefficients outside the predictive direction. Section C.2 gives the derivation.

Theorem 4.2 explains why pairing can help even when the population predictive direction is already known to be worth retaining. In this single-coordinate regime, no-pair selection with population moments identifies $u_1^*$, and an unpenalized refit recovers $\beta^*$. With finite data, empirical feature–response associations also contain chance correlations in other directions, and the term $v_* - b_*^2$ quantifies the limiting accumulated error outside $u_1^*$, which can persist when dimension and sample size grow together.

The predictive factor is shared by both views, whereas independent view noise has zero population cross-covariance. Pair alignment can therefore help distinguish predictive structure from chance feature–response correlations. In the decomposition of Theorem 4.2, this can reduce the error energy $v_* - b_*^2$ outside $u_1^*$. Result C.2 makes this precise for small $\gamma$: a small pair weight then lowers the limiting risk, in particular when $\rho \le 1$ or $\rho = \tau$. Section 5.1 gives empirical confirmation.

Section D.1 extends the risk formula to multiple responses and retained directions under the same finite-fourth-moment conditions.

### 4.1 Computing the two risk quantities

Given the model parameters $\gamma, a, \vartheta, \sigma^2$ and tuning parameters $\lambda, \tau, \rho$, set $s = \vartheta^2 + \sigma^2$ and compute $b_*, v_*$ in three scalar steps.

(i) Account for covariance estimation. For a trial value $z > \lambda$, find the unique positive solution $m = m(z)$ of

$$\frac{1}{m} = 2z\tau + (z-\lambda)r_- + (z+\lambda)r_+, \qquad r_\pm = r_\pm(z) = \frac{1}{1+2\gamma(z\pm\lambda)m}, \tag{16}$$

where $m(z)$ is the limiting average reciprocal eigenvalue of $M_n(z) = 2(z\widehat{\Sigma}_\tau - \lambda\widehat{\Sigma}_{\mathrm{ct}})$. The quantities $r_-$ and $r_+$ correspond to the sum and difference views, respectively. As $\gamma \to 0$, $r_\pm \to 1$ and $m \to \{2z(1+\tau)\}^{-1}$, the value without covariance-estimation error.

(ii) Locate the selected component. Using $m, r_-, r_+$ from step (i), evaluate

$$\alpha(z) = \frac{\sqrt{2a}\,\vartheta r_-}{m^{-1} + 2a(z-\lambda)r_-}, \qquad \phi(z) = 2\gamma s m r_- + \frac{2a\vartheta^2 r_-^2}{m^{-1} + 2a(z-\lambda)r_-}. \tag{17}$$

Solve $\phi(z_*) = 1$ for $z_* > \lambda$. Here $z_*$ is the probability limit of the largest generalized eigenvalue of $(\widehat{b}_{\mathrm{reg}}\widehat{b}_{\mathrm{reg}}^{\mathrm{T}} + \lambda\widehat{\Sigma}_{\mathrm{ct}}, \widehat{\Sigma}_\tau)$, the sample matrix pair in Equation (5). Assumption 4.1 guarantees a unique root. The value $\alpha(z_*)$ measures the selected direction's coefficient along $u_1^*$ before normalization.

(iii) Coefficient fitting. Compute

$$d_* = -\partial_z\phi(z_*),\ \chi_* = -\frac{\partial_\tau\phi(z_*)}{2z_*},\ b_* = \frac{\sqrt{2}\,\alpha(z_*)}{d_* + 2(\rho-\tau)\chi_*},\ v_* = \frac{2\chi_*}{\{d_* + 2(\rho-\tau)\chi_*\}^2}. \tag{18}$$

The partial derivatives hold the other arguments fixed before evaluation at $z_*$. To interpret them, let $\widehat{z}$ be the selected generalized eigenvalue and $\widetilde{u}_n = M_n(\widehat{z})^{-1}(\sqrt{2}\,\widehat{b}_{\mathrm{reg}})$ its unnormalized direction. Up to sign, $\widehat{U} = \widetilde{u}_n/(\widetilde{u}_n{}^{\mathrm{T}}\widehat{\Sigma}_\tau\widetilde{u}_n)^{1/2}$, and

$$2\widetilde{u}_n{}^{\mathrm{T}}\widehat{\Sigma}_\tau\widetilde{u}_n \longrightarrow d_*, \qquad \|\widetilde{u}_n\|_2^2 \longrightarrow \chi_*.$$

Thus $d_*$ determines the selection normalization and $\chi_*$ the squared Euclidean length. Since $\widehat{\Sigma}_\rho = \widehat{\Sigma}_\tau + (\rho-\tau)I_p$, the refit denominator $2\widetilde{u}_n{}^{\mathrm{T}}\widehat{\Sigma}_\rho\widetilde{u}_n$ converges to $d_* + 2(\rho-\tau)\chi_*$.

Finally, insert $b_*, v_*$ into Equations (13) and (14) to obtain excess and total prediction risk. Lemma B.3 gives the finite-sample identities, and Section C.2 proves their limits.

## 5 Experiments

This section first examines the accuracy of the high-dimensional risk formula and whether joint component selection improves prediction with complementary signals, and then evaluates predictive performance on five naturally paired spectroscopy datasets.

### 5.1 Synthetic studies

**Examining the limiting risk.** We first assess the limiting risk from Theorem 4.2. Section E.3 gives the model parameters and complete grids. Figure 2(a)–(b) compare empirical mean risk with $\mathcal{R}_\infty(\lambda,\tau,\rho)$ across pair weights and dimension ratios; the largest relative discrepancy is below $0.24\%$. In panel (c), the dimension ratio and pair weight are fixed as $n$ increases. Mean risk approaches the constant theoretical limit, while the central 90% range narrows from $0.063$ at $n = 200$ to $0.020$ at $n = 1600$, supporting both finite-sample accuracy and risk concentration. A three-response example appears in Section E.3.

**Complementary signals beyond pair-only preselection.** We compare the finite-sample performance of C-PCR with that of wider PCA++-PCR, and no-pair and pair-only endpoints.

We use 2 responses, $p = 256$ features, and a final representation of dimension $q = 2$. Each response depends on a separate pair of orthogonal shared factors with strengths $a_\ell < a_h$; these factors explain 80% of its variance. Independent view noise is chosen so that $\Sigma_{\mathrm{tot}} = 2I_p$, while the cross-view correlations along the predictive directions are $a_\ell/2$ and $a_h/2$. Thus marginal variance alone does not

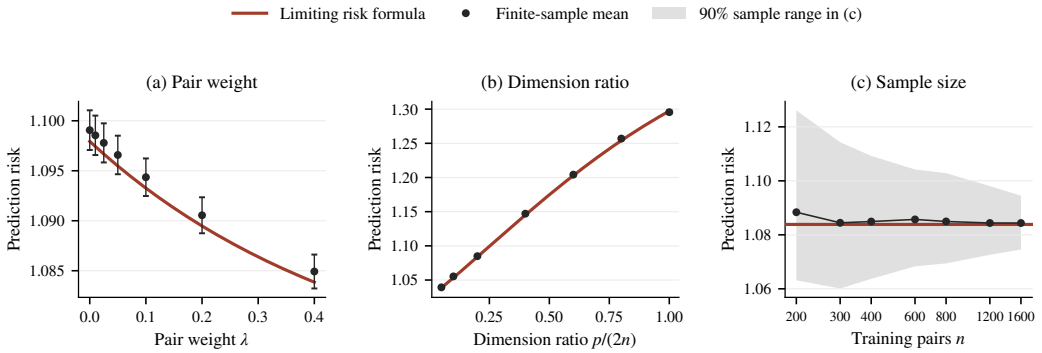


**Figure 2:** Numerical verification of the limiting risk in Theorem 4.2. (a) Pair weight $\lambda$ at $n = 400, p = 160$. (b) Dimension ratio $p/(2n)$ at $n = 400, \lambda = 0.4$. (c) Sample size $n \in \{200, 300, 400, 600, 800, 1200, 1600\}$ at $p/(2n) = 0.2, \lambda = 0.4$. Curves show the theoretical risk and points the mean conditional risk across 256 independent training samples. Error bars in (a)–(b) are pointwise 95% confidence intervals for the mean; shading in (c) is the empirical 5th–95th percentile range of risks over training samples.

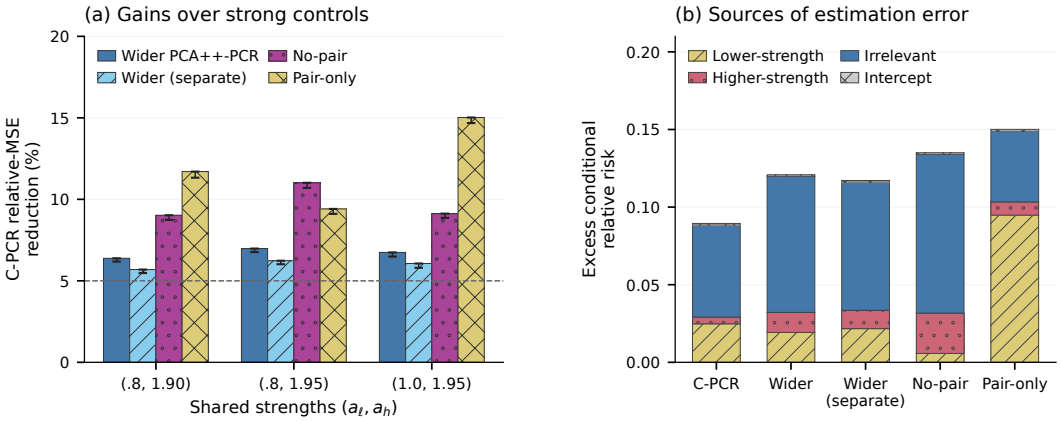


**Figure 3:** Complementary predictive signals with two responses and two retained components. (a) Reduction in mean test relative MSE of C-PCR relative to each indicated control, from 200 independent replications per setting. Whiskers extend down to simultaneous one-sided 95% paired-bootstrap lower bounds, adjusted over all 21 setting–control comparisons; the dashed line marks a 5% reduction. "Wider (separate)" tunes each response independently. (b) Exact conditional relative risk above the optimal linear-prediction risk for $(a_\ell, a_h) = (0.8, 1.90)$, decomposed into coefficient error on lower-strength, higher-strength, and irrelevant directions, plus intercept error, normalized by the population risk of the fitting-mean predictor for each response. Bars average the same 200 replications; they are an algebraic diagnostic, not a further fitted competitor.

distinguish the predictive directions. We consider $(a_\ell, a_h) \in \{(0.8, 1.90), (0.8, 1.95), (1.0, 1.95)\}$. The full model is specified in Section E.1.

Each replication provides every method with the same 384 fitting pairs and 1,536 independent validation pairs, followed by 5,000 test pairs. We define test relative mean squared error (relative MSE) response-wise as the test mean squared prediction error divided by that of the predictor that always returns the response's fitting-sample mean. Squared errors are averaged over the two single-view predictions; for multiple responses, we average the response-wise ratios. The wider PCA++-PCR control searches every initial width from 2 to 256, compresses the selected space to two supervised directions, and refits prediction coefficients. A response-wise version tunes each response separately; grids and further controls are in Section E.1.

Figure 3(a) shows gains of 6.38–6.98% over wider PCA++-PCR, retained at 5.69–6.23% with response-wise tuning, and 9.02–11.01% over jointly tuned no-pair fitting. Across all three settings

**Table 1:** Relative MSE on five naturally paired datasets; lower is better. Darker and lighter gray mark the best and second-best distinct displayed values per row; rounding ties share a shade. Corn and Mango average five fixed splits, Raman eight fits on a common test set, and the others one split. Preprocessing, validation and tuning protocols appear in Section F.

| Dataset | C-PCR | PCA++-PCR | No-pair | PLS | Ridge | PCR |
|---|---|---|---|---|---|---|
| Raman | 0.2140 | 0.2190 | 0.2303 | 0.2193 | 0.2305 | 0.2186 |
| Beef | 0.2268 | 0.2291 | 0.2410 | 0.2615 | 0.2410 | 0.3154 |
| Tablets | 0.6142 | 0.6142 | 0.6031 | 0.6054 | 0.6728 | 0.8224 |
| Corn | 0.1567 | 0.1616 | 0.1515 | 0.1738 | 0.1515 | 0.1924 |
| Mango | 0.1025 | 0.1363 | 0.1058 | 0.2067 | 0.1045 | 0.4719 |

and seven controls, simultaneous comparisons support reductions exceeding 5%. Validation selects an interior pair weight in all 600 fits, although both endpoints are available; Table 3 gives all comparisons.

Figure 3(b) explains the tradeoff: no-pair has the lowest error on weaker predictive directions but the largest irrelevant-direction error, while pair-only has the largest error on weaker signals. Wider controls also fit weaker signals slightly better than C-PCR, but larger errors on stronger and irrelevant directions outweigh that gain. C-PCR balances these errors better with finite data; with population moments known, no-pair selection and an unpenalized refit already recover the optimal linear predictor.

### 5.2 Natural paired measurements

We compare six methods on five naturally paired spectroscopy datasets: Raman mixtures, Beef patties, pharmaceutical Tablets, Corn, and Mango (Lange et al., 2026; Lange, 2025; León-Ecay et al., 2026; International Diffuse Reflectance Conference, 2002; Eigenvector Research, Inc., 2005; Anderson et al., 2024). All methods receive the same labeled fitting views and use the same available preprocessing choices within each dataset. Preprocessing, where varied, and parameters are selected on held-out validation data. Dataset sizes and the full protocol appear in Section F. Table 1 reports relative MSE as in Section 5.1, also averaged over fitting splits.

C-PCR has the lowest mean relative MSE on Raman, Beef and Mango. Mango gives the clearest separation from PCA++-PCR: its mean relative MSE is about 25% lower, and it outperforms PCA++-PCR, PLS and PCR in each of five splits, with smaller average gains over no-pair and ridge. Beef favors pairing over no-pair under both tuning grids. Raman improves on no-pair by about 7% with common SNV preprocessing, which was adopted after test results under the original scatter correction had been seen, so this row is retrospective (Section F.7). Tablets selects the pair-only endpoint, while Corn remains close to no-pair and ridge. These results show predictive gains on several datasets and competitive performance on the others.

## 6 Conclusions and discussion

C-PCR combines supervision and pair alignment to select predictive components. Our population result identifies when joint selection avoids losing a predictive direction, and the high-dimensional risk formula separates predictive-direction error from accumulated error elsewhere. Simulations support the formula and show gains over wider pair-only and no-pair controls; five naturally paired spectroscopy datasets further assess predictive utility. The datasets also indicate when pairing helps: gains are clearest on Mango, Beef and Raman, whereas validation selects the pair-only endpoint on Tablets and C-PCR stays close to no-pair on Corn. Because both endpoints remain available to held-out tuning, C-PCR can reduce to pair-only or no-pair selection when pairing adds little, at the cost of tuning the pair weight.

The theory assumes a paired-factor structure, fixed factor counts, strengths and tuning parameters, and spectral separation; the multiple-response extension also needs $q \leq L$, so neither result covers the validation-tuned fits of Section 5.2. Extensions to data-dependent tuning and growing representation dimensions remain open. Retrospective spectroscopic evidence motivates independent studies across domains and pairs of qualities, including augmentations designed to preserve predictive information (Tian et al., 2020b; von Kügelgen et al., 2021).

### AI USE STATEMENT

In this work, we used generative AI tools to assist with implementation, including experimental code for C-PCR, baseline methods, and simulations. We also used generative AI tools to polish parts of the manuscript for readability, search for and summarize related literature. All AI-assisted work was reviewed and verified by the authors and the authors take full responsibility for the final content of this work.

## REFERENCES


Nicholas Anderson, Kerry Walsh, Phul Subedi, and Jeremy Walsh. Mango DMC and NIR spectra. Mendeley Data, 2024. Version 5.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010. doi: 10.1007/978-1-4419-0661-8.

Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020.

Sijmen de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

Sijmen de Jong and Henk A. L. Kiers. Principal covariates regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems*, 14(1–3):155–164, 1992.

Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38): e2202113119, 2022.

Eigenvector Research, Inc. NIR of corn samples for standardization benchmarking. https://www.eigenvector.com/data/Corn/, 2005. Data originally collected at Cargill.

Alden Green and Elad Romanov. The high-dimensional asymptotics of principal component regression. *The Annals of Statistics*, 53(4):1697–1727, 2025.

Walid Hachem, Philippe Loubaton, Jamal Najim, and Pascal Vallet. On bilinear forms based on the resolvent of large random matrices. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(1):36–63, 2013.

Inge S. Helland. On the structure of partial least squares regression. *Communications in Statistics – Simulation and Computation*, 17(2):581–607, 1988.

International Diffuse Reflectance Conference. NIR spectra of pharmaceutical tablets from the 2002 shootout. Dataset distributed by Eigenvector Research, 2002. Accessed September 23, 2026.

Ian T. Jolliffe. 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.

Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007*, volume 4539 of *Lecture Notes in Computer Science*, pp. 82–96. Springer, 2007.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1–2):257–352, 2017. doi: 10.1007/s00440-016-0730-4.

Christoph Lange. SubstrateMixRaman. Hugging Face dataset, 2025. Dataset accompanying Lange et al., Measurement 258, 118884.

Christoph Lange, Madeline Altmann, Daniel Stors, Simon Seidel, Kyle Moynahan, Linda Cai, Stefan Born, Peter Neubauer, and M. Nicolas Cruz Bournazou. Deep learning for Raman spectroscopy: Benchmarking models for upstream bioprocess monitoring. *Measurement*, 258:118884, 2026.

Sara León-Ecay, Kizkitza Insausti, Silvia Arazuri, Carlos López-Molina, and Ainara López-Maestresalas. Dataset of near-infrared spectra for beef patties enriched with plant-based proteins: point-based vs hyperspectral imaging. Zenodo, 2026. Version 1.

William F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*, 4(6):612–622, 2011.

Karl Pearson. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011. doi: 10.1016/j.spl.2011.01.004.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision (ECCV 2020)*, volume 12356 of *Lecture Notes in Computer Science*, pp. 776–794, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, 2020b.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Mingqi Wu, Qiang Sun, and Archer Yang. PCA++: How uniformity induces robustness to background noise in contrastive learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shuo Yang, Yijun Dong, Rachel Ward, Inderjit S. Dhillon, Sujay Sanghavi, and Qi Lei. Sample efficiency of data augmentation consistency regularization. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3825–3853, 2023.

## SUPPLEMENTARY MATERIAL

This supplement provides the technical details and additional experiments supporting the main text. Section A summarizes the notation. Section B presents estimator identities, computational details, and connections to related methods. Section C proves the main results, and Section D develops the multiple-response extension. Section E gives simulation designs and additional results; Section F documents the natural-data experiments and sensitivity analyses.

**Guide to the proofs.**

(i) Section C.1 proves Proposition 3.1 by comparing population direction scores and refitting the retained predictor.

(ii) Section C.2 proves Theorem 4.2 directly: whole-pair deletion controls noise inverses, and a joint factor–response calculation gives alignment and coefficient norm. Section C.3 gives a second proof by specializing the multiple-response result, Theorem D.2.

(iii) Section D.3 proves Theorem D.2 using the same basic noise limits, a two-inverse kernel for cross-direction inner products, and the finite-sample identities in Lemma B.3.

(iv) Section C.4 proves Result C.2 by expanding the scalar risk limit for a small dimension ratio and a small pair weight.

## TABLE OF CONTENTS


**A Notation glossary** **13**

**B Estimator identities and computation** **14**
B.1 Additional related work . . . 14
B.2 The C-PCR algorithm . . . 14
B.3 Well-posedness and row-space computation . . . 14
B.4 Regression reductions and the pair-only limit . . . 17
B.5 Connection to prediction-consistency regression . . . 19

**C Proofs of the main results** **20**
C.1 Population risk and component selection . . . 20
C.2 A direct scalar risk proof . . . 21
C.3 A second proof from the multiple-response theorem . . . 24
C.4 Local effect of the pair weight . . . 24

**D Prediction risk with multiple responses** **25**
D.1 Model and prediction risk . . . 25
D.2 Computing the risk . . . 26
D.3 Proof of the multiple-response theorem . . . 27

**E Additional simulations** **29**
E.1 Complementary predictive signals: model and confirmation . . . 29
E.2 Population illustration of early truncation . . . 31
E.3 Fixed-parameter risk calibration . . . 32
E.4 Wider PCA++-PCR spaces and supervised controls . . . 32

**F Natural paired data** **34**
F.1 Shared evaluation protocol . . . 34
F.2 Raman carbon-source prediction . . . 36
F.3 Beef-patty prediction . . . 36
F.4 Pharmaceutical tablets . . . 37
F.5 Corn . . . 37
F.6 Mango dry matter . . . 37
F.7 Grid and protocol sensitivity . . . 37

## A NOTATION GLOSSARY

Table 2 summarizes the principal notation used throughout the paper, including the main model parameters, latent factors, loading vectors, covariance quantities, and asymptotic objects. To distinguish symbols with similar notation, $m$ denotes the number of background factors, whereas $m(z)$ denotes the bulk function defined in (16); similarly, $v_\ell^*$ denotes the loading vector associated with background factor $\ell$, whereas $v_*$ denotes a scalar limiting quantity.

