CoCA: Cooperative Component Analysis

Daisy Yi Ding[2*], Alden Green[1*], Min Woo Sun[2], and Robert Tibshirani[1,2]

[1]Department of Statistics, Stanford University

[2]Department of Biomedical Data Science, Stanford University

*Equal contribution (alphabetical order)

July 25, 2024

Abstract

We propose *Cooperative Component Analysis* (CoCA), a new method for unsupervised multi-view analysis: it identifies the component that simultaneously captures significant within-view variance and exhibits strong cross-view correlation. The challenge of integrating multi-view data is particularly important in biology and medicine, where various types of "-omic" data, ranging from genomics to proteomics, are measured on the same set of samples. The goal is to uncover important, shared signals that represent underlying biological mechanisms. CoCA combines an approximation error loss to preserve information within data views and an "agreement penalty" to encourage alignment across data views. By balancing the trade-off between these two key components in the objective, CoCA has the property of interpolating between the commonly-used principal component analysis (PCA) and canonical correlation analysis (CCA) as special cases at the two ends of the solution path. CoCA chooses the degree of agreement in a data-adaptive manner, using a validation set or cross-validation to estimate test error. Furthermore, we propose a sparse variant of CoCA that incorporates the Lasso penalty to yield feature sparsity, facilitating the identification of key features driving the observed patterns. We demonstrate the effectiveness of CoCA on simulated data and two real multiomics studies of COVID-19 and ductal carcinoma in situ of breast. In both real data applications, CoCA successfully integrates multiomics data, extracting components that are not only consistently present across different data views but also more informative and predictive of disease progression. CoCA offers a powerful framework for discovering important shared signals in multi-view data, with the potential to uncover novel insights in an increasingly multi-view data world.

1 Introduction

With technological advances in biomedicine, it is common to collect multiple types of data, or "data views" on a common set of samples. The multiple data views enable a more holistic characterization of the subjects under investigation. For example, omics data, ranging from genomics and epigenomics to transcriptomics and proteomics, can now be routinely generated for a given set of biological specimens. These omics data capture molecular variations from different dimensions, providing a comprehensive view of complex biological systems and offering the potential to uncover new insights that may be hidden in a single data modality.

Given the high dimensionality and complexity of these datasets, principal component analysis (PCA) is frequently used to reduce dimensionality and identify the most significant patterns within the data [1, 2, 3, 4, 5, 6, 7, 8, 9]. However, in multi-view settings, the challenge extends beyond finding the principal components (PCs) that explain the most variance within data views. It can be equally important to ensure that the identified components exhibit a strong correlation across different data views. This alignment of components across data views suggests that the uncovered patterns are not merely artifacts of a single data view but are consistent signals more representative of the underlying biology.

On the other hand, canonical correlation analysis (CCA) is a commonly used method to identify patterns that are strongly correlated across different data views [10]. It finds linear combinations of variables from two data views to maximize their correlation. However, CCA has limitations, especially when dealing with high-dimensional data: it may be sensitive to noise and can pick up small noise directions [11, 12, 13, 14].

1

Additionally, CCA may overlook important signals as it does not take into account the variance explained by the identified components.

We propose a new method called *Cooperative Component Analysis*, or short for CoCA, to identify the component that captures significant within-view variance and cross-view correlation in multi-view data. To illustrate our method, we applied CoCA to integrate CT scan-derived radiomics features and clinical features measured on a cohort of 127 COVID-19 patients [15], which will be described in more detail in Section 4. Figure 1 gives a visual illustration of the scores derived from radiomics and clinical data views, obtained from PCA, CCA, and CoCA, respectively. Each dot represents a patient and is colored by the patient's ICU admission outcome. As compared to PCA and CCA, CoCA shows a clearer separation between patients who required ICU admission and those who did not, suggesting that it captures more informative patterns associated with COVID-19 disease progression. Moreover, CoCA also shows better alignment of the scores derived from the two data modalities, reflecting a shared underlying biological signal.
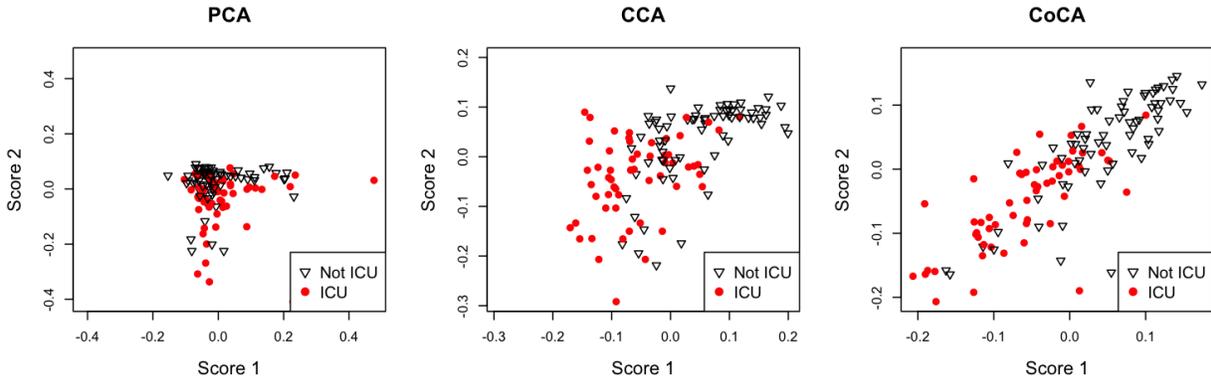