**Table 2:** Glossary of notation.

| Notation | Explanation |
|---|---|
| $\mathcal{D}_n = \{(x_i, x_i^+, y_i)\}_{i=1}^n$ | Training pairs: views $x_i, x_i^+ \in \mathbb{R}^p$ and response $y_i$ (§2) |
| $n,\ p,\ \gamma$ | Number of pairs, feature dimension, and $p/(2n) \to \gamma$ |
| $X,\ X^+,\ y$ | Views and responses stacked by row (4) |
| $U^* = (u_j^*),\ W^* = (v_\ell^*)$ | Shared and background loadings; $u_1^*$ is predictive (1) |
| $k,\ m$ | Numbers of shared and background factors |
| $w_i;\ h_i, h_i^+$ | Shared factors; view-specific background factors |
| $a_j,\ b_\ell;\ \Gamma,\ \Delta$ | Factor strengths and their diagonal matrices; $a = a_1$ |
| $\epsilon_i, \epsilon_i^+;\ \xi_i,\ \sigma^2$ | Feature noises; response noise and its variance |
| $\theta^*,\ \vartheta,\ s$ | Response coefficients, $\theta^* = \vartheta e_1$, and $s = \vartheta^2 + \sigma^2$ |
| $\Sigma_{\mathrm{tot}},\ \Sigma_{\mathrm{ct}},\ b_{\mathrm{reg}}$ | Total and contrastive covariances; feature–response covariance (3) |
| $q,\ \lambda,\ \tau,\ \rho$ | Representation dimension, pair weight, selection and coefficient ridges |
| $\widehat{\Sigma}_\tau,\ \widehat{\Sigma}_\rho$ | $\widehat{\Sigma}_{\mathrm{tot}} + \tau I_p$ and $\widehat{\Sigma}_{\mathrm{tot}} + \rho I_p$ |
| $\widehat{M}_\lambda,\ \widehat{H}_{\lambda,\tau}$ | Selection matrix $\widehat{b}_{\mathrm{reg}}\widehat{b}_{\mathrm{reg}}^{\mathrm{T}} + \lambda\widehat{\Sigma}_{\mathrm{ct}}$ and its whitened form |
| $\widehat{U},\ \widehat{\theta},\ \widehat{\beta}$ | Selected directions (5), refit coefficients, and $\widehat{\beta} = \widehat{U}\widehat{\theta}$ (6) |
| $\eta_j,\ \lambda_c$ | Pair-only eigenvalue of $u_j^*$; critical pair weight (Prop. 3.1) |
| $q_0$ | Initial pair-only width of wider PCA++-PCR |
| $P,\ J_{\mathrm{swap}},\ r_X$ | Stacked design, view swap, and $\mathrm{rank}(P)$ (Result B.1) |
| $\mathcal{R}(\beta),\ \mathcal{R}^*_{\mathrm{lin}}$ | Conditional test risk (2); best linear risk |
| $b_*,\ v_*$ | Limits of $(u_1^*)^{\mathrm{T}}\widehat{\beta}$ and $\|\widehat{\beta}\|_2^2$ (Thm. 4.2) |
| $\mathcal{R}_\infty(\lambda, \tau, \rho)$ | Limiting test risk (14) |
| $M_n(z)$ | Selection pencil $2(z\widehat{\Sigma}_\tau - \lambda\widehat{\Sigma}_{\mathrm{ct}})$, $z > \lambda$ |
| $m(z),\ r_\pm(z)$ | Solutions of the bulk equation (16) |
| $\alpha(z),\ \phi(z)$ | Alignment and response maps (17) |
| $z_*,\ \widehat{z}$ | Root of $\phi(z_*) = 1$; selected sample eigenvalue, $\widehat{z} \to z_*$ |
| $\widetilde{u}_n;\ d_*,\ \chi_*$ | Unnormalized direction; its normalization and length limits (18) |
| $\lambda_{\mathrm{ex}}$ | Upper bound on admissible pair weights (12) |
| $\zeta,\ c_\pm,\ t_z,\ \iota$ | $2\gamma$, $z \pm \lambda$, $2z\tau$, and the sign index $\iota \in \{-, +\}$ |
| $\mu,\ \kappa$ | Disagreement weight; direction ridge in consistency and wider controls |
| $L,\ \Theta^*,\ \Sigma_\xi,\ \Omega$ | Number of responses, coefficient matrix, noise and response covariances |
| $F,\ r,\ V$ | $(U^*, W^*)$, $r = k + m$, and factor-space test covariance |
| $\Phi,\ A_z,\ \psi,\ \Xi$ | Response map, factor alignment, two-root bulk term, and response kernel |
| $H,\ E,\ J,\ N;\ \mathcal{R}_q$ | Direction limits, $N = E - H^{\mathrm{T}}H$, and limiting risk (Thm. D.2) |
| $a_\ell,\ a_h$ | Lower and higher strengths in §5.1; labels, not indices |
| $q_{\max};\ \mathcal{G}_\tau, \mathcal{G}_\lambda, \mathcal{G}_\rho$ | Maximal dimension and tuning grids (Alg. 2) |
| $L_{\mathrm{val}}$, rMSE | Validation loss; test relative MSE (59) |

# B ESTIMATOR IDENTITIES AND COMPUTATION

## B.1 ADDITIONAL RELATED WORK

Multiview regression uses unlabeled pairs to construct a canonical-correlation-based regression penalty (Kakade & Foster, 2007). Cooperative learning couples view-specific predictors through an agreement penalty and establishes a local prediction benefit from alignment (Ding et al., 2022); augmentation consistency penalizes changes in predictions under transformations (Yang et al., 2023). The population quantities $\Sigma_{\mathrm{tot}}$, $\Sigma_{\mathrm{ct}}$, and $b_{\mathrm{reg}}$ of Section 2 connect to different component-selection principles. Classical PCR selects directions by total variance (Massy, 1965). Wu et al. (2025) use pair alignment under total covariance normalization to extract shared structure. The response-association term also appears in supervised PCA with a linear response kernel, under Euclidean normalization (Barshan et al., 2011).

C-PCR formulates supervised dimension reduction: responses and pair alignment jointly determine which components to retain, and coefficients are then fitted in the selected space for single-view prediction. Our analysis quantifies the prediction error of this full procedure when feature dimension and sample size grow together. We provide a computable risk formula for one component that accounts for using the same responses in selection and coefficient fitting, with an extension to multiple responses under the same factor model. The precise one-component connection to consistency-regularized regression is given in Section B.5.

## B.2 THE C-PCR ALGORITHM

Algorithm 1 states the estimator of Section 3.1 for fixed tuning parameters. After empirical centering, it follows the two steps of that section: component selection solves Equation (5), and coefficient fitting solves Equation (6). Selection has a dense route for $p < 2n$ and a row-space route for $p \geq 2n$ that never forms a $p \times p$ matrix (Result B.1). In the pseudocode, $\mathrm{eig}_q(A)$ returns the $q$ leading orthonormal eigenvectors of a symmetric matrix $A$ and their eigenvalues $z_{1:q} = (z_1 \geq \cdots \geq z_q)$, and $\mathrm{svd}(P)$ returns the factors of the compact singular value decomposition $P = U_X \Lambda_X V_X^{\mathrm{T}}$. Algorithm 2 gives the validation search used for the natural data.

Algorithm 2 places $\tau$ and $\lambda$ in the outer loops because each pair requires a new eigenproblem, while the inner loops over $q$ and $\rho$ reuse it: leading eigenvectors are nested in $q$, and each $\rho$ needs only a $q \times q$ solve. With the row-space route, the SVD of $P$ does not depend on $(\tau, \lambda)$, so the search costs one compact SVD, $|\mathcal{G}_\tau||\mathcal{G}_\lambda|$ eigenproblems of size $r_X = \mathrm{rank}(P)$, and at most $|\mathcal{G}_\tau||\mathcal{G}_\lambda| q_{\max} |\mathcal{G}_\rho|$ small solves. The mean predictor is always a candidate, so the selected model never has a larger validation loss than it. Ties in validation loss keep the earlier candidate, and the validation loss is

$$L_{\mathrm{val}}(\beta) = \frac{1}{|\mathcal{D}_{\mathrm{val}}|} \sum_{(x,y) \in \mathcal{D}_{\mathrm{val}}} \left\{ y - \bar{y} - (x - \bar{x})^{\mathrm{T}} \beta \right\}^2,$$

where each validation measurement is predicted from one view. The natural-data grids are $\mathcal{G}_\tau = s_X \mathcal{G} \cup \{\infty\}$, $\mathcal{G}_\lambda = \{0\} \cup s_Y \mathcal{H} \cup \{\infty\}$, $\mathcal{G}_\rho = \{0\} \cup s_X \mathcal{G}$, and $q_{\max} = 20$ (Section F). When preprocessing is also selected, Algorithm 2 runs once per preprocessing and the lowest validation loss is kept.

## B.3 WELL-POSEDNESS AND ROW-SPACE COMPUTATION

For multiple responses, let $Y \in \mathbb{R}^{n \times L}$ collect the response vectors by row and set $\widehat{B}_{\mathrm{reg}} = (X^{\mathrm{T}} Y + X^{+\mathrm{T}} Y)/(2n)$. The matrix versions of Equations (5) and (6) are

$$\widehat{U} \in \arg \max_{U^{\mathrm{T}} \widehat{\Sigma}_\tau U = I_q} \left\langle \widehat{B}_{\mathrm{reg}} \widehat{B}_{\mathrm{reg}}^{\mathrm{T}} + \lambda \widehat{\Sigma}_{\mathrm{ct}}, U U^{\mathrm{T}} \right\rangle. \tag{19}$$

Here $\Theta \in \mathbb{R}^{q \times L}$ contains the response coefficients on the selected component scores, and $\widehat{B}_q \in \mathbb{R}^{p \times L}$ is the fitted prediction coefficient matrix.

$$\widehat{\Theta} = (\widehat{U}^{\mathrm{T}} \widehat{\Sigma}_\rho \widehat{U})^{-1} \widehat{U}^{\mathrm{T}} \widehat{B}_{\mathrm{reg}}, \qquad \widehat{B}_q = \widehat{U} \widehat{\Theta}, \qquad \widehat{y}(x) = \widehat{B}_q^{\mathrm{T}} x. \tag{20}$$

At $\rho = 0$, use the Moore–Penrose inverse if needed. For a fixed representation $U$, define the pooled ridge loss

$$\ell_\rho(\Theta; U) = \frac{\|Y - XU\Theta\|_F^2 + \|Y - X^+ U\Theta\|_F^2}{2n} + \rho \|U\Theta\|_F^2. \tag{21}$$

For any invertible $R \in \mathbb{R}^{q \times q}$, replacing $(U, \Theta)$ by $(UR, R^{-1}\Theta)$ leaves the predictor and penalty unchanged.

**Algorithm 1** C-PCR estimator for fixed tuning parameters

**Input:** pairs $\mathcal{D}_n$; dimension $q \geq 1$; pair weight $\lambda \in [0, \infty]$; selection ridge $\tau \in (0, \infty]$; coefficient ridge $\rho \geq 0$

**Output:** directions $\widehat{U} \in \mathbb{R}^{p \times q}$; coefficients $\widehat{\beta} \in \mathbb{R}^p$; prediction function $\widehat{y}(\cdot)$

**Stage 1: empirical centering**

$\triangleright$ *Center both views by one pooled mean, so they keep the same feature coordinates.*

$\bar{x} \leftarrow$ mean of all $2n$ views $x_i, x_i^+$

$\bar{y} \leftarrow$ mean of the $n$ responses $y_i$

$X, X^+ \leftarrow$ matrices with rows $x_i - \bar{x}$ and $x_i^+ - \bar{x}$

$y \leftarrow$ vector with entries $y_i - \bar{y}$

**Stage 2: component selection** (Equation (5))

$\triangleright$ *Rank directions by response association plus $\lambda$ times pair alignment, relative to $\widehat{\Sigma}_\tau$.*

**if** $p < 2n$ **then** $\triangleright$ *dense route, Equation (22)*

  form $\widehat{\Sigma}_{\mathrm{tot}}, \widehat{\Sigma}_{\mathrm{ct}}, \widehat{b}_{\mathrm{reg}}$ as in Equation (4)

  **if** $\lambda = \infty$ **then** $\triangleright$ *pair-only endpoint*

    $\widehat{M}_\lambda \leftarrow \widehat{\Sigma}_{\mathrm{ct}}$

  **else**

    $\widehat{M}_\lambda \leftarrow \widehat{b}_{\mathrm{reg}} \widehat{b}_{\mathrm{reg}}^{\mathrm{T}} + \lambda \widehat{\Sigma}_{\mathrm{ct}}$

  **if** $\tau = \infty$ **then** $\triangleright$ *Euclidean metric*

    $\widehat{\Sigma}_\tau \leftarrow I_p$

  **else**

    $\widehat{\Sigma}_\tau \leftarrow \widehat{\Sigma}_{\mathrm{tot}} + \tau I_p$

  $\widehat{H}_{\lambda,\tau} \leftarrow \widehat{\Sigma}_\tau^{-1/2} \widehat{M}_\lambda \widehat{\Sigma}_\tau^{-1/2}$ $\triangleright$ *whitened selection matrix*

  $(Q_q, z_{1:q}) \leftarrow \mathrm{eig}_q(\widehat{H}_{\lambda,\tau})$ $\triangleright$ *leading eigenpairs*

  $\widehat{U} \leftarrow \widehat{\Sigma}_\tau^{-1/2} Q_q$ $\triangleright$ *so $\widehat{U}^{\mathrm{T}} \widehat{\Sigma}_\tau \widehat{U} = I_q$*

**else** $\triangleright$ *row-space route, Result B.1*

  $P \leftarrow (2n)^{-1/2} \left[\begin{smallmatrix} X \\ X^+ \end{smallmatrix}\right]$ $\triangleright$ *$\widehat{\Sigma}_{\mathrm{tot}} = P^{\mathrm{T}} P$*

  $y_{\mathrm{pool}} \leftarrow (2n)^{-1/2} \left[\begin{smallmatrix} y \\ y \end{smallmatrix}\right]$ $\triangleright$ *$\widehat{b}_{\mathrm{reg}} = P^{\mathrm{T}} y_{\mathrm{pool}}$*

  $(U_X, \Lambda_X, V_X) \leftarrow \mathrm{svd}(P)$ $\triangleright$ *compact, rank $r_X \leq 2n$*

  **if** $\tau = \infty$ **then** $\triangleright$ *Euclidean metric*

    $G_\tau \leftarrow \Lambda_X$

  **else**

    $G_\tau \leftarrow \Lambda_X (\Lambda_X^2 + \tau I_{r_X})^{-1/2}$

  **if** $\lambda = \infty$ **then** $\triangleright$ *pair-only endpoint*

    $\widehat{H}_{\lambda,\tau} \leftarrow G_\tau U_X^{\mathrm{T}} J_{\mathrm{swap}} U_X G_\tau$

  **else**

    $\widehat{H}_{\lambda,\tau} \leftarrow G_\tau U_X^{\mathrm{T}} (\lambda J_{\mathrm{swap}} + y_{\mathrm{pool}} y_{\mathrm{pool}}^{\mathrm{T}}) U_X G_\tau$

  $(\overline{Q}_q, z_{1:q}) \leftarrow \mathrm{eig}_q(\widehat{H}_{\lambda,\tau})$ $\triangleright$ *same nonzero spectrum as $\widehat{H}_{\lambda,\tau}$*

  $\widehat{U} \leftarrow V_X \Lambda_X^{-1} G_\tau \overline{Q}_q$ $\triangleright$ *$\Lambda_X^{-1} G_\tau = (\Lambda_X^2 + \tau I_{r_X})^{-1/2}$ for $\tau < \infty$*

**if** $z_q \leq 0$ **then** $\triangleright$ *$\widehat{\Sigma}_{\mathrm{ct}}$ can be indefinite*

  **return** REJECT $\triangleright$ *fewer than $q$ positive roots*

**Stage 3: coefficient fitting** (Equation (6))

$\triangleright$ *Fit $y$ on the scores $\widehat{U}^{\mathrm{T}} x$ of both views with penalty $\rho \|\widehat{U}\theta\|_2^2$ (via $P\widehat{U}$ if $p \geq 2n$).*

**if** $\rho > 0$ **then**

  $\widehat{\theta} \leftarrow (\widehat{U}^{\mathrm{T}} \widehat{\Sigma}_\rho \widehat{U})^{-1} \widehat{U}^{\mathrm{T}} \widehat{b}_{\mathrm{reg}}$ $\triangleright$ *Cholesky solve*

**else**

  $\widehat{\theta} \leftarrow (\widehat{U}^{\mathrm{T}} \widehat{\Sigma}_{\mathrm{tot}} \widehat{U})^{+} \widehat{U}^{\mathrm{T}} \widehat{b}_{\mathrm{reg}}$ $\triangleright$ *Moore–Penrose inverse*

$\widehat{\beta} \leftarrow \widehat{U} \widehat{\theta}$

**Stage 4: prediction**

**return** $\widehat{U}$, $\widehat{\beta}$, and $\widehat{y}(x) = \bar{y} + (x - \bar{x})^{\mathrm{T}} \widehat{\beta}$ $\triangleright$ *$x$ from either view*

**Algorithm 2** Validation search for C-PCR

**Input:** fitting pairs $\mathcal{D}_{\text{fit}}$; validation data $\mathcal{D}_{\text{val}}$; grids $\mathcal{G}_\tau, \mathcal{G}_\lambda, \mathcal{G}_\rho$ for $\tau, \lambda, \rho$; maximal dimension $q_{\max}$
**Output:** selected $(\widehat{q}, \widehat{\lambda}, \widehat{\tau}, \widehat{\rho})$; prediction function $\widehat{y}(\cdot)$

▷ *Centering and scaling use $\mathcal{D}_{\text{fit}}$ only (Section F).*
center $\mathcal{D}_{\text{fit}}$ as in lines 2–5 of Algorithm 1
▷ *Start from the fitting-response-mean predictor, $\beta = 0$.*
$L^\star \leftarrow L_{\text{val}}(0)$
$\widehat{\beta}^\star \leftarrow 0$
$\widehat{q} \leftarrow 0$ ▷ *$\widehat{q} = 0$: mean predictor*
**for** $\tau \in \mathcal{G}_\tau$ **do**
  **for** $\lambda \in \mathcal{G}_\lambda$ **do**
    ▷ *One eigenproblem per $(\lambda, \tau)$ serves every $q$: leading eigenspaces are nested.*
    $(\widehat{U}, z_1, z_2, \ldots) \leftarrow$ lines 7–33 of Algorithm 1 with $q = \min\{q_{\max}, r_X\}$
    $\bar{q} \leftarrow \#\{j : z_j > 0\}$ ▷ *positive roots only*
    **for** $q = 1, \ldots, \bar{q}$ **do**
      $\widehat{U}_q \leftarrow$ first $q$ columns of $\widehat{U}$
      **for** $\rho \in \mathcal{G}_\rho$ **do**
        $\widehat{\beta} \leftarrow$ Stage 3 of Algorithm 1 with $(\widehat{U}_q, \rho)$ ▷ *$q \times q$ solve*
        $L \leftarrow L_{\text{val}}(\widehat{\beta})$
        **if** $L < L^\star - 10^{-12}$ **then** ▷ *strict improvement*
          $L^\star \leftarrow L$
          $\widehat{\beta}^\star \leftarrow \widehat{\beta}$
          $(\widehat{q}, \widehat{\lambda}, \widehat{\tau}, \widehat{\rho}) \leftarrow (q, \lambda, \tau, \rho)$
**return** $(\widehat{q}, \widehat{\lambda}, \widehat{\tau}, \widehat{\rho})$ and $\widehat{y}(x) = \bar{y} + (x - \bar{x})^{\mathrm{T}} \widehat{\beta}^\star$ ▷ *no refit on $\mathcal{D}_{\text{val}}$*

For computation, write $\widehat{M}_\lambda = \widehat{B}_{\text{reg}} \widehat{B}_{\text{reg}}^{\mathrm{T}} + \lambda \widehat{\Sigma}_{\text{ct}}$. The component-selection problem in Equation (19) has a spectral solution. Let $Q_q$ contain the leading $q$ orthonormal eigenvectors of $\widehat{H}_{\lambda,\tau}$, where

$$\widehat{H}_{\lambda,\tau} = \widehat{\Sigma}_\tau^{-1/2} \widehat{M}_\lambda \widehat{\Sigma}_\tau^{-1/2}, \qquad \widehat{U} = \widehat{\Sigma}_\tau^{-1/2} Q_q. \tag{22}$$

For $\tau > 0$ and $1 \le q \le p$, the change of variables $Q = \widehat{\Sigma}_\tau^{1/2} U$ maps the constraint to the compact Stiefel manifold $Q^{\mathrm{T}} Q = I_q$. The objective becomes $\operatorname{tr}(Q^{\mathrm{T}} \widehat{H}_{\lambda,\tau} Q)$; the Ky Fan principle gives its leading invariant space. It is unique at a strict rank-cutoff eigengap, and is the whole feature space when $q = p$.

The coefficient fit minimizes Equation (21). Its normal equation is $(U^{\mathrm{T}} \widehat{\Sigma}_\rho U)\Theta = U^{\mathrm{T}} \widehat{B}_{\text{reg}}$, yielding Equation (20) whenever the restricted matrix is nonsingular. This holds for $\rho > 0$. At $\rho = 0$, the same formula uses a Moore–Penrose inverse if needed. For a scalar response, $Y = y$, $\widehat{B}_{\text{reg}} = \widehat{b}_{\text{reg}}$, $\widehat{\Theta} = \widehat{\theta}$, and $\widehat{B}_q = \widehat{\beta}$, recovering Equation (6).

**Result B.1 (Exact row-space formulation)** Let

$$P = \frac{1}{\sqrt{2n}} \begin{bmatrix} X \\ X+ \end{bmatrix}, \qquad Y_{\text{pool}} = \frac{1}{\sqrt{2n}} \begin{bmatrix} Y \\ Y \end{bmatrix}, \qquad J_{\text{swap}} = \begin{bmatrix} 0 & I_n \\ I_n & 0 \end{bmatrix}.$$

Write $P = U_X \Lambda_X V_X^{\mathrm{T}}$ for its compact SVD of rank $r_X > 0$, with $\Lambda_X$ positive diagonal. Put $G_\tau = \Lambda_X (\Lambda_X^2 + \tau I_{r_X})^{-1/2}$ and define

$$\widehat{\overline{H}}_{\lambda,\tau} = G_\tau U_X^{\mathrm{T}} (\lambda J_{\text{swap}} + Y_{\text{pool}} Y_{\text{pool}}^{\mathrm{T}}) U_X G_\tau.$$

Then

$$\widehat{H}_{\lambda,\tau} = V_X \widehat{\overline{H}}_{\lambda,\tau} V_X^{\mathrm{T}}.$$

Hence the nonzero C-PCR spectrum is obtained from the $r_X \times r_X$ matrix $\widehat{\overline{H}}_{\lambda,\tau}$. If $\overline{Q}_q$ contains $q$ leading eigenvectors with positive eigenvalues, the corresponding feature loadings are

$$\widehat{U} = V_X (\Lambda_X^2 + \tau I_{r_X})^{-1/2} \overline{Q}_q.$$

Thus for positive selected eigenvalues and $p \gg n$, C-PCR requires no $p \times p$ matrix: a compact SVD of the $2n \times p$ stacked design is followed by $r_X \times r_X$ eigenproblems, where $r_X \le 2n$.

*Proof of Result B.1.* The definitions give $\widehat{\Sigma}_{\mathrm{tot}} = P^{\mathrm{T}}P$, $\widehat{\Sigma}_{\mathrm{ct}} = P^{\mathrm{T}}J_{\mathrm{swap}}P$, and $\widehat{B}_{\mathrm{reg}} = P^{\mathrm{T}}Y_{\mathrm{pool}}$. On the row space of $P$,

$$\widehat{\Sigma}_{\tau}^{-1/2}P^{\mathrm{T}} = V_X(\Lambda_X^2 + \tau I_{r_X})^{-1/2}\Lambda_X U_X^{\mathrm{T}} = V_X G_\tau U_X^{\mathrm{T}}.$$

Since $\widehat{M}_\lambda = P^{\mathrm{T}}(\lambda J_{\mathrm{swap}} + Y_{\mathrm{pool}}Y_{\mathrm{pool}}^{\mathrm{T}})P$, substitution yields the displayed factorization of $\widehat{H}_{\lambda,\tau}$. The operator vanishes on the orthogonal complement of the design row space. Multiplying its positive-eigenvalue vectors by $\widehat{\Sigma}_{\tau}^{-1/2}$ gives the stated feature loadings. □

For $p \geq 2n$, a dense compact SVD costs $O\{p(2n)^2\}$, followed by $O(r_X^3)$ for each dense spectral solve.

### B.4 Regression reductions and the pair-only limit

This subsection derives the two endpoints described in Section 3.2 and the joint loss formulation of C-PCR when $\rho = \tau$. Throughout, the training data, representation dimension $q$, and ridge parameters are fixed.

**Remark B.2 (A joint optimization problem for $\rho = \tau$.)** When $\rho = \tau$, the normalization gives $\widehat{\Theta} = \widehat{U}^{\mathrm{T}}\widehat{B}_{\mathrm{reg}}$, and the multiple-response procedure in Equations (19) and (20) jointly solves

$$\min_{U^{\mathrm{T}}\widehat{\Sigma}_\tau U = I_q,\, \Theta \in \mathbb{R}^{q\times L}} \left\{ \frac{\|Y - XU\Theta\|_F^2 + \|Y - X^+U\Theta\|_F^2}{2n} + \tau\|U\Theta\|_F^2 - \lambda\,\mathrm{tr}(U^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{ct}}U) \right\}. \quad (23)$$

With $B = U\Theta$, the first two terms are the pooled ridge loss, while the last term rewards pair alignment. Minimizing over $\Theta$ yields the component-selection problem in Equation (19). At $\lambda = 0$, this reduces to the rank-constrained pooled ridge problem derived below.

*Proof of the equivalence in Remark B.2.* To verify the equivalence of Equation (23) and the procedure in Equations (19) and (20) at $\rho = \tau$, fix $U$ with $U^{\mathrm{T}}\widehat{\Sigma}_\tau U = I_q$. Completing the square gives $\widehat{\Theta} = U^{\mathrm{T}}\widehat{B}_{\mathrm{reg}}$. In terms of the pooled ridge loss in Equation (21),

$$\min_{\Theta}\{\ell_\tau(\Theta; U) - \lambda\,\mathrm{tr}(U^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{ct}}U)\} = \frac{\mathrm{tr}(Y^{\mathrm{T}}Y)}{n} - \mathrm{tr}(U^{\mathrm{T}}\widehat{M}_\lambda U).$$

Minimizing over $U$ gives Equation (19), proving the equivalence. For $\rho \neq \tau$, Equation (20) retains these directions and refits their coefficients by minimizing $\ell_\rho$; the resulting pair need not minimize Equation (23). □

**No-pair endpoint: reduced-rank ridge regression.** At $\lambda = 0$ and $\rho = \tau$, write $B = U\Theta$ in Equation (23). Every matrix $B$ of rank at most $q$ admits such a factorization with $U^{\mathrm{T}}\widehat{\Sigma}_\tau U = I_q$: choose a $q$-dimensional basis containing the column space of $B$ and normalize it in the $\widehat{\Sigma}_\tau$ metric. Conversely, every product $U\Theta$ has rank at most $q$. Thus the fitted coefficient solves the pooled reduced-rank ridge problem (Mukherjee & Zhu, 2011),

$$\widehat{B}_q \in \arg\min_{\mathrm{rank}(B)\leq q} \left\{ \frac{\|Y - XB\|_F^2 + \|Y - X^+B\|_F^2}{2n} + \tau\|B\|_F^2 \right\}.$$

The rank constraint means that the $L$ responses are predicted through at most $q$ common components. To see the solution explicitly, set $G = \widehat{\Sigma}_\tau^{-1/2}\widehat{B}_{\mathrm{reg}}$. Up to a term independent of $B$, the objective is $\|\widehat{\Sigma}_\tau^{1/2}B - G\|_F^2$, so its minimizer is obtained by retaining the leading $q$ singular directions of $G$. Equivalently, if $Q_q$ contains leading orthonormal eigenvectors of $GG^{\mathrm{T}}$, then

$$\widehat{B}_q = \widehat{\Sigma}_\tau^{-1/2}Q_qQ_q^{\mathrm{T}}G.$$

This is exactly Equations (19) and (20) at $\lambda = 0$ and $\rho = \tau$. The no-pair criterion determines at most $\mathrm{rank}(\widehat{B}_{\mathrm{reg}})$ directions; any additional directions have zero response association and do not change this fitted coefficient. The response-rank distinction is stated in Result B.4.

For a scalar response with $L = q = 1$, the rank constraint is automatic, giving ordinary pooled ridge, $\widehat{B}_1 = \widehat{\Sigma}_\tau^{-1}\widehat{b}_{\mathrm{reg}}$. More generally, pooled ridge is recovered whenever $q \geq \mathrm{rank}(\widehat{B}_{\mathrm{reg}})$. These identities for the fitted coefficient require $\rho = \tau$; choosing $\rho \neq \tau$ retains the same selected space but changes the coefficient fit.

**Pair-only endpoint and the PCA++-PCR limit.** For $\lambda > 0$, dividing the selection objective by $\lambda$ leaves its maximizers unchanged. In whitened coordinates,

$$\lambda^{-1}\widehat{H}_{\lambda,\tau} = \widehat{\Sigma}_\tau^{-1/2}\widehat{\Sigma}_{\mathrm{ct}}\widehat{\Sigma}_\tau^{-1/2} + \lambda^{-1}GG^{\mathrm{T}} \longrightarrow \widehat{\Sigma}_\tau^{-1/2}\widehat{\Sigma}_{\mathrm{ct}}\widehat{\Sigma}_\tau^{-1/2} \quad \text{in operator norm.}$$

If the limiting matrix has a strict gap between its $q$th and $(q+1)$st eigenvalues, its leading $q$-dimensional space is unique, and the selected spaces converge to it. Multiplication by $\widehat{\Sigma}_\tau^{-1/2}$ then gives convergence to the pair-only feature space in Equation (7). For $q = p$, the selected space is already the whole feature space. If the cutoff eigenvalue is tied, every accumulation point still maximizes the pair-only criterion, but the response term can resolve the tie.
Applying Equation (20) to the limiting space gives PCA++-PCR. The fitted predictor converges as well whenever the limiting restricted Gram matrix is nonsingular, which is automatic for $\rho > 0$. Thus pair-only refers to component selection: regression coefficients continue to use the responses.

**PLS endpoint.** At $\lambda = 0$, the selected direction is $\widehat{\Sigma}_\tau^{-1}\widehat{b}_{\mathrm{reg}}$. Since the refit depends only on the selected span, as $\tau \to \infty$,

$$\tau\widehat{\Sigma}_\tau^{-1}\widehat{b}_{\mathrm{reg}} = (I_p + \widehat{\Sigma}_{\mathrm{tot}}/\tau)^{-1}\widehat{b}_{\mathrm{reg}} \longrightarrow \widehat{b}_{\mathrm{reg}}, \qquad \widehat{\beta} \longrightarrow \frac{\widehat{b}_{\mathrm{reg}}^{\mathrm{T}}\widehat{b}_{\mathrm{reg}}}{\widehat{b}_{\mathrm{reg}}^{\mathrm{T}}\widehat{\Sigma}_\rho\widehat{b}_{\mathrm{reg}}}\,\widehat{b}_{\mathrm{reg}}.$$

The first PLS weight of the stacked views $\widetilde{X} = (X^{\mathrm{T}}, X^{+\mathrm{T}})^{\mathrm{T}}$ and responses $\widetilde{Y} = (Y^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ is $\widetilde{X}^{\mathrm{T}}\widetilde{Y} = 2n\widehat{B}_{\mathrm{reg}}$, so at $\rho = 0$ this limit is the one-component PLS coefficient. For several responses, the same limit gives the leading left singular vectors of $\widehat{B}_{\mathrm{reg}}$, the linear-kernel supervised PCA space (Barshan et al., 2011). Beyond one component the methods differ: PLS adds the Krylov directions $\widehat{\Sigma}_{\mathrm{tot}}\widehat{b}_{\mathrm{reg}}, \widehat{\Sigma}_{\mathrm{tot}}^2\widehat{b}_{\mathrm{reg}}, \ldots$ (Helland, 1988), whereas no-pair C-PCR determines only $\mathrm{rank}(\widehat{B}_{\mathrm{reg}})$ directions and, for $\lambda > 0$, further directions come from pair alignment.

**Exterior directions and coefficient refitting.** For the scalar and multiple-response risk calculations, write the paired design in sum and difference coordinates,

$$Z = \frac{X + X^+}{\sqrt{2}}, \qquad D = \frac{X - X^+}{\sqrt{2}},$$

with Gram matrices $A_Z = Z^{\mathrm{T}}Z/n$ and $A_D = D^{\mathrm{T}}D/n$. Shared variation enters only $A_Z$. The moment matrices and the response quantities are

$$\widehat{\Sigma}_{\mathrm{tot}} = \tfrac{1}{2}(A_Z + A_D), \quad \widehat{\Sigma}_{\mathrm{ct}} = \tfrac{1}{2}(A_Z - A_D), \quad g_n = Z^{\mathrm{T}}Y/n = \sqrt{2}\,\widehat{B}_{\mathrm{reg}}, \quad \Omega_n = Y^{\mathrm{T}}Y/n.$$

For $z > \lambda$, the selection pencil $M_n(z) = 2(z\widehat{\Sigma}_\tau - \lambda\widehat{\Sigma}_{\mathrm{ct}})$ and the response map are

$$M_n(z) = (z-\lambda)A_Z + (z+\lambda)A_D + 2z\tau I_p, \qquad \Phi_n(z) = g_n^{\mathrm{T}}M_n(z)^{-1}g_n. \tag{24}$$

An exterior root is a generalized eigenvalue greater than the pair-only bound $\lambda$.

**Lemma B.3 (Finite-sample exterior identities)** For $\lambda \geq 0$, $\tau > 0$, and $z > \lambda$, $M_n(z) \succ 0$. Exterior roots satisfy $\det\{I_L - \Phi_n(z)\} = 0$, and their number is at most $\mathrm{rank}(\widehat{B}_{\mathrm{reg}}) \leq L$. Moreover,

$$-\Phi_n'(z) \succeq \frac{\Phi_n(z)}{z+\lambda}, \qquad 0 \preceq \Phi_n(z) \preceq \frac{\Omega_n}{z-\lambda}. \tag{25}$$

At a simple root $\widehat{z}_j$, choose a unit $\widehat{a}_j$ with $\Phi_n(\widehat{z}_j)\widehat{a}_j = \widehat{a}_j$, and put $d_{nj} = -\widehat{a}_j^{\mathrm{T}}\Phi_n'(\widehat{z}_j)\widehat{a}_j$. The metric-normalized direction obeys

$$\widehat{U}_j = \sqrt{2/d_{nj}}\,M_n(\widehat{z}_j)^{-1}g_n\widehat{a}_j, \qquad \widehat{U}_j^{\mathrm{T}}\widehat{B}_{\mathrm{reg}} = \widehat{a}_j^{\mathrm{T}}/\sqrt{d_{nj}}. \tag{26}$$

For $q$ such directions in decreasing root order, the coefficient refit is nonsingular even at $\rho = 0$, since

$$\widehat{U}_q^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{tot}}\widehat{U}_q \succeq \frac{\widehat{z}_q}{\lambda + \|\Omega_n\|_{\mathrm{op}}} I_q, \qquad \widehat{\Theta}_q = \{I_q + (\rho - \tau)\widehat{U}_q^{\mathrm{T}}\widehat{U}_q\}^{-1}\widehat{U}_q^{\mathrm{T}}\widehat{B}_{\mathrm{reg}}. \tag{27}$$

*Proof of Lemma B.3.* The identities $\widehat{\Sigma}_{\mathrm{tot}} \pm \widehat{\Sigma}_{\mathrm{ct}} \succeq 0$ bound the pair-only spectrum by $\lambda$; the response perturbation has rank $\mathrm{rank}(\widehat{B}_{\mathrm{reg}})$. The determinant lemma gives the root equation. Since $M_n' = 2\widehat{\Sigma}_\tau$, differentiation yields $-\Phi_n' = 2g_n^{\mathrm{T}} M_n^{-1} \widehat{\Sigma}_\tau M_n^{-1} g_n$. The bound $M_n \preceq 2(z+\lambda)\widehat{\Sigma}_\tau$ gives the first inequality in Equation (25); $M_n \succeq (z-\lambda) Z^{\mathrm{T}} Z/n$ gives the second. Substituting the root equation verifies Equation (26). Finally, $\widehat{B}_{\mathrm{reg}} \widehat{B}_{\mathrm{reg}}^{\mathrm{T}} \preceq \|\Omega_n\|_{\mathrm{op}} \widehat{\Sigma}_{\mathrm{tot}}$ and $\widehat{\Sigma}_{\mathrm{ct}} \preceq \widehat{\Sigma}_{\mathrm{tot}}$ imply $\mathrm{diag}(\widehat{z}_1, \ldots, \widehat{z}_q) \preceq (\lambda + \|\Omega_n\|_{\mathrm{op}}) \widehat{U}_q^{\mathrm{T}} \widehat{\Sigma}_{\mathrm{tot}} \widehat{U}_q$. Metric normalization then gives Equation (27) and its positive-definite denominator. □

Uniform response rescaling by $c > 0$, with pair weight rescaled by $c^2$, multiplies the objective by $c^2$, preserving its selected space and multiplying predictions by $c$; separate rescaling of outputs need not preserve that space.

### B.5 Connection to prediction-consistency regression

For a scalar response and one retained component, write $\widehat{B}_{\mathrm{reg}} = \widehat{b}_{\mathrm{reg}}$ and $D_\Delta = \widehat{\Sigma}_{\mathrm{tot}} - \widehat{\Sigma}_{\mathrm{ct}} \succeq 0$. Consider the paired ridge-consistency estimator

$$\begin{aligned} \widehat{\beta}_{\mathrm{con}}(\mu, \kappa) &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|y - X\beta\|_2^2 + \|y - X^+\beta\|_2^2}{2n} + \mu \beta^{\mathrm{T}} D_\Delta \beta + \kappa \|\beta\|_2^2 \right\} \\ &= (\widehat{\Sigma}_{\mathrm{tot}} + \mu D_\Delta + \kappa I_p)^{-1} \widehat{b}_{\mathrm{reg}}, \end{aligned} \tag{28}$$

where $\mu \geq 0$ and $\kappa > 0$. Write $A(\mu, \kappa) = \widehat{\Sigma}_{\mathrm{tot}} + \mu D_\Delta + \kappa I_p$ for the inverted matrix, which is positive definite on that range. The term $\beta^{\mathrm{T}} D_\Delta \beta = \|(X - X^+)\beta\|_2^2/(2n)$ penalizes disagreement between paired predictions. This is a symmetrically supervised ridge adaptation of augmentation consistency (Yang et al., 2023): it places the supervised loss on both views and includes a ridge penalty.

Suppose the selected C-PCR direction $\widehat{u}$ has generalized eigenvalue $z > \lambda$. Its eigen-equation gives

$$\{(z-\lambda)\widehat{\Sigma}_{\mathrm{tot}} + \lambda D_\Delta + z\tau I_p\}\widehat{u} = \widehat{b}_{\mathrm{reg}} (\widehat{b}_{\mathrm{reg}}^{\mathrm{T}} \widehat{u}), \qquad \mu_z = \frac{\lambda}{z-\lambda}, \quad \kappa_z = \frac{z\tau}{z-\lambda}. \tag{29}$$

The matrix on the left is positive definite, so $\widehat{b}_{\mathrm{reg}}^{\mathrm{T}} \widehat{u} \neq 0$ and $\mathrm{span}(\widehat{u}) = \mathrm{span}\{\widehat{\beta}_{\mathrm{con}}(\mu_z, \kappa_z)\}$. This is a correspondence of directions after matching penalties through the sample eigenvalue, not equality of the fitted coefficients. Indeed, for any nonzero $u$ spanning this direction, the C-PCR coefficient refit is

$$\widehat{\beta} = u \frac{u^{\mathrm{T}} \widehat{b}_{\mathrm{reg}}}{u^{\mathrm{T}} (\widehat{\Sigma}_{\mathrm{tot}} + \rho I_p) u}. \tag{30}$$

Applying this same refit, with the same $\rho$ and training data, to the consistency direction therefore gives identical predictions. The raw coefficient in Equation (28) generally does not. This one-component identity does not by itself identify the multicomponent selection-and-refit procedure with a single consistency-regularized regression; nor does it supply the joint high-dimensional prediction-risk limits derived in Sections 4 and D.

**Several responses and the response-rank budget.** For multiple responses, the consistency reduced-rank regression (RRR) comparison uses

$$\min_{\mathrm{rank}(B) \leq q} \left\{ \frac{\|Y - XB\|_F^2 + \|Y - X^+B\|_F^2}{2n} + \mu\, \mathrm{tr}(B^{\mathrm{T}} D_\Delta B) + \kappa \|B\|_F^2 \right\}.$$

Set $A = \widehat{\Sigma}_{\mathrm{tot}} + \mu D_\Delta + \kappa I_p \succ 0$. Its selected space is $\mathrm{col}(A^{-1/2} Q_q)$, where $Q_q$ contains leading left singular vectors of $A^{-1/2} \widehat{B}_{\mathrm{reg}}$. The comparison refits coefficients in that space using the independently chosen $\rho$ in Equation (20).

**Result B.4 (Direction-dependent penalties and response rank)** Let $r_Y = \mathrm{rank}(\widehat{B}_{\mathrm{reg}})$. For any positive-definite $A$, consistency RRR determines at most $r_Y$ response-bearing directions, all in $\mathrm{col}(A^{-1}\widehat{B}_{\mathrm{reg}})$; if $q > r_Y$, any $A$-orthonormal completion has the same selection objective. Each exterior C-PCR direction at root $z_j$ instead belongs to $\mathrm{col}\{A(\mu_{z_j}, \kappa_{z_j})^{-1} \widehat{B}_{\mathrm{reg}}\}$, with penalties from Equation (29). For at least two distinct exterior roots, their penalty pairs coincide if and

only if $\lambda = 0$. In that case selection gives the reduced-rank ridge space, and the fitted coefficients agree when $\rho = \tau$.

*Proof of Result B.4.* Completing the square and whitening by $A^{1/2}$ gives the rank-$r_Y$ matrix $A^{-1/2}\widehat{B}_{\mathrm{reg}}\widehat{B}_{\mathrm{reg}}^{\mathrm{T}}A^{-1/2}$; its remaining eigenvalues are zero. Mapping its eigenspaces back proves the first claim, with the zero-response space $\{u : \widehat{B}_{\mathrm{reg}}^{\mathrm{T}}u = 0\}$ equal to the $A$-orthogonal complement. The selected eigen-equation gives the root-specific column space, as in Equation (29). For $\lambda > 0$, $\mu_z = \lambda/(z - \lambda)$ is strictly decreasing; for $\lambda = 0$, the penalty pair is $(0, \tau)$. The coefficient identity follows from Section B.4. □

Different penalty pairs need not produce different spaces. The distinction is that positive pairing can determine further directions beyond the response rank when the selected spectrum has a strict cutoff gap.

# C Proofs of the main results

## C.1 Population risk and component selection

**Verification of the best linear benchmark.** For a general response coefficient $\theta^*$ in the paired-factor model, the best linear predictor has coefficient and risk

$$\beta^* = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \mathcal{R}(\beta) = U^*\Gamma^{1/2}(I_k + \Gamma)^{-1}\theta^*, \qquad \mathcal{R}^*_{\mathrm{lin}} = \sigma^2 + (\theta^*)^{\mathrm{T}}(I_k + \Gamma)^{-1}\theta^*. \tag{31}$$

*Proof of Equations (11) and (31).* The identities in Equations (11) and (31) use only the population second moments. Since $\Sigma_{\mathrm{tot}} \succ 0$, the unique minimizer satisfies $\Sigma_{\mathrm{tot}}\beta^* = b_{\mathrm{reg}}$. Substituting Equation (3) and using orthonormality of the loadings gives the coefficient in Equation (31). Its risk is $\mathbb{E}[y_{\mathrm{te}}^2] - {b_{\mathrm{reg}}}^{\mathrm{T}}{\Sigma_{\mathrm{tot}}}^{-1}b_{\mathrm{reg}}$, yielding the displayed benchmark risk. Conditional on the training data, completing the square gives Equation (11). Equivalently, the excess risk is $\mathbb{E}[\{x_{\mathrm{te}}^{\mathrm{T}}(\widehat{\beta} - \beta^*)\}^2 \mid \mathcal{D}_n]$. Setting $\theta^* = \vartheta e_1$ recovers Equation (10). □

This expresses the prediction target directly through the shared loadings and response coefficients, as in the latent-factor PCR formulation of Green & Romanov (2025).

*Proof of Proposition 3.1.* Under $\theta^* = \vartheta e_1$, Equation (3) gives $b_{\mathrm{reg}} = U^*\Gamma^{1/2}\vartheta e_1 = \sqrt{a_1}\vartheta u_1^*$, with $\Sigma_{\mathrm{tot}}$ and $\Sigma_{\mathrm{ct}}$ as in Equation (3). The population version of $\widehat{b}_{\mathrm{reg}}$ equals $b_{\mathrm{reg}}$. Hence the pair-only generalized eigenvalues are $\eta_j = a_j/(1 + a_j + \tau)$ on $u_j^*$ and zero elsewhere. The joint numerator $\lambda\Sigma_{\mathrm{ct}} + b_{\mathrm{reg}}{b_{\mathrm{reg}}}^{\mathrm{T}}$ changes only the predictive direction's score, to $(\lambda + \vartheta^2)\eta_1$. If $\eta_2 > \eta_1$, the predictor therefore leads exactly when $(\lambda + \vartheta^2)\eta_1 > \lambda\eta_2$, which is Equation (8). Equality is a tie. Within the retained predictor span, population ridge refitting gives

$$\beta_{\text{C-PCR}} = \frac{\sqrt{a_1}\vartheta}{1 + a_1 + \rho}u_1^*.$$

Every retained shared distractor has zero covariance with $y$, so the population rank-one PCA++-PCR coefficient satisfies $\beta_{\text{PCA++-PCR}} = 0$. Expanding $\mathcal{R}(\beta) = \vartheta^2 + \sigma^2 - 2\beta^{\mathrm{T}}b_{\mathrm{reg}} + \beta^{\mathrm{T}}\Sigma_{\mathrm{tot}}\beta$ gives

$$\mathcal{R}(\beta_{\text{PCA++-PCR}}) - \mathcal{R}(\beta_{\text{C-PCR}}) = \frac{2a_1\vartheta^2}{1 + a_1 + \rho} - \frac{a_1\vartheta^2(1 + a_1)}{(1 + a_1 + \rho)^2} = \frac{a_1\vartheta^2(1 + a_1 + 2\rho)}{(1 + a_1 + \rho)^2} > 0.$$

If an enlarged pair-only space contains $u_1^*$, its remaining population coordinates are uncorrelated with both $u_1^*$ and $y$. Matched ridge refitting assigns them zero coefficients and recovers the same $\beta_{\text{C-PCR}}$. □

For completeness, the population ordering extends directly to higher retained rank. Let $t_q^{(-1)}$ be the $q$th largest competing factor strength in $\{a_j : 2 \le j \le k\} \cup \{b_\ell\}$, taking it to be zero if fewer than $q$ competitors exist. Away from ties, rank-$q$ marginal PCR retains the predictor exactly when

$$a_1 > t_q^{(-1)}. \tag{32}$$

Its variance score is $1 + a_1$, compared with $1 + a_j$, $1 + b_\ell$, and the isotropic score one. For the pair-only ordering, background scores vanish, and the predictor is excluded exactly when at least $q$ shared distractors have larger $\eta_j$. A retained space excluding $u_1^*$ has response risk $\vartheta^2 + \sigma^2$ after optimal population refitting. These are truncation errors at the population level; sample directions and fitted coefficients introduce additional estimation errors.

**Wider pair-only preselection.** The same counting applies to wider PCA++-PCR and to C-PCR. Collect the shared distractors that outrank $u_1^*$ under each score:

$$\mathcal{N} = \{j \geq 2 : \eta_j > \eta_1\}, \qquad \mathcal{N}_\lambda = \{j \geq 2 : \lambda\eta_j > (\lambda + \vartheta^2)\eta_1\} \subseteq \mathcal{N}.$$

Away from ties, the predictive direction $u_1^*$ is then retained exactly when

$$\begin{array}{ll} |\mathcal{N}| < q & \text{for rank-}q\text{ PCA++-PCR,} \\ |\mathcal{N}| < q_0 & \text{for wider PCA++-PCR with } q_0 \text{ initial directions,} \\ |\mathcal{N}_\lambda| < q & \text{for rank-}q\text{ C-PCR.} \end{array}$$

Whenever $u_1^*$ is retained, the refit returns $\beta_{\text{C-PCR}}$: the other retained directions are orthogonal to $u_1^*$ and have no response association, so they receive zero coefficients. Hence, when rank-$q$ C-PCR retains $u_1^*$, wider PCA++-PCR matches it in the population once $q_0 > |\mathcal{N}|$, and any advantage of C-PCR must come from finite-sample estimation. Section E.4 compares the two in finite samples.

### C.2 A DIRECT SCALAR RISK PROOF

**More details on the excess-risk decomposition.** We derive the finite-sample decomposition in Equation (15) and explain how its terms converge.

*Proof of Equation (15).* Substitute Equations (3) and (10) into Equation (11) and decompose $\widehat{\beta}$ along $u_1^*$ and its orthogonal complement. Orthonormality of the shared and background loadings gives the four displayed terms. □

In Equation (15), with $\widehat{b} = u_1^{*\mathrm{T}}\widehat{\beta}$ and $\widehat{\beta}_\perp = \widehat{\beta} - \widehat{b}\,u_1^*$, the first term is the squared coefficient error $\{\widehat{b} - \sqrt{a}\,\vartheta/(1+a)\}^2$, multiplied by the test variance along $u_1^*$. That variance is $1 + a$, because the model gives

$$(u_1^*)^{\mathrm{T}} x_{\mathrm{te}} = \sqrt{a}\, w_{\mathrm{te},1} + (u_1^*)^{\mathrm{T}} \epsilon_{\mathrm{te}}, \qquad \mathrm{Var}\{(u_1^*)^{\mathrm{T}} x_{\mathrm{te}}\} = a + 1,$$

where the two summands are independent and have variances $a$ and one. The last two sums in Equation (15) account for the additional variance along the nonpredictive shared directions $u_j^*$, $j \geq 2$, and the background directions $v_\ell^*$. These directions have zero population covariance with the response; the theorem shows that sampling fluctuations leave no nonzero limiting fitted coefficient along them. Since their number and strengths are fixed, both sums vanish.
In contrast, the remaining squared norm satisfies

$$\|\widehat{\beta}_\perp\|_2^2 = \|\widehat{\beta}\|_2^2 - \widehat{b}^2 \longrightarrow v_* - b_*^2,$$

which need not be zero: there are increasingly many noise directions in which small estimation errors can accumulate. For example, a vector with $p$ entries equal to $p^{-1/2}$ has every entry tending to zero but squared norm equal to one. This explains why vanishing coefficients in any fixed collection of nonpredictive directions do not imply vanishing total error outside $u_1^*$.
We first record the noise and quadratic-form limits used in both risk proofs. All limits in this appendix are in probability. Write $\zeta = 2\gamma$, $c_\pm(z) = z \pm \lambda$, and $t_z = 2z\tau$.

**Lemma C.1 (Paired-noise inverse and finite latent blocks)** Under the feature-noise conditions of Assumption 2.1, let $E_-$ and $E_+$ have rows $(\epsilon_i + \epsilon_i^+)^{\mathrm{T}}/\sqrt{2n}$ and $(\epsilon_i - \epsilon_i^+)^{\mathrm{T}}/\sqrt{2n}$, and put

$$Q_z = \{t_z I + c_-(z)E_-^{\mathrm{T}}E_- + c_+(z)E_+^{\mathrm{T}}E_+\}^{-1}. \tag{33}$$

For deterministic bounded vectors $u, v$, uniformly on compact subsets of $z > \lambda, \tau > 0$,

$$\frac{\mathrm{tr}\, Q_z}{p} \to m(z), \quad u^{\mathrm{T}}Q_z v - m(z)u^{\mathrm{T}}v \to 0, \quad \frac{1}{n}\mathrm{tr}(E_\iota Q_z E_\iota^{\mathrm{T}}) \to \zeta m(z) r_\iota(z), \tag{34}$$

where $\iota \in \{-, +\}$ indexes the sum and difference arrays, and $m, r_\pm$ solve Equation (16). The convergence extends to small complex neighborhoods and derivatives. Independently of these noise arrays, let $a_i, b_i$ be coordinates of centered iid finite-dimensional rows with uniformly bounded fourth moments. For a noise-measurable matrix $T$ with $\|T\|_{\mathrm{op}} = O_{\mathrm{Pr}}(1)$,

$$\frac{a^{\mathrm{T}}Tb}{n} - \mathbb{E}[a_i b_i]\frac{\mathrm{tr}\, T}{n} \to 0. \tag{35}$$

Coordinates within each latent row may be dependent, as may the two noise rows of a pair; the latent block remains independent of the noise arrays.

*Proof of Lemma C.1.* Let $V_i = (e_{-,i}, e_{+,i})$ contain the transposed noise rows, and delete the whole pair to obtain $Q_z^{(i)}$. Independent original coordinates give, for an independent test matrix $T$,

$$\mathbb{E}_i \left| e_{\iota,i}^{\mathrm{T}} T e_{\iota',i} - \mathbf{1}\{\iota = \iota'\} \operatorname{tr} T/n \right|^2 \le C\|T\|_F^2/n^2, \qquad \mathbb{E}|u^{\mathrm{T}} e_{\iota,i}|^4 \le C\|u\|^4/n^2. \tag{36}$$

Here $\mathbb{E}_i$ conditions on all other pairs. The same bounds hold for complex or nonsymmetric tests. With $C_z = \operatorname{diag}(c_-, c_+)$ and $K_i = V_i^{\mathrm{T}} Q_z^{(i)} V_i$,

$$Q_z V_i = Q_z^{(i)} V_i (I_2 + C_z K_i)^{-1}, \qquad (I_2 + C_z^{1/2} K_i C_z^{1/2})^{-1} \preceq I_2.$$

The linear-form bound makes each squared pair-replacement difference of a bounded deterministic bilinear test $O(n^{-2})$. Summing over independent pairs gives variance $O(n^{-1})$; bounded rank gives the same order for normalized traces. Thus, with $\bar{e} = n^{-1} \operatorname{tr} \mathbb{E} Q_z$, $\mathbb{E}\|K_i - \bar{e} I_2\|_F^2 = O(n^{-1})$. In the expected resolvent identity, replace the update denominator by $I_2 + C_z \bar{e}$. For each unit-vector bilinear test, Cauchy–Schwarz combines an $O(n^{-1})$ linear-product bound in $L^2$ with the $O(n^{-1/2})$ denominator error, giving $O(n^{-3/2})$ per pair. Conditioning the positive rank-two update also gives $\|\mathbb{E} Q_z^{(i)} - \mathbb{E} Q_z\|_{\mathrm{op}} = O(n^{-1})$. Consequently,

$$\left\| \mathbb{E} Q_z - \left( t_z + \sum_\iota \frac{c_\iota}{1 + c_\iota \bar{e}} \right)^{-1} I_p \right\|_{\mathrm{op}} = O(n^{-1/2}).$$

The trace equation $1 = t_z m + \sum_\iota c_\iota m/(1 + \zeta c_\iota m)$ has a unique positive solution because its right side is strictly increasing. This proves the first two limits; averaging diagonal entries of the pair-update identity proves the weighted trace limit. These are averaged error bounds, not simultaneous rowwise concentration.

For Equation (35), conditional expansion over observation indices gives mean zero and variance at most $C\|T\|_F^2/n^2$. First restrict to bounded-norm events and then let the bound grow. Within-row coordinate dependence is allowed.

We use the following analytic argument for derivative limits. On an exterior compact set, $\|Q_z\| \le t_z^{-1}$, $\|Q_z E_\iota^{\mathrm{T}}\| \le (t_z c_\iota)^{-1/2}$, and $\|E_\iota Q_z E_{\iota'}^{\mathrm{T}}\| \le (c_\iota c_{\iota'})^{-1/2}$. Positive real parts preserve these bounds up to fixed constants in a small complex neighborhood of $(z, \tau)$. From any subsequence, pointwise convergence on a countable dense real set gives an almost-surely convergent further subsequence. Local analytic boundedness and uniqueness then give uniform convergence on smaller neighborhoods; Cauchy's formula gives derivatives. The same argument applies to the response kernels below once their local bounds are established. □

**Admissible pair weights.** The scalar response map in Equation (17) crosses one above $\lambda$ exactly when

$$\lim_{z \downarrow \lambda} \phi(z) > 1. \tag{37}$$

For $\lambda > 0$, write $A = 1 + \tau - 2\gamma$. At the boundary, $r_- = 1$ and $m = c/\lambda$, where the bulk equation gives

$$8\gamma\tau c^2 + 2Ac = 1, \qquad c = \{A + \sqrt{A^2 + 8\gamma\tau}\}^{-1}, \qquad \phi(\lambda+) = \frac{(2\gamma s + 2a\vartheta^2)c}{\lambda} = \frac{\lambda_{\mathrm{ex}}}{\lambda}.$$

At $\lambda = 0$, substituting $m = C_0/z$ gives a positive constant $C_0$, constant $r_\pm$, and $\phi(z) \to \infty$ as $z \downarrow 0$. Thus Equation (12) is exactly the crossing condition. Strict decrease follows from the proof below.

*Proof of Theorem 4.2 (direct).* Use Equation (24) with $Y = y$, write $\phi_n = \Phi_n$, $Q = Q_z$, and take $F = (U^*, W^*)$. The scalar and matrix response maps use exactly the same normalization. For this fixed $z > \lambda$, abbreviate $m = m(z)$, $r_\iota = r_\iota(z)$, $c_\iota = c_\iota(z)$, and $t = t_z$.

*Retaining the latent factors and response.* Let $H_-, H_+ \in \mathbb{R}^{n \times (k+m)}$ have rows

$$(H_-)_{i,:} = n^{-1/2} \big( \sqrt{2}\, w_i^{\mathrm{T}} \Gamma^{1/2}, (h_i + h_i^+)^{\mathrm{T}} \Delta^{1/2}/\sqrt{2} \big),$$
$$(H_+)_{i,:} = n^{-1/2} \big( 0, (h_i - h_i^+)^{\mathrm{T}} \Delta^{1/2}/\sqrt{2} \big).$$

Thus $Z/\sqrt{n} = E_- + H_- F^{\mathrm{T}}$ and $D/\sqrt{n} = E_+ + H_+ F^{\mathrm{T}}$. Put $y_0 = y/\sqrt{n}$ and define

$$S_- = \operatorname{diag}(2\Gamma, \Delta), \qquad S_+ = \operatorname{diag}(0_k, \Delta), \qquad b = \sqrt{2a}\, \vartheta\, e_1 \in \mathbb{R}^{k+m}.$$

The finite latent block is independent of the feature noises. Its rows have bounded fourth moments and second moments

$$H_\iota^\top H_\iota \to S_\iota, \quad H_-^\top H_+ \to 0, \quad H_-^\top y_0 \to b, \quad H_+^\top y_0 \to 0, \quad y_0^\top y_0 \to s.$$

These are laws of large numbers, not assertions of independence between the transformed latent variables. Apply Equation (35) to this entire latent block. The required kernels obey $\|E_\iota Q E_{\iota'}^\top\| \le (c_\iota c_{\iota'})^{-1/2}$; linear tests involving $F^\top Q E_\iota^\top$ vanish by their conditional second moments.
Set

$$T = c_- E_-^\top H_- + c_+ E_+^\top H_+, \qquad g_\epsilon = E_-^\top y_0,$$
$$R_n = c_- H_-^\top H_- + c_+ H_+^\top H_+, \qquad b_n = H_-^\top y_0.$$

The covariance pencil and response vector satisfy the exact identities

$$M_n = Q^{-1} + F R_n F^\top + F T^\top + T F^\top, \qquad g_n = F b_n + g_\epsilon. \tag{38}$$

This also verifies the normalized trace interpretation used in the main-text calculation. The difference $M_n - Q^{-1}$ has rank at most $2(k+m)$, and both inverse norms are at most $t^{-1}$. Hence $p^{-1}\operatorname{tr} M_n^{-1} - p^{-1}\operatorname{tr} Q \to 0$. For the weighted traces, stack $\sqrt{c_-}Z/\sqrt{n}$ and $\sqrt{c_+}D/\sqrt{n}$ into a matrix $\mathcal{X}$. Its difference from the corresponding pure-noise matrix $\mathcal{X}_0$ has rank at most $k+m$. The matrices $\mathcal{X}(tI + \mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top$ and its pure-noise counterpart lie between zero and the identity, and their difference has rank at most $2(k+m)$. Taking the traces of their two diagonal blocks and using Equation (34) therefore gives

$$\frac{\operatorname{tr}(Z M_n^{-1} Z^\top)}{np} \to m r_-, \qquad \frac{\operatorname{tr}(D M_n^{-1} D^\top)}{np} \to m r_+.$$

Combining the preceding conditional bounds with Equation (34) gives

$$F^\top Q T \to 0, \quad F^\top Q g_\epsilon \to 0, \quad g_\epsilon^\top Q g_\epsilon \to s\zeta m r_-,$$
$$T^\top Q T \to \Lambda := \zeta m(c_-^2 r_- S_- + c_+^2 r_+ S_+), \qquad T^\top Q g_\epsilon \to \zeta m c_- r_- b.$$

The cross terms vanish because the corresponding latent covariances are zero; independence of $E_-$ and $E_+$ is unnecessary.
Write $l_n = F^\top M_n^{-1} g_n$ and $j_n = T^\top M_n^{-1} g_n$. Projecting Equation (38) yields a finite linear system whose limiting equations are

$$l = m\{b - (c_- S_- + c_+ S_+) l - j\}, \qquad j = \zeta m c_- r_- b - \Lambda l.$$

Indeed, $x_n = M_n^{-1} g_n$ satisfies $Q^{-1} x_n = g_n - (F R_n F^\top + F T^\top + T F^\top) x_n$ by Equation (38), that is, $x_n = Q\{F(b_n - R_n l_n - j_n) + g_\epsilon - T l_n\}$; applying $F^\top$ and $T^\top$ and the limits above gives these equations. Its Schur complement is $I + m(c_- r_- S_- + c_+ r_+ S_+) \succ 0$, so the limiting system is invertible and substitution of the converging coefficients is justified. Since $1 - \zeta m c_- r_- = r_-$, it gives

$$l = r_-\{m^{-1} I + c_- r_- S_- + c_+ r_+ S_+\}^{-1} b.$$

Only its predictive coordinate is nonzero. Consequently,

$$\begin{aligned} F^\top M_n^{-1} g_n &\longrightarrow \frac{\sqrt{2a}\,\vartheta\, r_-}{m^{-1} + 2a c_- r_-}\, e_1, \\ g_n^\top M_n^{-1} g_n &\longrightarrow s\zeta m r_- + \frac{2a\vartheta^2 r_-^2}{m^{-1} + 2a c_- r_-}. \end{aligned} \tag{39}$$

These are exactly $\alpha(z)e_1$ and $\phi(z)$ from Equation (17).
*Selection and coefficient fitting.* The full pencil also has a positive real part locally, and

$$|\phi_n(z,\tau)| \le \frac{\|y/\sqrt{n}\|^2}{\operatorname{Re} c_-}, \qquad \|F^\top M_n^{-1} g_n\| \le \frac{\|y/\sqrt{n}\|}{\sqrt{\operatorname{Re} t\, \operatorname{Re} c_-}}.$$

The analytic argument in Lemma C.1 gives locally uniform limits and derivatives in $z, \tau$. By Lemma B.3, $-\phi'(z) \ge \phi(z)/(z+\lambda) > 0$, and $\phi_n(z) \le (y^\top y/n)/(z-\lambda)$ prevents roots escaping to infinity. The boundary condition above therefore gives a unique simple root $z_*$, and the sample root $\widehat{z} \to z_*$.

Put $\widetilde{u}_n = M_n(\widehat{z})^{-1} g_n$ and $d_n = -\phi_n'(\widehat{z})$. Taking the $\tau$ derivative at fixed $z$ gives $\partial_\tau M_n(z) = 2zI$; substituting the selected root afterwards yields

$$\|\widetilde{u}_n\|^2 = -\frac{\partial_\tau \phi_n(\widehat{z})}{2\widehat{z}} \to \chi_*, \qquad d_n \to d_*, \qquad F^{\mathrm{T}} \widetilde{u}_n \to \alpha(z_*) e_1.$$

Here $d_* \geq (z_* + \lambda)^{-1} > 0$ and $\chi_* \geq \alpha(z_*)^2 > 0$. The normalization and refit identities in Lemma B.3 give directly

$$\widehat{\beta} = \frac{\sqrt{2}\,\widetilde{u}_n}{d_n + 2(\rho - \tau)\|\widetilde{u}_n\|^2}, \qquad d_* + 2(\rho - \tau)\chi_* > 0.$$

Consequently $F^{\mathrm{T}}\widehat{\beta} \to b_* e_1$ and $\|\widehat{\beta}\|^2 \to v_*$, with the quantities in Equation (18). Finally,

$$\mathcal{R}(\widehat{\beta}) = s - 2\sqrt{a}\,\vartheta (u_1^*)^{\mathrm{T}} \widehat{\beta} + \|\widehat{\beta}\|^2 + \widehat{\beta}^{\mathrm{T}} F \operatorname{diag}(\Gamma, \Delta) F^{\mathrm{T}} \widehat{\beta}.$$

Substitution proves Equation (14); subtracting the linear benchmark and completing the square gives Equation (13). The term $v_* - b_*^2$ is nonnegative as the limit of $\|(I_p - u_1^*(u_1^*)^{\mathrm{T}})\widehat{\beta}\|^2$. □

### C.3 A SECOND PROOF FROM THE MULTIPLE-RESPONSE THEOREM

*Proof of Theorem 4.2 (via the multiple-response theorem).* In Theorem D.2, take $L = q = 1$, $\Theta^* = \vartheta e_1$, and $\Sigma_\xi = \sigma^2$, retaining all $k$ shared and $m$ background factors. Then $\Omega = s$, $B_0 = \sqrt{2a}\,\vartheta e_1$, and $\mathcal{T}_z$ in Equation (48) is diagonal. Hence $A_z = \alpha(z) e_1$ and $\Phi(z) = \phi(z)$. The threshold calculation Equation (37) supplies the simple exterior root under the original scalar assumptions.
The exact identity $g_n^{\mathrm{T}} M_n(z)^{-2} g_n = -(2z)^{-1} \partial_\tau \Phi_n(z)$ gives $\Xi(z_*, z_*) = \chi_*$. Thus

$$H = \sqrt{2/d_*}\,\alpha(z_*) e_1, \quad J = d_*^{-1/2}, \quad E = 2\chi_*/d_*, \quad H\Theta_1 = b_* e_1, \quad \Theta_1^2 E = v_*.$$

With $V = \operatorname{diag}(I_k + \Gamma, I_m + \Delta)$ and $B_F^* = \sqrt{a}\,\vartheta(1+a)^{-1} e_1$, the risk decomposition in Equation (46) is precisely $(1+a)\{b_* - \sqrt{a}\,\vartheta/(1+a)\}^2 + v_* - b_*^2$. This proves the excess-risk formula and, after adding $\mathcal{R}_{\mathrm{lin}}^* = s - a\vartheta^2/(1+a)$, the total-risk formula. □

### C.4 LOCAL EFFECT OF THE PAIR WEIGHT

Write $\mathcal{R}_\infty(\lambda, \tau, \rho; \gamma)$ for the limit $\mathcal{R}_\infty(\lambda, \tau, \rho)$ in Equation (14), with its dependence on the dimension ratio $\gamma$ made explicit.

**Result C.2 (Local effect of pairing)** Fix $a, \vartheta, \tau > 0$, $\sigma^2 \geq 0$, and $\rho \geq 0$. As $\gamma \to 0$, $\mathcal{R}_\infty(\lambda, \tau, \rho; \gamma)$ converges for every admissible $\lambda$ to the population risk $s - a\vartheta^2(1 + a + 2\rho)/(1 + a + \rho)^2$, and

$$\partial_\lambda \mathcal{R}_\infty(0, \tau, \rho; \gamma) = -\gamma C + O(\gamma^2), \qquad C = \frac{2\sigma^2(1 + a + \tau)\,P}{\vartheta^2(1+\tau)^2(1+a+\rho)^3}, \tag{40}$$

where $P = a^2 + a\{2(1 + \rho + \rho^2) + \tau(1 - \rho)\} + (1 + \rho)^2(1 + \tau)$. Hence, for small $\gamma$ and $\sigma^2 > 0$, a small pair weight lowers the limiting risk when $P > 0$, in particular when $\rho \leq 1$ or $\rho = \tau$.

*Proof of Result C.2.* All quantities in Equations (16) to (18) are analytic in $(\gamma, \lambda)$ near $(0, 0)$. At $\gamma = 0$, $r_\pm = 1$ and

$$z_* = \frac{a(\vartheta^2 + \lambda)}{1 + a + \tau}, \qquad d_* = \frac{1 + a + \tau}{a\vartheta^2}, \qquad \chi_* = \frac{1}{2a\vartheta^2}, \qquad \alpha(z_*) = \frac{1}{\sqrt{2a}\,\vartheta},$$

so $b_* = \sqrt{a}\,\vartheta/(1 + a + \rho)$, $v_* = b_*^2$, and the risk is the population risk for every $\lambda$. Differentiating Equation (16) gives $\partial_\gamma m = 4m^3(z^2 + \lambda^2)$ at $\gamma = 0$; propagating it through Equations (17) and (18) gives, at $(\gamma, \lambda) = (0, 0)$,

$$\partial_\lambda \partial_\gamma d_* = 0, \qquad \partial_\lambda \partial_\gamma \chi_* = -\frac{\sigma^2(1 + a + \tau)}{a\vartheta^6(1+\tau)^2}, \qquad \partial_\lambda \partial_\gamma \alpha(z_*) = \frac{\sigma^2(1 + a + \tau)}{\sqrt{2}\,a^{3/2}\vartheta^5(1+\tau)}.$$

Substituting in $b_* = \sqrt{2}\,\alpha(z_*)/D$ and $v_* = 2\chi_*/D^2$, with $D = d_* + 2(\rho - \tau)\chi_*$, and then in Equation (14), gives $\partial_\lambda \partial_\gamma \mathcal{R}_\infty = -C$. If $\rho \leq 1$, the coefficient of $a$ in $P$ is at least 2; if $\rho = \tau$, it equals $2 + 3\rho + \rho^2$. □

The effect is proportional to $\gamma\sigma^2$: a small pair weight reduces the energy $v_* - b_*^2$ that the fitted coefficient spreads over noise directions, and pairing has no effect at $\gamma = 0$. We checked Equation (40) against the scalar steps in 50-digit arithmetic, including a case with $P < 0$.

# D PREDICTION RISK WITH MULTIPLE RESPONSES

The paired-factor model extends to vector responses while preserving the sources of variation in Equation (1). This section gives the resulting joint prediction risk and its proof.

## D.1 MODEL AND PREDICTION RISK

**Model and dimensions.** Keep the two feature equations in Equation (1) and replace the scalar response coefficient by a matrix:

$$x_i = U^*\Gamma^{1/2}w_i + W^*\Delta^{1/2}h_i + \epsilon_i, \quad x_i^+ = U^*\Gamma^{1/2}w_i + W^*\Delta^{1/2}h_i^+ + \epsilon_i^+, \quad y_i = \Theta^{*\mathrm{T}}w_i + \xi_i. \tag{41}$$

Here $w_i \in \mathbb{R}^k$, $y_i, \xi_i \in \mathbb{R}^L$, and $\Theta^* \in \mathbb{R}^{k\times L}$; the feature and background dimensions are unchanged. The scalar entry $\Theta^*_{j\ell}$ is the coefficient of factor $w_{ij}$ in response $y_{i\ell}$, so

$$y_{i\ell} = \sum_{j=1}^{k} \Theta^*_{j\ell} w_{ij} + \xi_{i\ell}, \qquad \ell = 1, \dots, L.$$

Each response may depend on several coordinates of the single vector $w_i$. Its coefficient column is $\Theta^*_{:,\ell} \in \mathbb{R}^k$. If $W_{\mathrm{lat}} \in \mathbb{R}^{n\times k}$, $Y \in \mathbb{R}^{n\times L}$, and $\Xi \in \mathbb{R}^{n\times L}$ collect $w_i^{\mathrm{T}}, y_i^{\mathrm{T}}, \xi_i^{\mathrm{T}}$ by row, the response equation is $Y = W_{\mathrm{lat}}\Theta^* + \Xi$. The latent matrix $W_{\mathrm{lat}}$ differs from the background loading matrix $W^*$. Setting $L = 1$ and $\Theta^* = \theta^*$ recovers Equation (1); setting $\theta^* = \vartheta e_1$ further recovers the prediction regime in Assumption 4.1.

**Assumption D.1 (Paired-factor structure with multiple responses)** (i) The feature variables and loadings satisfy Assumption 2.1. The scalar observation noise in that assumption is replaced by a centered vector $\xi_i \in \mathbb{R}^L$, independent of all factors and feature noises, with covariance $\Sigma_\xi \succeq 0$. Its coordinates may be dependent. Observations are iid within each $(n, p)$, and all primitive scalar fourth moments are uniformly bounded. The response covariance

$$\Omega = \Theta^{*\mathrm{T}}\Theta^* + \Sigma_\xi$$

is positive definite.

(ii) As $p, n \to \infty$, $p/(2n) \to \gamma \in (0, \infty)$. The dimensions $k, m, L$, the retained rank $1 \le q \le L$, and $\Gamma, \Delta, \Theta^*, \Sigma_\xi$ are fixed. The loadings may vary with $p$, preserving orthonormality. Primitive distributions may vary subject to the stated moments and independence. Means are known and zero, and $\lambda \ge 0, \tau > 0, \rho \ge 0$ are fixed.

The assumptions describe the same sources of variation as the scalar model: shared factors carry response information, and independent backgrounds and feature noises contribute variation. Responses can be correlated through the shared factors and through $\Sigma_\xi$. No independence is imposed after taking sums and differences of the views or projecting factors onto responses. The proof keeps each observation pair and its latent variables together; finite fourth moments control their quadratic forms. The theorem below states the multiple-response counterpart of the scalar exterior-root condition.

**Prediction target and linear benchmark.** For an independent test unit with the training distribution and a fitted coefficient matrix $B \in \mathbb{R}^{p\times L}$, the multiple-response version of Equation (2) is

$$\mathcal{R}(B) = \mathbb{E}\left[\|y_{\mathrm{te}} - B^{\mathrm{T}}x_{\mathrm{te}}\|_2^2 \mid \mathcal{D}_n\right],$$

with $B$ held fixed conditional on $\mathcal{D}_n$. Write $F = (U^*, W^*) \in \mathbb{R}^{p\times r}$, $r = k + m$, and $K_x = \mathrm{diag}(\Gamma, \Delta)$. The test moments are

$$C_{xy} = U^*\Gamma^{1/2}\Theta^*, \qquad \Sigma_{\mathrm{tot}} = I_p + FK_xF^{\mathrm{T}}, \qquad V = F^{\mathrm{T}}\Sigma_{\mathrm{tot}}F = I_r + K_x. \tag{42}$$

The optimal linear coefficient is $FB_F^*$, where

$$B_F^* = \begin{pmatrix} (I_k + \Gamma)^{-1}\Gamma^{1/2}\Theta^* \\ 0_{m\times L} \end{pmatrix}, \qquad \mathcal{R}_{\mathrm{lin}}^* = \mathrm{tr}\,\Omega - \|V^{1/2}B_F^*\|_F^2. \tag{43}$$

Completing the square gives the exact identity

$$\mathcal{R}(B) - \mathcal{R}_{\mathrm{lin}}^* = \|V^{1/2}(F^{\mathrm{T}}B - B_F^*)\|_F^2 + \|(I_p - FF^{\mathrm{T}})B\|_F^2. \tag{44}$$

Thus prediction error within the factor space is weighted by its test covariance, while the remaining coefficient error is weighted by the identity.

Use the sample response map $\Phi_n$ and metric-normalized eigenvectors $\widehat{U}_q$ from Lemma B.3. Its deterministic limit $\Phi$, and the matrices $H, E, J$ below, are computed in Section D.2.

**Theorem D.2 (Prediction risk with multiple responses)** Under Assumption D.1, suppose the largest $q$ roots of $\det\{I_L - \Phi(z)\} = 0$ are distinct and simple, with $z_1 > \cdots > z_q > \lambda$ and a positive gap at the selected-rank cutoff. The deterministic matrices $H \in \mathbb{R}^{r\times q}$, $E \in \mathbb{R}^{q\times q}$, and $J \in \mathbb{R}^{q\times L}$ in Equation (52) satisfy, after consistent choices of component signs,

$$F^{\mathrm{T}}\widehat{U}_q \longrightarrow H, \qquad \widehat{U}_q^{\mathrm{T}}\widehat{U}_q \longrightarrow E, \qquad \widehat{U}_q^{\mathrm{T}}\widehat{B}_{\mathrm{reg}} \longrightarrow J \quad \text{in probability.}$$

The matrix $N = E - H^{\mathrm{T}}H$ is positive semidefinite, and $I_q + (\rho - \tau)E$ is positive definite, including at $\rho = 0$. Define

$$\Theta_q = \{I_q + (\rho - \tau)E\}^{-1}J. \tag{45}$$

Then $\widehat{\Theta}_q \to \Theta_q$, and the conditional prediction risk converges in probability to

$$\mathcal{R}(\widehat{B}_q) \longrightarrow \mathcal{R}_q = \mathcal{R}^*_{\mathrm{lin}} + \underbrace{\|V^{1/2}(H\Theta_q - B^*_F)\|_F^2}_{\text{error within the factor space}} + \underbrace{\mathrm{tr}(\Theta_q^{\mathrm{T}}N\Theta_q)}_{\text{error outside it}}. \tag{46}$$

Here $H$ records alignment with the factor loadings, $E$ records Euclidean inner products between the selected directions, and $J$ records their training response associations. The matrix $\Theta_q \in \mathbb{R}^{q\times L}$ describes coefficient fitting on the selected scores; it differs from the population factor coefficient $\Theta^* \in \mathbb{R}^{k\times L}$. The root condition keeps these directions separated above the pair-only spectrum. Section D.3 proves the result directly under the paired-factor assumptions, and setting $L = q = 1$ and $\Theta^* = \vartheta e_1$ gives the scalar theorem; Section C.3 carries out this specialization.
For example, when $q = 2$, the last risk term is

$$N_{11}\|(\Theta_q)_{1,:}\|_2^2 + N_{22}\|(\Theta_q)_{2,:}\|_2^2 + 2N_{12}\langle(\Theta_q)_{1,:}, (\Theta_q)_{2,:}\rangle.$$

Orthogonality in the selection metric does not force $N_{12} = 0$, so several retained directions must be fitted and evaluated jointly. Even at $\rho = \tau$, when $\Theta_q = J$, the cross term can remain.

### D.2 Computing the risk

Put $\zeta = 2\gamma$. The finite-dimensional factor covariances and factor–response cross-covariance are

$$S_- = \mathrm{diag}(2\Gamma, \Delta), \qquad S_+ = \mathrm{diag}(0_k, \Delta), \qquad B_0 = \begin{pmatrix} \sqrt{2}\,\Gamma^{1/2}\Theta^* \\ 0_{m\times L} \end{pmatrix}. \tag{47}$$

For $z > \lambda$, compute $m(z), r_\pm(z)$ from Equation (16), and set $c_\pm(z) = z \pm \lambda$, $t_z = 2z\tau$. All sums below range over $\iota \in \{-, +\}$. Define the $r \times r$ matrix $\mathcal{T}_z$ and the $r \times L$ factor-alignment map $A_z$ by

$$\mathcal{T}_z = m(z)^{-1}I_r + \sum_\iota c_\iota(z)r_\iota(z)S_\iota, \qquad A_z = r_-(z)\mathcal{T}_z^{-1}B_0. \tag{48}$$

The $L \times L$ response map is

$$\Phi(z) = \zeta m(z)r_-(z)\Omega + r_-(z)B_0^{\mathrm{T}}A_z. \tag{49}$$

The first term accounts for response association with estimated noise directions; the second includes the shared factors. This map is symmetric positive definite. Its eigenvalues crossing one identify the selected exterior roots.
To evaluate inner products between directions at different roots, define

$$\psi(z, w) = \left\{ \frac{1}{m(z)m(w)} - \zeta \sum_\iota c_\iota(z)c_\iota(w)r_\iota(z)r_\iota(w) \right\}^{-1}. \tag{50}$$

The bulk equation gives $m(z)/\psi(z, w) = t_w + \sum_\iota c_\iota(w)r_\iota(z)r_\iota(w) > 0$, so $\psi$ is positive and symmetric. Set

$$C_-(z, w) = \Omega - c_-(w)B_0^{\mathrm{T}}A_w - c_-(z)A_z^{\mathrm{T}}B_0 + c_-(z)c_-(w)A_z^{\mathrm{T}}S_-A_w,$$
$$C_+(z, w) = c_+(z)c_+(w)A_z^{\mathrm{T}}S_+A_w.$$

The required response kernel is

$$\Xi(z, w) = \frac{\psi(z, w)}{m(z)m(w)}A_z^{\mathrm{T}}A_w + \zeta\psi(z, w)\sum_\iota r_\iota(z)r_\iota(w)C_\iota(z, w). \tag{51}$$

The two response–factor cross terms in $C_-$ account for using the same responses in selection and fitting. Since $\Xi(z,w)^\top = \Xi(w,z)$, the resulting Gram matrix is symmetric.
At each root $z_j$, choose a unit $a_j \in \mathbb{R}^L$ satisfying $\Phi(z_j)a_j = a_j$ and put $d_j = -a_j^\top \Phi'(z_j)a_j > 0$, holding model and tuning parameters fixed in the derivative. Then

$$J_{j,:} = \frac{a_j^\top}{\sqrt{d_j}}, \qquad H_{:,j} = \sqrt{\frac{2}{d_j}} A_{z_j} a_j, \qquad E_{jk} = \frac{2a_j^\top \Xi(z_j, z_k) a_k}{\sqrt{d_j d_k}}. \tag{52}$$

Substituting these matrices in Equations (45) and (46) completes the risk calculation.

### D.3 Proof of the multiple-response theorem

The one-inverse limits and finite latent-block quadratic forms are supplied by Lemma C.1. The extra noise calculation needed for several selected directions is a product of inverses at two spectral arguments.

**Lemma D.3 (Products of paired-noise inverses)** For $Q_z$ from Equation (33), deterministic bounded vectors $u, v$, and $z, w > \lambda$,

$$\begin{gathered} u^\top Q_z Q_w v - \psi(z,w) u^\top v \to 0, \qquad p^{-1}\operatorname{tr}(Q_z Q_w) \to \psi(z,w), \\ \frac{1}{n}\operatorname{tr}(E_\iota Q_z Q_w E_\iota^\top) \to \zeta \psi(z,w) r_\iota(z) r_\iota(w). \end{gathered} \tag{53}$$

These limits hold locally uniformly, including derivatives in $z, w, \tau$, under the same noise assumptions as Lemma C.1.

*Proof of Lemma D.3.* Introduce a scalar coupling $\delta$ near zero and the inverse

$$\mathcal{Q}_\delta = \begin{pmatrix} Q_z^{-1} & \delta I_p \\ \delta I_p & Q_w^{-1} \end{pmatrix}^{-1}, \qquad T_\delta = \begin{pmatrix} t_z & \delta \\ \delta & t_w \end{pmatrix}.$$

For $|\delta| < \min(t_z, t_w)/2$, the matrix is uniformly positive definite. With $f_1, f_2$ the standard basis of $\mathbb{R}^2$, one observation pair contributes four columns,

$$\mathcal{V}_i = (f_1 \otimes e_{-,i}, f_2 \otimes e_{-,i}, f_1 \otimes e_{+,i}, f_2 \otimes e_{+,i}), \quad \mathcal{C} = \operatorname{diag}(C_-, C_+), \quad C_\iota = \operatorname{diag}(c_\iota(z), c_\iota(w)).$$

Delete these columns to obtain $\mathcal{Q}_\delta^{(i)}$, and set $K_i = \mathcal{V}_i^\top \mathcal{Q}_\delta^{(i)} \mathcal{V}_i$. Its conditional mean has entries

$$\mathbb{E}_i(K_i)_{\iota u, \iota' v} = \mathbf{1}\{\iota = \iota'\} \frac{\operatorname{tr}[\mathcal{Q}_\delta^{(i)}]_{uv}}{n}, \qquad u, v \in \{1, 2\}.$$

Let $\bar{R}_{uv} = p^{-1} \operatorname{tr} \mathbb{E}[\mathcal{Q}_\delta]_{uv}$, $\zeta_n = p/n$, and $K_0 = \operatorname{diag}(\zeta_n \bar{R}, \zeta_n \bar{R})$. The quadratic-form bounds in Equation (36), pair-replacement trace concentration, and rank-four deletion imply $\|K_i - K_0\|_{L^2} = O(n^{-1/2})$. In particular, opposite noise signs contribute zero conditional mean even though they are dependent.
For $D_i = (\mathcal{C}^{-1} + K_i)^{-1}$ and $D_0 = (\mathcal{C}^{-1} + K_0)^{-1}$,

$$0 \preceq D_i, D_0 \preceq \mathcal{C}, \qquad D_i - D_0 = -D_i(K_i - K_0)D_0.$$

Thus the denominator replacement costs $O(n^{-1/2})$ in $L^2$. For deterministic unit vectors $u, v \in \mathbb{R}^{2p}$, each numerator product $(u^\top \mathcal{V}_i)_j (\mathcal{V}_i^\top \mathcal{Q}_\delta^{(i)} v)_k$ has $L^2$ norm $O(n^{-1})$ by conditional linear-form fourth moments. Cauchy–Schwarz gives a replacement error $O(n^{-3/2})$ per pair, without any numerator–denominator independence. The positive Woodbury update also gives $\|\mathbb{E}\mathcal{Q}_\delta^{(i)} - \mathbb{E}\mathcal{Q}_\delta\|_{\mathrm{op}} = O(n^{-1})$. Summing the expected resolvent identity over pairs, uniformly over unit-vector tests, yields

$$\left\| \mathbb{E}\mathcal{Q}_\delta - \left[ T_\delta + \sum_\iota (C_\iota^{-1} + \zeta_n \bar{R})^{-1} \right]^{-1} \otimes I_p \right\|_{\mathrm{op}} = O(n^{-1/2}).$$

The limiting block-trace equation is therefore

$$R_\delta^{-1} = T_\delta + \sum_\iota (C_\iota^{-1} + \zeta R_\delta)^{-1}, \qquad R_0 = \operatorname{diag}(m(z), m(w)). \tag{54}$$

At zero coupling, the linearization on entry $(u, v)$, now indexed by $u, v \in \{z, w\}$, has multiplier

$$1 - \zeta m(u)m(v) \sum_\iota c_\iota(u)c_\iota(v)r_\iota(u)r_\iota(v) = m(v)\left[t_v + \sum_\iota c_\iota(v)r_\iota(u)r_\iota(v)\right] > 0.$$

The implicit function theorem gives a unique stable local branch. The resolvent identity bounds the sample block traces' change from zero coupling by $C|\delta|$; their vanishing equation residual forces convergence to this branch. Pair-replacement concentration gives the corresponding bilinear equivalent. Differentiating the off-diagonal block at zero gives

$$-\partial_\delta[\mathcal{Q}_\delta]_{12}|_0 = Q_z Q_w, \qquad \psi = m(z)m(w)\left[1 + \zeta\psi \sum_\iota c_\iota(z)c_\iota(w)r_\iota(z)r_\iota(w)\right],$$

which proves the unweighted limits in Equation (53).

For the weighted trace, let $V_i = (e_{-,i}, e_{+,i})$, $K_{iu} = V_i^\top Q_u^{(i)} V_i$, and $C_u = \mathrm{diag}(c_-(u), c_+(u))$. The pair-update identity gives

$$V_i^\top Q_z Q_w V_i = (I_2 + K_{iz}C_z)^{-1} V_i^\top Q_z^{(i)} Q_w^{(i)} V_i (I_2 + C_w K_{iw})^{-1}.$$

The middle block differs in $L^2$ by $O(n^{-1/2})$ from $n^{-1}\,\mathrm{tr}(Q_z^{(i)}Q_w^{(i)})I_2$; deletion changes this normalized trace by $O(n^{-1})$. The denominator limits are $\mathrm{diag}(r_-(u), r_+(u))$. Averaging diagonal entries proves the weighted limit.

Finally, positive real parts keep the coupled inverse and the weighted kernels locally bounded for small complex $\delta, z, w, \tau$; its stable branch is analytic. The analytic argument in Lemma C.1 justifies the differentiation above and the claimed uniform derivative limits. □

*Proof of Theorem D.2.* Let $H_-, H_+ \in \mathbb{R}^{n\times r}$ have rows

$$(H_-)_{i,:} = n^{-1/2}(\sqrt{2}\, w_i^\top \Gamma^{1/2}, (h_i + h_i^+)^\top \Delta^{1/2}/\sqrt{2}),$$
$$(H_+)_{i,:} = n^{-1/2}(0, (h_i - h_i^+)^\top \Delta^{1/2}/\sqrt{2}).$$

Then $Z/\sqrt{n} = E_- + H_- F^\top$, $D/\sqrt{n} = E_+ + H_+ F^\top$, and $Y_0 = Y/\sqrt{n}$. The entire latent-response block is independent of the original feature noises. Its second moments give

$$H_\iota^\top H_\iota \to S_\iota, \quad H_-^\top H_+ \to 0, \quad H_-^\top Y_0 \to B_0, \quad H_+^\top Y_0 \to 0, \quad Y_0^\top Y_0 \to \Omega.$$

Set $T_z = \sum_\iota c_\iota E_\iota^\top H_\iota$, $g_\epsilon = E_-^\top Y_0$, $B_n = H_-^\top Y_0$, and $R_{n,z} = \sum_\iota c_\iota H_\iota^\top H_\iota$. The exact lift is

$$M_n(z) = Q_z^{-1} + F R_{n,z} F^\top + F T_z^\top + T_z F^\top, \qquad g_n = F B_n + g_\epsilon. \tag{55}$$

We next evaluate the noise-dependent terms of this lift. Each is a bilinear form in the finite latent block $(H_-, H_+, Y_0)$ whose kernel depends only on the feature noise. Conditional on the noise, Equation (35) replaces such a form by the latent second moment times the normalized kernel trace, and Equation (34) gives the trace limits. Three cases arise.

First, forms that are linear in the latent block vanish. Since $F$ is deterministic and the latent block is centered and independent of the noise,

$$F^\top Q_z T_z \to 0, \qquad F^\top Q_z g_\epsilon \to 0.$$

Second, the response term uses the kernel $E_- Q_z E_-^\top$, whose normalized trace converges to $\zeta m r_-$. Together with $Y_0^\top Y_0 \to \Omega$, this gives

$$g_\epsilon^\top Q_z g_\epsilon = Y_0^\top (E_- Q_z E_-^\top) Y_0 \to \zeta m r_- \Omega.$$

Third, substituting $T_z = \sum_\iota c_\iota E_\iota^\top H_\iota$ yields

$$T_z^\top Q_z T_z = \sum_{\iota,\iota'} c_\iota c_{\iota'}\, H_\iota^\top (E_\iota Q_z E_{\iota'}^\top) H_{\iota'}, \qquad T_z^\top Q_z g_\epsilon = \sum_\iota c_\iota\, H_\iota^\top (E_\iota Q_z E_-^\top) Y_0.$$

In the same-sign terms, the kernel traces converge to $\zeta m r_\iota$, and the latent moments converge to $S_\iota$ and $B_0$. Hence

$$T_z^\top Q_z T_z \to \Lambda_z := \zeta m \sum_\iota c_\iota^2 r_\iota S_\iota, \qquad T_z^\top Q_z g_\epsilon \to \zeta m c_- r_- B_0.$$

The opposite-sign terms vanish because their latent second moments, $H_-^{\mathrm{T}} H_+$ and $H_+^{\mathrm{T}} Y_0$, converge to zero; this holds even though the sum and difference noises are dependent. Finally, the kernel bounds $\|E_\iota Q_z E_{\iota'}^{\mathrm{T}}\|_{\mathrm{op}} \le (c_\iota c_{\iota'})^{-1/2}$ in Lemma C.1 ensure that finite fourth moments suffice for these conditional calculations.

Writing $l_n = F^{\mathrm{T}} M_n^{-1} g_n$ and $j_n = T_z^{\mathrm{T}} M_n^{-1} g_n$, the exact lift gives

$$M_n^{-1} g_n = Q_z\{F(B_n - R_{n,z} l_n - j_n) + g_\epsilon - T_z l_n\}.$$

The limiting equations for its two projections are

$$l = m\{B_0 - (c_- S_- + c_+ S_+) l - j\}, \qquad j = \zeta m c_- r_- B_0 - \Lambda_z l.$$

Their Schur complement $I_r + m\sum_\iota c_\iota r_\iota S_\iota \succ 0$ keeps the small inverse bounded. Hence $l_n \to A_z$, $\Phi_n(z) \to \Phi(z)$, and $B_n - R_{n,z} l_n - j_n \to m^{-1} A_z$. Substitution in the exact lift gives, in Frobenius norm,

$$M_n(z)^{-1} g_n = Q_z\left[F m(z)^{-1} A_z + E_-^{\mathrm{T}}\{Y_0 - c_-(z) H_- A_z\} - c_+(z) E_+^{\mathrm{T}} H_+ A_z\right] + o_{\Pr}(1). \quad (56)$$

The norm error follows from $\|Q_z F\| \le t_z^{-1}$, $\|Q_z E_\iota^{\mathrm{T}}\| \le (t_z c_\iota)^{-1/2}$, and $\|H_\iota\|, \|Y_0\| = O_{\Pr}(1)$. Expand the product of Equation (56) at $z, w$. Conditional quadratic forms replace the same-sign latent products by $C_\iota(z, w)$; the opposite-sign covariances are zero. Deterministic-factor/latent cross terms vanish by conditional second moments. Lemma D.3 therefore gives

$$g_n^{\mathrm{T}} M_n(z)^{-1} M_n(w)^{-1} g_n \to \Xi(z, w),$$

with every cross-direction entry retained. All these limits are locally uniform with derivatives: the full pencil has positive real part locally, $\|\Phi_n\| \le C\|Y_0\|^2$, and its alignment and weighted kernels have the same local bounds as in the scalar proof. Apply the analytic argument of Lemma C.1.

By Lemma B.3, the crossings satisfy $d_j \ge 1/(z_j + \lambda) > 0$, and $\Phi_n(z) \preceq \Omega_n/(z - \lambda)$ excludes escaping roots. The assumed simple roots and gaps give convergence of the selected roots and response vectors after sign choices. Substitution in Equation (26) gives $H, E, J$; projection implies $N = E - H^{\mathrm{T}} H \succeq 0$. The bound Equation (27) keeps the refit inverse bounded and gives $I_q + (\rho - \tau) E \succ 0$. Finally substitute these limits into Equation (44) to obtain the risk limit. □

# E Additional simulations

## E.1 Complementary predictive signals: model and confirmation

**Generating model.** The second experiment in Section 5.1 uses four orthonormal loading vectors $u_{1\ell}, u_{1h}, u_{2\ell}, u_{2h} \in \mathbb{R}^{256}$, obtained as the first four columns of the QR factor of an iid $N(0, 1)$ matrix. For each unit $i$, let $w_{ij\ell}, w_{ijh}, \xi_{ij}$, $j = 1, 2$, be independent $N(0, 1)$ variables. Set

$$z_i = \sum_{j=1}^{2} \left(\sqrt{a_\ell}\, u_{j\ell} w_{ij\ell} + \sqrt{a_h}\, u_{jh} w_{ijh}\right), \qquad x_i = z_i + e_i, \qquad x_i^+ = z_i + e_i^+, \quad (57)$$

where the view noises $e_i, e_i^+ \overset{\text{iid}}{\sim} N(0, D)$ are independent of all factors and response noises, with $D = 2I_p - K$ and $K = \sum_{j=1}^{2}(a_\ell u_{j\ell} u_{j\ell}^{\mathrm{T}} + a_h u_{jh} u_{jh}^{\mathrm{T}})$. Response $j$ uses only the two factors of its own block,

$$y_{ij} = \nu\left(\sqrt{a_\ell}\, w_{ij\ell} + \sqrt{a_h}\, w_{ijh}\right) + \sqrt{0.2}\, \xi_{ij}, \quad (58)$$

with $\nu = \{0.8/(a_\ell + a_h)\}^{1/2}$, so these factors explain 80% of its unit variance. The loading frame is regenerated in each replication. For all three specified strength pairs, $D \succ 0$, $\Sigma_{\mathrm{tot}} = 2I_p$, $\Sigma_{\mathrm{ct}} = K$, and $\mathrm{Cov}(y_i) = I_2$. Since $\mathbb{E}(z_i w_{ij\ell}) = \sqrt{a_\ell}\, u_{j\ell}$ and $\mathbb{E}(z_i w_{ijh}) = \sqrt{a_h}\, u_{jh}$, the feature–response cross-covariance matrix $B_{\mathrm{reg}} = \mathbb{E}(x_i y_i^{\mathrm{T}})$ has columns

$$b_j = \nu\left(a_\ell u_{j\ell} + a_h u_{jh}\right), \qquad j = 1, 2.$$

The two response blocks share no factors. There are no additional shared distractors. The upper-strength pair correlation is 0.95 or 0.975; thus this construction studies strongly reproducible signal with a weaker predictive complement. It extends the paired-factor construction to non-isotropic view noise and does not test the assumptions of the scalar theorem. This is a multivariate, non-isotropic-noise prediction experiment, separate from the scalar risk calibration in Section 5.1. The optimal linear predictor uses two response-specific feature directions, so $q = 2$ can represent it without discarding predictive information despite the four latent factors.

**Fitting and tuning.** Responses retain their original scales. A common pooled feature mean and the two fitting-response means are estimated from the 384 fitting pairs; the same fitted centering is applied to validation and test data. All methods use both views of each labeled fitting and validation unit. The 1,536 validation pairs are independent of fitting data; they stabilize model selection and are counted in the labeled-data budget. Validation relative MSE divides each response's squared error, averaged over both single-view predictions, by $n_v^{-1}\sum_{i=1}^{n_v}(y_{ij}^{\mathrm{val}}-\bar{y}_j)^2$, where $n_v = 1536$ and $\bar{y}_j$ is its fitting mean. Joint selection averages the two response ratios; response-wise selection minimizes its own ratio. Let $s_X = \mathrm{tr}(\widehat{\Sigma}_{\mathrm{tot}})/r_X$, where $r_X = 256$ is the pooled fitting-design rank, and $s_Y = \|Y_c\|_F^2/(2n)$, where $Y_c$ is the centered $n \times 2$ response matrix. Each finite ridge parameter is selected from

$$\mathcal{G}_X = \{0\} \cup \{s_X 2^e : e = -8, -7, \ldots, 8\}.$$

Direction searches also include the Euclidean-metric endpoint; coefficient refits include the fitting-mean predictor. To cover both pair-weight endpoints, selection uses the leading two generalized eigenvectors of

$$\left((1-\omega)\widehat{B}_{\mathrm{reg}}\widehat{B}_{\mathrm{reg}}^{\mathrm{T}}/s_Y + \omega\widehat{\Sigma}_{\mathrm{ct}}, \widehat{\Sigma}_{\mathrm{tot}} + \tau I_p\right),$$

$$\omega \in \{0, .02, .05, .1, .2, .35, .5, .65, .8, .9, .95, .98, 1\}.$$

For $\omega < 1$, this selects the same space as Equation (19) with $\lambda = s_Y\omega/(1-\omega)$; $\omega = 0$ is no-pair and $\omega = 1$ is pair-only. Zero selection ridge is evaluated on the pooled row space. Directions are not discarded merely because their pair eigenvalues are negative. Refitting uses the feature-coefficient penalty in Equation (20), so changing a basis does not change the penalty.

For each initial selection ridge $\tau_0$, wider PCA++-PCR retains the leading $q_0$ pair directions for every integer $q_0 = 2, \ldots, 256$. With an orthonormal basis $W$ for that space, compression ridge $\kappa$ gives the two response-bearing directions

$$\widetilde{U} = W(W^{\mathrm{T}}\widehat{\Sigma}_{\mathrm{tot}}W + \kappa I_{q_0})^{-1}W^{\mathrm{T}}\widehat{B}_{\mathrm{reg}}.$$

The Euclidean compression endpoint is $\widetilde{U} = WW^{\mathrm{T}}\widehat{B}_{\mathrm{reg}}$. A coefficient ridge $\rho$ refits both responses in $\mathrm{span}(\widetilde{U})$. Width, $\tau_0$, $\kappa$, and $\rho$ jointly minimize validation relative MSE. Width two reduces to the rank-two pair-only refit; at full width, the family equals no-pair with its selection and coefficient ridges, an identity checked in every replication. The response-wise wider control instead computes each column's scalar direction and refit separately, selecting its own $(q_0, \tau_0, \kappa, \rho)$, with $q_0 = 1$ also permitted. Response-wise no-pair independently tunes its selection and coefficient ridges. These scalar fits are not jointly refitted in the union of their directions. For each wider or no-pair family, a validation-union control selects between the complete jointly tuned fit and the complete response-wise fit using the same mean validation relative MSE.

**Independent confirmation and uncertainty.** Exploratory comparisons motivated the three settings; their results are not included in the confirmation estimates or intervals. Before confirmation, the settings, grids, sample sizes, seven controls, and 200 replications per setting were fixed. All confirmation seeds are disjoint from the exploratory runs. Every fitted coefficient, intercept, and selection decision is frozen before the independent 5,000-pair test sample is generated. The test loss averages two single-view predictions, not a prediction using the mean of the two test features. If $R_{b,r}$ and $R_{c,r}$ denote baseline and C-PCR test relative MSE in replication $r$, the reported improvement is

$$100\left(1 - \frac{\overline{R}_c}{\overline{R}_b}\right), \qquad \overline{R}_c = \frac{1}{200}\sum_{r=1}^{200} R_{c,r}, \quad \overline{R}_b = \frac{1}{200}\sum_{r=1}^{200} R_{b,r}.$$

The prespecified 5% threshold is tested with the paired differences $0.95R_{b,r} - R_{c,r}$. One-sided Student $t$ lower bounds use level $1 - 0.05/21$ for each of the three settings and seven controls. All 21 lower bounds exceed zero. As a sensitivity check, 100,000 paired bootstrap resamples of the 200 replication rows give the percentile lower bounds in Table 3 at tail probability $0.05/21$. Each resample retains all methods from the same replication together. These are approximate Monte Carlo intervals; neither the two views nor the test observations are counted as independent replications. No replication is removed and the sample size is not increased after examining confirmation results.

**Exact risk decomposition.** For a fitted coefficient matrix $\widehat{B} = (\widehat{\beta}_1, \widehat{\beta}_2)$ and intercepts $\widehat{c}_j$, let $\beta_j^\star = b_j/2$ be the optimal linear coefficient. Because $\Sigma_{\mathrm{tot}} = 2I_p$, the conditional risk for response $j$ satisfies

$$\mathcal{R}_j(\widehat{\beta}_j, \widehat{c}_j) = 1 - \tfrac{1}{2}\|b_j\|^2 + 2\|\widehat{\beta}_j - \beta_j^\star\|^2 + \widehat{c}_j^2.$$

**Table 3:** Independent confirmation of complementary prediction signals. The first row reports C-PCR mean test relative MSE. Every other entry is its percentage reduction relative to the indicated control, with the simultaneous one-sided 95% paired-bootstrap lower bound in brackets. Darker and lighter gray mark the best and second-best methods by mean test relative MSE within each column; tied displayed reductions share a shade. There are 200 independent replications per column; all 21 comparisons enter the multiplicity adjustment. "Separate" denotes response-wise tuning, and "union" validation selection between joint and separate fits.

| Control | $(0.8, 1.90)$ | $(0.8, 1.95)$ | $(1.0, 1.95)$ |
|---|---|---|---|
| C-PCR relative MSE | 0.4587 | 0.4297 | 0.4341 |
| Wider PCA++-PCR | 6.38 [6.18] | 6.98 [6.76] | 6.74 [6.48] |
| Wider, separate | 5.69 [5.48] | 6.23 [6.02] | 6.06 [5.80] |
| Wider, union | 5.69 [5.48] | 6.23 [6.02] | 6.06 [5.80] |
| No-pair | 9.02 [8.74] | 11.01 [10.70] | 9.12 [8.86] |
| No-pair, separate | 8.91 [8.62] | 10.89 [10.58] | 9.00 [8.74] |
| No-pair, union | 8.91 [8.62] | 10.89 [10.58] | 9.01 [8.74] |
| Pair-only PCA++-PCR | 11.70 [11.33] | 9.41 [9.11] | 15.02 [14.68] |

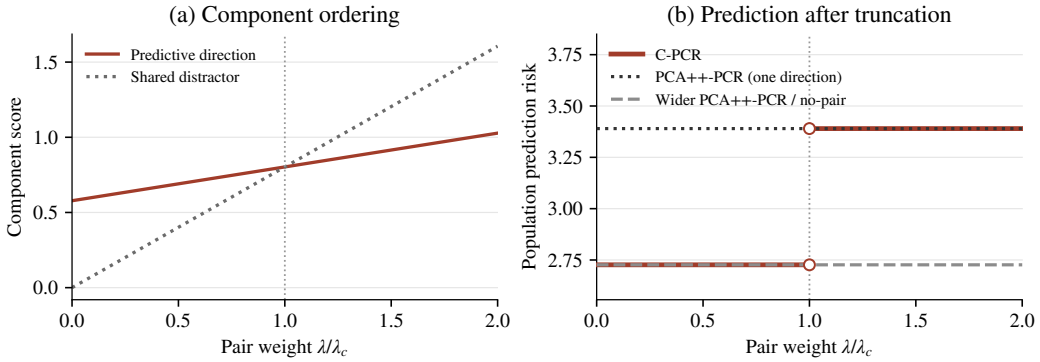


**Figure 4:** Population selection with a weak predictive factor. (a) Joint component scores cross at $\lambda = \lambda_c$. (b) Below the crossing, rank-one C-PCR retains the predictor that rank-one PCA++-PCR discards. The wider PCA++-PCR space contains both factors and has the same risk as no-pair. Open endpoints indicate nonuniqueness at the crossing.

The first term is its optimal linear-prediction risk. The squared coefficient error splits orthogonally across $P_\ell = \sum_j u_{j\ell} u_{j\ell}^\top$, $P_h = \sum_j u_{jh} u_{jh}^\top$, and $P_0 = I_p - P_\ell - P_h$. The plotted contributions are the corresponding $2\|P_t(\widehat{\beta}_j - \beta_j^\star)\|^2$, and the intercept term, divided by $1 + \bar{y}_j^2$ and averaged over the two responses and replications. Here $\bar{y}_j$ is the fitting-response mean; $1 + \bar{y}_j^2$ is the population loss of that mean predictor. This conditional-risk diagnostic therefore has no test-sampling noise and is distinct from the test relative MSE in panel (a). It is not used to tune any method. The decomposition describes estimation error in the fitted coefficients, not a causal attribution to separate algorithmic stages.

### E.2 Population illustration of early truncation

The population example has a predictive shared factor of strength $a_1 = 0.3$, a shared distractor of strength $a_2 = 3$, response coefficient $\vartheta = 1.7$, and response-noise variance $0.5$. We use selection ridge $\tau = 0.2$, coefficient ridge $\rho = 0.1$, and one retained direction. The predictive factor is weaker in the pair-only ordering but carries all response association. Below $\lambda_c = 1.1239$, the response term makes C-PCR retain it, reducing population risk by $0.6635$ relative to rank-one PCA++-PCR (Figure 4). Retaining both shared directions before regression removes this population truncation gap. At $\lambda = 0$ (no-pair endpoint), selection uses response association alone: the predictive factor has a positive population score, whereas the distractor has score zero. Thus, supervision already preserves the weak predictive direction before truncation.

### E.3 Fixed-parameter risk calibration

**One predictive factor.** The three experiments in Figure 2 use the one-factor model with independent $N(0,1)$ factor and feature-noise coordinates, $\xi_i \sim N(0,\sigma^2)$, no background factors, and known zero means. In every panel, $a = 3$, $\vartheta = 1.2$, $\sigma = 0.8$, $q = 1$, and $\tau = \rho = 0.7$. The fixed designs are

$$\begin{aligned}
&\text{(a)} \quad (n,p) = (400,160), \quad \lambda \in \{0, 0.01, 0.025, 0.05, 0.1, 0.2, 0.4\};\\
&\text{(b)} \quad n = 400, \quad \lambda = 0.4, \quad \gamma = \frac{p}{2n} \in \{0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1\};\\
&\text{(c)} \quad \gamma = 0.2, \quad \lambda = 0.4, \quad n \in \{200, 300, 400, 600, 800, 1200, 1600\}.
\end{aligned}$$

Each setting uses 256 independent training samples. The common setting $(n,p,\lambda) = (400,160,0.4)$ is reused across panels, and the same training sample is reused across the weights in (a). All settings satisfy the exterior-root condition in Equation (12). Conditional risk is calculated from exact population moments; no validation, parameter selection, or test-sample noise is involved.
In Figure 2(a)–(b), each interval is the mean risk plus or minus $t_{0.975,255}$ times its sample standard deviation divided by $\sqrt{256}$, where $t_{0.975,255}$ is the 97.5th percentile of the Student $t$ distribution with 255 degrees of freedom. The shaded band in panel (c) of that figure instead shows the empirical 5th and 95th percentiles of conditional risk across training samples, describing variability of the fitted predictor rather than uncertainty in its mean risk. The deterministic curves evaluate Equation (14); the limit is constant in (c) because $\gamma$ and the model and tuning parameters stay fixed.

**Three predictive channels.** We calibrate Theorem D.2 directly under the paired-factor model Equation (41), using the same $w_i$ to generate both views and all three responses. Set $k = m = L = 3$ and

$$\Gamma = \mathrm{diag}(5,3,1.5), \qquad \Delta = \mathrm{diag}(2,1,0.5), \qquad \Theta^* = \begin{pmatrix} 1.6 & .25 & .1 \\ .15 & 1.3 & .2 \\ .1 & .2 & 1 \end{pmatrix}, \qquad \Sigma_\xi = 0.25 I_3.$$

For each feature dimension, QR factorization of a fixed seeded Gaussian matrix supplies the six orthonormal columns of $(U^*, W^*)$, held fixed across innovation laws and replications. The coordinates of the primitive variables $w_i, h_i, h_i^+, \epsilon_i, \epsilon_i^+$, and $2\xi_i$ are sampled independently from one of three standardized laws: $N(0,1)$, Rademacher (equal probabilities on $-1, 1$), or an asymmetric law taking $\sqrt{2}$ and $-1/\sqrt{2}$ with probabilities $1/3$ and $2/3$. Thus all laws have finite fourth moments, while non-Gaussian sum and difference noises need not be independent.
Use $n = 200, 400, 800$, $p = 0.4n$, known zero means, and fixed $(\lambda,\tau,\rho) = (0.25, 0.7, 0.3)$, without response standardization or validation tuning. The computed roots are $2.4506, 1.2508, 0.5952$, separated above $\lambda = 0.25$ as simple crossings with $d_j > 0$. The limiting risks for ranks one, two, and three are $3.8103, 2.6865, 2.2357$, respectively; these targets stay fixed as $n$ changes because the aspect ratio and factor parameters are fixed. Each setting uses 64 independent training samples, with ranks one through three retaining nested prefixes of the same fitted eigenvectors. Conditional risk is computed exactly from the population moments. The relative mean-risk discrepancy decreases from $0.57\%$–$1.24\%$ at $n = 200$ to $0.17\%$–$0.37\%$ at $n = 800$ across the nine law/rank combinations (Figure 5); the finite-sample means remain slightly above the limiting risks.

### E.4 Wider PCA++-PCR spaces and supervised controls

This comparison asks how much expanding the initial pair-only space changes finite-sample prediction. It uses two predictive channels in three responses, $n = p = 128$, known zero means, final rank two, and 512 independent validation units. Generate $y_i \sim N(0,\Omega)$ independently of two residual innovation arrays with independent coordinates following one of the three standardized laws above, and set

$$Z_i = A_0 \Omega^{-1/2} y_i + \Sigma_-^{1/2} \xi_{-,i}, \quad D_i = \Sigma_+^{1/2} \xi_{+,i}, \quad x_i = (Z_i + D_i)/\sqrt{2}, \quad x_i^+ = (Z_i - D_i)/\sqrt{2}.$$

The covariance square roots are symmetric. Let

$$\Omega = \mathrm{diag}(o_1, o_2, 1/100), \qquad A_0 = (\sqrt{s_1} e_1, \sqrt{s_2} e_2, 0),$$

where $e_j$ is the $j$th coordinate vector. Set

$$D_\pm = \mathrm{diag}(d_{\pm,1}, d_{\pm,2}, 1, \ldots, 1), \qquad \Sigma_\pm = D_\pm^{1/2} (I + 0.2 Q_\pm \Lambda Q_\pm^{\mathrm{T}}) D_\pm^{1/2}.$$

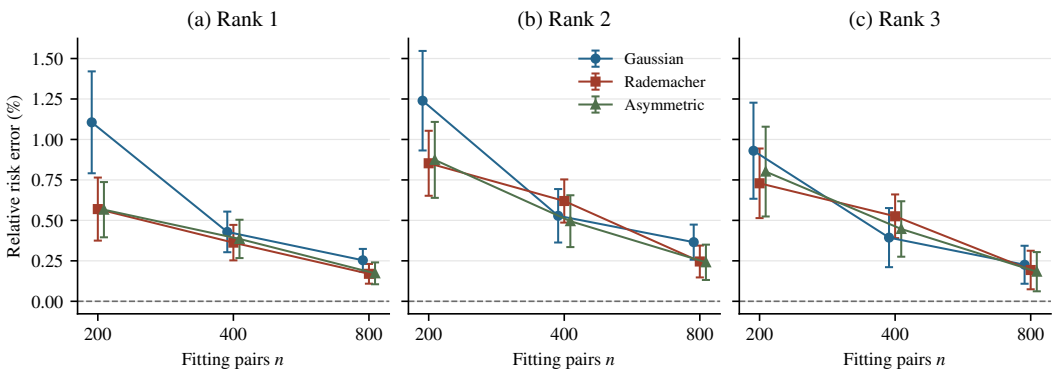


**Figure 5:** Calibration of the paired-factor risk formula at ranks (a) one, (b) two, and (c) three. Relative risk error is $100(\overline{\mathcal{R}} - \mathcal{R}_q)/\mathcal{R}_q$, where $\overline{\mathcal{R}}$ is the mean exact conditional risk over 64 independent training samples and $\mathcal{R}_q$ is its deterministic limit. Error bars are pointwise 95% Student $t$ intervals. All three innovation laws are sampled at the original factor and noise levels. The sample-size axis uses logarithmic spacing; small horizontal offsets distinguish the laws.

Here $\Lambda$ is diagonal, with $p$ equally spaced diagonal entries from $-1$ to 1, and $Q_-, Q_+$ are fixed independent orthogonal matrices obtained by QR factorization of iid $N(0,1)$ matrices. Channel parameters are

$$(d_{-,1}, d_{-,2}) = (2/25, 1/25), \quad (d_{+,1}, d_{+,2}) = (3/1000, 7/10),$$
$$(s_1, s_2) = (3/25, 48), \qquad (o_1, o_2) = (621/1948, 13044/7261).$$

This heterogeneous design was chosen in preliminary experiments, with evaluation on independent simulations. Validation averages the two single-view losses per unit. Test risk is evaluated exactly from population moments without test-sample noise.

For each selection ridge, the wider control retains $q_0 \in \{2, 4, 8, 16, 32, 64, 128\}$ leading pair directions, then selects two response-bearing directions by regularized reduced-rank regression (RRR) within that space. With an orthonormal basis $W$ for the initial space and finite direction ridge $\kappa$, this compression minimizes $\ell_\kappa(\Theta; W)$ subject to $\operatorname{rank}(\Theta) \le 2$. The penalty $\kappa\|W\Theta\|_F^2$ acts on feature coefficients and is independent of the initial basis. Coefficients in the resulting two-dimensional space are then refitted with an independently selected ridge. Widths exceeding the number of positive pair roots may retain negative roots, so the full-width control accesses the entire feature space; at $q_0 = 128$, its predictions agree with rank-two no-pair fitting. At $q_0 = 2$, it agrees with the rank-two PCA++-PCR refit.

All tuned methods use the rule-generated grids in Section F, with scales $s_X = \operatorname{tr}(\widehat{\Sigma}_{\text{tot}})/r_X$ and $s_Y = \|Y\|_F^2/(nL)$ computed from each fitting sample. The model responses retain their stated covariance; they are not individually standardized in this known-mean simulation. The no-pair, exact pair-only, zero coefficient ridge, Euclidean-metric, and zero-predictor endpoints are included. All component methods have final rank two and share the coefficient grid. The wider control uses the same selection ridges and the same direction-ridge grid for supervised compression. Candidate counts differ across families and are retained with the selected fits. Each method minimizes validation squared loss, and both the integer-exponent main grid and the prespecified half-integer sensitivity grid are run before conditional population risks are evaluated. There are 64 independent replications for each innovation law. The reported intervals are pointwise 95% Student $t$ intervals computed from paired risk differences across the 64 independent fitting/validation replications.

PCovR (de Jong & Kiers, 1992) selects score directions of

$$\varpi \frac{\widetilde{X}\widetilde{X}^{\mathsf{T}}}{\|\widetilde{X}\|_F^2} + (1-\varpi)\frac{P_{\widetilde{X}}\widetilde{Y}\widetilde{Y}^{\mathsf{T}}P_{\widetilde{X}}}{\|\widetilde{Y}\|_F^2} \quad \text{on } \operatorname{col}(\widetilde{X}),$$

where $\widetilde{X}$ stacks the views, $\widetilde{Y}$ repeats their responses, and $P_{\widetilde{X}}$ projects onto the column space of $\widetilde{X}$. Scores are mapped to feature loadings with the design pseudoinverse, followed by coefficient refitting. The main grid for the mixing weight $\varpi$ is $\{0, 0.1, \ldots, 1\}$, including prediction-only and reconstruction-only endpoints; the sensitivity grid uses increments of 0.05. Ridge-PCovR additionally

**Table 4:** Mean conditional population risks over 64 independent fitting/validation replications per innovation law, using the rule-generated main grid. All component methods have final rank two and the same coefficient-refitting candidates. Wider PCA++-PCR retains $q_0$ pair directions before supervised compression; the full-width row also reports its numerically identical no-pair fit.

| Method | $N(0,1)$ | Rademacher | Asymmetric |
|---|---|---|---|
| C-PCR | 0.2541 | 0.2517 | 0.2540 |
| PCovR + refit | 0.2663 | 0.2657 | 0.2663 |
| Ridge-PCovR + refit | 0.2663 | 0.2657 | 0.2663 |
| PCA++-PCR, $q_0 = 2$, then RRR | 0.3624 | 0.3627 | 0.3627 |
| PCA++-PCR, $q_0 = 4$, then RRR | 0.3644 | 0.3647 | 0.3646 |
| PCA++-PCR, $q_0 = 8$, then RRR | 0.3684 | 0.3684 | 0.3683 |
| PCA++-PCR, $q_0 = 16$, then RRR | 0.3625 | 0.3633 | 0.3649 |
| PCA++-PCR, $q_0 = 32$, then RRR | 0.3095 | 0.3068 | 0.3101 |
| PCA++-PCR, $q_0 = 64$, then RRR | 0.2820 | 0.2816 | 0.2810 |
| PCA++-PCR, $q_0 = 128$, then RRR / No-pair | 0.2755 | 0.2746 | 0.2751 |
| Ridge | 0.3281 | 0.3275 | 0.3257 |

**Table 5:** Dense-grid sensitivity for the heterogeneous simulation; main-grid risks appear in Table 4. N, R, and A denote $N(0,1)$, Rademacher, and asymmetric innovations. The dense grid inserts half-integer log-grid exponents. Each entry averages 64 conditional population risks, using the same training/validation replications as the main grid.

| Method | N | R | A |
|---|---|---|---|
| C-PCR | 0.2486 | 0.2458 | 0.2474 |
| PCovR + refit | 0.2657 | 0.2654 | 0.2659 |
| No-pair, rank two | 0.2746 | 0.2741 | 0.2746 |

uses zero and the same positive direction-ridge candidates for preliminary response regularization, with both observed-response and fitted-response normalizations under shrinkage.

Here conditional risk means the population squared prediction loss conditional on the entire frozen fit, including the fitting and validation data. Table 4 shows that sufficiently wide preselection ($q_0 \geq 32$) substantially narrows the risk gap between PCA++-PCR and C-PCR, while C-PCR retains lower mean risk under the rule-generated main grid. For $N(0,1)$ innovations, the paired no-pair-minus-C-PCR difference is $0.0213$ $[0.0188, 0.0238]$, and the PCovR difference is $0.0122$ $[0.0093, 0.0151]$. The full-width control agrees numerically with no-pair in every replication. The grid-density check in Table 5 preserves the ordering against no-pair and PCovR under all three innovation laws. This finite-sample comparison complements the two-factor population example, where retaining both factors removes the truncation gap.

# F Natural Paired Data

## F.1 Shared Evaluation Protocol

All methods receive the same two labeled fitting views and predict from one view. Preprocessing, feature centering, and response standardization use fitting data only; preprocessing, where varied, and parameters are selected by held-out validation loss. Coefficients remain fitted on the fitting partition. All views of a recorded evaluation unit stay in one partition. After a spectral transformation, both views are centered by the pooled fitting mean $\bar{x}$ and divided by

$$d_x = \left\{ \frac{1}{2n} \sum_{i=1}^{n} (\|x_i - \bar{x}\|_2^2 + \|x_i^+ - \bar{x}\|_2^2) \right\}^{1/2}.$$

Each response uses its fitting mean and standard deviation, with variance denominator $n$. Raman fixes standard normal variate normalization (SNV), which centers and scales each spectrum. The other datasets select among raw spectra, SNV, and an 11-channel, degree-three Savitzky–Golay first derivative. No artificial noise is added. Table 6 lists the evaluation sizes.

**Table 6:** Natural-data sizes per split; $p$ and $L$ are the numbers of spectral channels and responses. Raman uses eight fitting splits on a common test set; Corn and Mango use five overlapping sample-level splits. Tablets retains its supplied official rows.

| Dataset | Fitting | Validation | Test | $p$ | $L$ | Splits |
|---|---|---|---|---|---|---|
| Raman | 50 | 150 | 1088 | 704 | 3 | 8 |
| Beef | 72 | 72 | 96 | 125 | 3 | 1 |
| Tablets | 155 | 40 | 460 | 650 | 3 | 1 |
| Corn | 48 | 16 | 16 | 700 | 4 | 5 |
| Mango | 278 | 92 | 94 | 278 | 1 | 5 |

**Table 7:** Candidate evaluations per fitting split, including preprocessing choices and mean-predictor fallbacks. Raman uses fixed SNV; Beef, Tablets, Corn, and Mango each search three preprocessors. Counts are identical across each dataset's splits; Beef and Tablets share counts. Shared fits can be reused, so these are family-specific evaluations, not independent models or matched computational budgets. Beef's consistency RRR and consistency ridge evaluate 18,273/111,621 and 465/1,485 candidates on main/dense grids, respectively.

| Dataset | Grid | C-PCR | PCA++-PCR | No-pair | PLS | Ridge | PCR |
|---|---|---|---|---|---|---|---|
| Raman | Main | 39789 | 3921 | 589 | 21 | 15 | 21 |
| Raman | Dense | 245389 | 13521 | 2029 | 21 | 27 | 21 |
| Beef / Tablets | Main | 119367 | 11763 | 1767 | 63 | 45 | 63 |
| Beef / Tablets | Dense | 736167 | 40563 | 6087 | 63 | 81 | 63 |
| Corn | Main | 119955 | 11763 | 2355 | 63 | 45 | 63 |
| Corn | Dense | 738195 | 40563 | 8115 | 63 | 81 | 63 |
| Mango | Main | 118191 | 11763 | 591 | 63 | 45 | 63 |
| Mango | Dense | 732111 | 40563 | 2031 | 63 | 81 | 63 |

**A rule-generated tuning grid.** Let $r_X$ be the numerical rank of the pooled fitting design and define

$$s_X = \frac{\operatorname{tr}(\widehat{\Sigma}_{\text{tot}})}{r_X}, \qquad s_Y = \frac{\|Y\|_F^2}{nL}, \quad \mathcal{G} = \{10^e : e = -6, \dots, 6\}, \quad \mathcal{H} = \{10^e : e = -4, \dots, 4\}.$$

Here $s_X$ is the average variance over the observed row space, and $s_Y = 1$ for standardized natural-data responses. All five datasets and the simulation in Section E.4 use

$$\tau \in s_X\mathcal{G} \cup \{\infty\}, \qquad \rho \in \{0\} \cup s_X\mathcal{G}, \qquad \lambda \in \{0\} \cup s_Y\mathcal{H} \cup \{\infty\}.$$

The infinite values are exact Euclidean-metric and pair-only selection. Every search includes the fitting-response-mean predictor. C-PCR and PCA++-PCR search ranks 1–20, limited by the numerical design rank and the number of positive selected roots; no-pair retains at most $L$ response-bearing directions. PCR and PLS search ranks 1–20, using score ordinary least squares and native PLS coefficients, respectively, with PLS feature rescaling disabled. Ridge and coefficient-refitting methods share the $\rho$ grid.

**Consistency-control tuning.** For consistency RRR (Section B.5), direction ridge $\kappa$ uses the $\tau$ grid, disagreement weight $\mu$ uses $\{0\} \cup \mathcal{H} \cup \{\infty\}$, and coefficient ridge $\rho$ is selected independently. Consistency ridge uses the same $\mu, \rho$ grids without a separate refit. The $\kappa = \infty$ endpoint uses Euclidean selection; $\mu = \infty$ restricts fitting to the difference nullspace within the observed training row space when nonempty. All inverses and selected spaces use the row-space reduction.

**Selection and search budget.** The main grid uses integer exponents; the sensitivity grid inserts half-integers in $\mathcal{G}, \mathcal{H}$. Model selection minimizes validation squared error in fitting-standardized response coordinates, with strict-improvement tolerance $10^{-12}$. Both grids and their fits are frozen before evaluation, and the main grid is retained regardless of test performance. A denser grid cannot increase minimum validation loss but may increase test loss. Table 7 gives the current protocol's candidate evaluations. Shared grids do not match total search budgets: C-PCR includes both no-pair and pair-only families, so lower validation loss alone does not demonstrate a pairing benefit.

**Table 8:** Beef rMSE on 96 patties under the main grid. Differences are comparator minus C-PCR; 95% paired Student $t$ intervals condition on the frozen fits and normalization denominators. The final row is a rank-restricted diagnostic: $q \leq L$ selects $q = 3$ and gives the same frozen fit as fixed $q = 3$.

| Method | Relative MSE | Difference | 95% interval |
|---|---|---|---|
| C-PCR | 0.2268 | – | – |
| PCA++-PCR | 0.2291 | 0.0023 | [0.0002, 0.0044] |
| No-pair | 0.2410 | 0.0143 | [0.0004, 0.0282] |
| PLS | 0.2615 | 0.0347 | [0.0150, 0.0544] |
| Ridge | 0.2410 | 0.0143 | [0.0004, 0.0282] |
| PCR | 0.3154 | 0.0887 | [0.0460, 0.1313] |
| Consistency RRR + refit | 0.2393 | 0.0126 | [0.0010, 0.0241] |
| Consistency ridge | 0.2365 | 0.0097 | [-0.0020, 0.0215] |
| C-PCR, $q \leq L$ (selects $q = 3$) | 0.2258 | -0.0009 | [-0.0081, 0.0062] |

**Relative MSE and uncertainty.** With predictions in their original response units, define

$$\mathrm{rMSE} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{(NV)^{-1}\sum_{i=1}^{N}\sum_{v=1}^{V}\{\widehat{f}_\ell(x_i^{(v)}) - y_{i\ell}\}^2}{N^{-1}\sum_{i=1}^{N}(y_{i\ell} - \bar{y}_{\ell,\mathrm{fit}})^2}, \tag{59}$$

where $N$ is the number of test units, $V = 2$ separately predicted views except $V = 5$ for Beef, and $\bar{y}_{\ell,\mathrm{fit}}$ is the fitting-response mean. The mean predictor has rMSE one. Response-wise ratios are averaged across fitting splits; denominators enter only reporting, not selection. Thus reporting weights differ from the fitting-only response weights used for validation.
Differences are comparator rMSE minus C-PCR rMSE. Beef intervals use paired Student $t$ calculations across test units, conditional on the fits and observed normalization denominators; they are descriptive, pointwise, and unadjusted. Repeated scans are not independent test units. Raman losses are averaged over its eight frozen fits within each composition, and these fits are not independent; Corn and Mango splits overlap. Their results are therefore reported without independence-based confidence intervals.

### F.2 Raman carbon-source prediction

The public substrate-mixture data contain 6,960 spectra, of which 5,566 have complete eight-response compositions (Lange et al., 2026; Lange, 2025). Exactly 2,720 compositions have two spectra each; 21 compositions with six spectra are excluded by the exactly-two rule. Without physical specimen identifiers, composition defines the evaluation group. The grouped partitions preserve all eight 50/150 fitting/validation schedules and the 1,088 test compositions, leaving 1,632 development compositions; equal compositions stay together.
The three targets—glucose, glycerol, and acetate—and the 400–1800 $\mathrm{cm}^{-1}$ window excluding 1708–1718 $\mathrm{cm}^{-1}$ are fixed analysis choices, leaving 704 channels. With common SNV, C-PCR selects ranks 7–20, including three pair-only fits. The protocol sensitivities below retain every test composition. Performance concerns these composition groups, not transfer to new instruments or acquisition batches. Common SNV preprocessing removes the instability of the original scatter correction and yields a modest advantage over PLS and a larger gain over no-pair.

### F.3 Beef-patty prediction

The public near-infrared dataset has 240 patties with five scans each (León-Ecay et al., 2026). We use all 125 native micro-NIR channels from 908.1 to 1676.2 nm, without interpolation, and retain the 72/72/96 fitting/validation/test identifiers. Responses are the three ingredient concentrations. Training uses scans one and two; evaluation averages five single-scan losses.
The main grid selects raw spectra, $q = 19$, $\lambda = 10$, and $\tau = \rho = 8 \times 10^{-8}$. The $q \leq L$ diagnostic selects $q = 3$ and exactly reproduces the fixed-$q = 3$ fit (Table 8). The dense grid selects four components; its no-pair-minus-C-PCR difference is 0.0121 $[0.0030, 0.0213]$, while the smaller PCA++-PCR difference has an interval containing zero. All scans remain grouped by patty. Acquisition day and concentration are not balanced, so these results concern the recorded campaign; cross-day or new-batch performance is not established.

**Table 9:** Main-grid rMSE for the five Corn and Mango splits. Splits overlap and are not independent replicates. Darker and lighter gray mark the best and second-best distinct displayed values; rounding ties share a shade. Shading does not indicate significance.

| Dataset | Split | C-PCR | PCA++-PCR | No-pair | PLS | Ridge | PCR |
|---|---|---|---|---|---|---|---|
| Corn | 1 | 0.1442 | 0.1350 | 0.1268 | 0.1349 | 0.1268 | 0.1653 |
| Corn | 2 | 0.1140 | 0.1290 | 0.1129 | 0.1189 | 0.1131 | 0.1555 |
| Corn | 3 | 0.1068 | 0.1271 | 0.1068 | 0.1263 | 0.1068 | 0.1350 |
| Corn | 4 | 0.2459 | 0.2502 | 0.2461 | 0.2810 | 0.2461 | 0.3079 |
| Corn | 5 | 0.1727 | 0.1666 | 0.1648 | 0.2080 | 0.1648 | 0.1983 |
| Mango | 1 | 0.0735 | 0.1200 | 0.0811 | 0.1791 | 0.0811 | 0.4292 |
| Mango | 2 | 0.1253 | 0.1736 | 0.1188 | 0.1888 | 0.1183 | 0.4159 |
| Mango | 3 | 0.1014 | 0.1086 | 0.0977 | 0.1765 | 0.0977 | 0.4223 |
| Mango | 4 | 0.1136 | 0.1489 | 0.1295 | 0.2506 | 0.1295 | 0.4852 |
| Mango | 5 | 0.0986 | 0.1303 | 0.1019 | 0.2387 | 0.0962 | 0.6067 |

### F.4 Pharmaceutical tablets

The IDRC 2002 data provide two-instrument spectra and responses for weight, hardness, and assay (International Diffuse Reflectance Conference, 2002). We retain the official 155/40/460 fitting/validation/test rows and all 650 channels from 600 to 1898 nm. The arrays contain 655 rows although the source describes 654 tablets; no row is deleted to reconcile that discrepancy. Physical sample labels are absent. No exactly duplicated paired spectrum occurs across partitions, but this cannot establish physical identity independence.

The main grid selects raw spectra, $q = 9$, $\lambda = \infty$, and $\rho = 0$, giving the PCA++-PCR fit; the dense grid retains it. The response-wise C-PCR rMSEs are 1.0292, 0.7216, and 0.0919 for Weight, Hardness, and Assay, respectively. Weight remains the most difficult response; its rMSE exceeds that of the mean predictor.

### F.5 Corn

We use all 80 samples, instruments m5 and mp5, and four responses: moisture, oil, protein, and starch (Eigenvector Research, Inc., 2005). All 700 channels from 1100 to 2498 nm are retained. Five fixed random permutations allocate 48/16/16 samples to fitting/validation/test, keeping both views and all responses together. The individual results in Table 9 show the variability behind the small differences in mean performance.

### F.6 Mango dry matter

The public collection supplies dry-matter responses and repeated near-infrared measurements (Anderson et al., 2024). We retain the 464 fruit from the 2015 season observed on all three instruments 15006, 15014, and 15016, using the fixed pair 15006/15014. The cohort is therefore three-instrument complete. Technical repeats are averaged within fruit and instrument before splitting. All 278 channels from 309 to 1140 nm are retained. The same five seeds as Corn give 278/92/94 fitting/validation/test fruit. Neither dataset uses response-based cohort restriction or test-selected instrument pairs.

C-PCR improves on PCA++-PCR, PLS, and PCR in all five main-grid splits; gains over no-pair and ridge do not hold in every split. The dense grid retains the lowest average among the six methods, with a favorable PCA++-PCR comparison in four of five splits. This evidence concerns the selected season and recorded instrument pair, not transfer to unseen seasons or instruments.

### F.7 Grid and protocol sensitivity

Table 10 reports dense-grid performance and its change from the main results in Table 1. Table 11 retains the distinct Raman and Tablets alternatives. These analyses are retrospective: the Raman preprocessing repair and Tablets cross-validation diagnostics followed inspection of the original test results. Freezing the follow-up batches before evaluation does not make them independent confirmations. Every protocol evaluates all test units and all six methods; none combines methods from different protocols chosen by test performance.

**Raman preprocessing.** The original five-preprocessor search includes raw spectra, SNV, quadratic baseline removal, differentiation, and multiplicative scatter correction (MSC). Quadratic removal subtracts the fitted constant, linear, and quadratic wavenumber terms. MSC regresses each spectrum

**Table 10:** Dense-grid rMSE, with percentage change from the main grid in parentheses: $100(\mathrm{rMSE}_{\mathrm{dense}}/\mathrm{rMSE}_{\mathrm{main}} - 1)$, computed before rounding. Negative changes favor the dense grid. Raman uses fixed SNV, and all datasets retain held-out validation. PCR and PLS searches are unchanged. Shading ranks the dense risks within each dataset as in Table 1; it does not indicate significance.

| Dataset | C-PCR | PCA++-PCR | No-pair | PLS | Ridge | PCR |
|---|---|---|---|---|---|---|
| Raman | 0.2137 (-0.11%) | 0.2186 (-0.15%) | 0.2251 (-2.25%) | 0.2193 (0.00%) | 0.2227 (-3.37%) | 0.2186 (0.00%) |
| Beef | 0.2203 (-2.87%) | 0.2262 (-1.25%) | 0.2324 (-3.58%) | 0.2615 (0.00%) | 0.2410 (0.00%) | 0.3154 (0.00%) |
| Tablets | 0.6142 (0.00%) | 0.6142 (0.00%) | 0.6884 (+14.14%) | 0.6054 (0.00%) | 0.6275 (-6.74%) | 0.8224 (0.00%) |
| Corn | 0.1502 (-4.12%) | 0.1598 (-1.11%) | 0.1475 (-2.65%) | 0.1738 (0.00%) | 0.1491 (-1.60%) | 0.1924 (0.00%) |
| Mango | 0.1037 (+1.15%) | 0.1355 (-0.60%) | 0.1062 (+0.40%) | 0.2067 (0.00%) | 0.1053 (+0.75%) | 0.4719 (0.00%) |

**Table 11:** Retrospective protocol sensitivities; the main protocols appear in Table 1 and their dense-grid results in Table 10. Original Raman includes local-window MSC; non-MSC selection excludes only that preprocessor. Except where marked dense, grids use integer exponents. Tablets rows use the three cross-validation rules in this subsection. In the response-ridge row, † marks PLS and PCR with additional ridge refitting.

| Dataset | Protocol | C-PCR | PCA++-PCR | No-pair | PLS | Ridge | PCR |
|---|---|---|---|---|---|---|---|
| Raman | Original, main | 0.7005 | 0.7022 | 0.7408 | 0.2193 | 2.1220 | 0.4185 |
| Raman | Original, dense | 0.8773 | 0.6938 | 1.2809 | 0.2193 | 1.1421 | 0.4185 |
| Raman | Non-MSC selection | 0.2140 | 0.2190 | 0.2362 | 0.2193 | 0.2334 | 0.2186 |
| Raman | Non-MSC, dense | 0.2137 | 0.2186 | 0.2267 | 0.2193 | 0.2256 | 0.2186 |
| Raman | Full-spectrum MSC | 0.2470 | 0.2505 | 0.2591 | 0.2542 | 0.2578 | 0.2529 |
| Tablets | Five-fold CV | 0.6102 | 0.6102 | 0.6333 | 0.6896 | 0.6333 | 0.6856 |
| Tablets | One-SE rule | 0.6402 | 0.5496 | 0.6406 | 0.6866 | 0.6333 | 0.6839 |
| Tablets | Response ridges† | 0.6163 | 0.6163 | 0.6363 | 0.6459 | 0.6404 | 0.6833 |

on the fitting mean spectrum over the retained 1500–1800 $\mathrm{cm}^{-1}$ channels, then subtracts the offset and divides by the slope across the spectrum. A nearly zero slope on one test spectrum caused the original large prediction error. Common SNV avoids this division while retaining every composition. Excluding MSC but selecting among the other four preprocessors also selects SNV for every C-PCR fit. The full-spectrum MSC alternative uses all 704 retained channels in the correction regression. All reference spectra use fitting data only.

**Tablets selection rules.** Five-fold cross-validation partitions the 155 fitting rows into folds of 31 paired units using seed 2026092307; each fold estimates scales on its 124 fitting units. Candidates use normalized grid exponents and must be feasible in every fold. The integer grid, mean fold loss, and deterministic candidate identifiers for ties select SNV, $q = 5$, $\lambda = \infty$, and $\rho = 0$, refitted on all 155 rows. The 40 official validation rows serve only as checks for these diagnostics; test rows never select candidates.

The one-standard-error rule admits candidates within one estimated fold standard error of the minimum, then favors lower rank and larger coefficient ridge. This standard error is a selection heuristic, not an independent-replicate interval. The response-ridge alternative selects a separate coefficient ridge for each response while sharing its retained space; all six methods receive this flexibility, including ridge-refitted PLS and PCR. Neither alternative improves the minimum-CV C-PCR result.