Figure 1: *Comparison of multi-view scores derived from radiomics and clinical data views for a cohort of COVID-19 patients, using PCA, CCA, and CoCA, respectively.* Each point represents a patient, colored by ICU admission outcome. CoCA achieves clearer separation between ICU and non-ICU patients and better alignment between the two data views.

The paper is organized as follows. In Section 2, we introduce the formulation of CoCA and characterize its solution. We show that it generalizes both PCA and CCA and further motivate the conceptual underpinnings of CoCA through an illustrative simulation study. In Section 3, we present sparse CoCA that incorporates a Lasso penalty to improve the interpretability of the identified component, along with an efficient optimization algorithm for solving sparse CoCA. We discuss its relation with other existing approaches and demonstrate its effectiveness in more simulation studies. In Section 4, we illustrate the practical utility of CoCA through its application to two real-world multiomic datasets: (1) integration of radiomics and laboratory measurements of COVID-19 patients, and (2) integration of epithelial and stromal gene expression data of breast ductal carcinoma in situ patients. In both applications, we show that CoCA uncovers components that capture significant signals within data views while exhibiting strong correlations across views. The paper ends with a discussion in Section 5 and an Appendix.

2 Cooperative component analysis (CoCA)

2.1 CoCA

We begin with a simple form of cooperative component analysis (CoCA) without sparsity. Let $\boldsymbol{X}_1 \in \mathbb{R}^{n \times p_1}, \boldsymbol{X}_2 \in \mathbb{R}^{n \times p_2}$ — representing two data views — and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the concatenation of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Fixing the hyperparameter $\rho \geq 0$, we propose to minimize the following objective:

$$\min_{u,v,d} \frac{1}{2}\|\boldsymbol{X} - duv^{\top}\|_F^2 + \frac{\rho}{2}\|d\boldsymbol{X}_1 v_1 - d\boldsymbol{X}_2 v_2\|_2^2, \quad \text{subject to } \|v\|_2^2 = 1, \|u\|_2^2 = 1, \tag{1}$$

where $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$ are vectors, and d is a scalar. v is partitioned as $v = (v_1, v_2)$, where $v_1 \in \mathbb{R}^{p_1}$

corresponds to $\boldsymbol{X}_1$ and $v_2 \in \mathbb{R}^{p_2}$ to $\boldsymbol{X}_2$. The rows of the data matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_1$ represent the common set of observations, and the columns correspond to features. The objective of CoCA is comprised of two key components:

1. **Approximation error**: $\|\boldsymbol{X} - duv^\top\|_F^2$ measures the Frobenius norm of the difference between the original data $\boldsymbol{X}$ and its low-rank approximation $duv^\top$. Specifically, given a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with singular value decomposition (SVD) $\boldsymbol{X} = \sum_{k=1}^r d_k u_k v_k^\top$, by the Eckart-Young Theorem [16], the leading singular value d_1, and left and right singular vectors u_1, v_1, are the solution to the optimization problem

$$\min_{u,v,d} \|\boldsymbol{X} - duv^\top\|_F^2, \quad \text{subject to } \|v\|_2^2 = 1, \|u\|_2^2 = 1. \tag{2}$$

The constraints $\|v\|_2^2 = 1$ and $\|u\|_2^2 = 1$ ensure the normalization of the singular vectors to guarantee a unique solution. Note that PCA can be performed via SVD of the data matrix: the principal components are the right singular vectors of the data matrix obtained from SVD.

2. **Agreement Penalty**: $\|d\boldsymbol{X}_1 v_1 - d\boldsymbol{X}_2 v_2\|_2^2$ introduces a penalty to encourage alignment across different data views. Intuitively, this term aligns two sets of scores for the samples, obtained by computing $\boldsymbol{X}_1 v_1$ and $\boldsymbol{X}_2 v_2$. These scores could represent, for example, disease severity measures, derived from radiomics and clinical data views as in the previous motivating example. We refer to these as multi-view scores. By minimizing the difference between these two sets of scores, the agreement penalty encourages similarity between the disease severity measures obtained from radiomics data and those from clinical data, thereby promoting agreement across data views. $\rho \geq 0$ controls the relative importance of the agreement penalty.

CoCA formulation reflects a balance between capturing important patterns within data views and encouraging alignment across data views to reveal common underlying signals. The hyperparameter ρ controls the trade-off between the approximation error and the agreement penalty. When $\rho = 0$, Problem 1 reduces to the standard SVD, which focuses solely on minimizing the approximation error. As ρ increases, more emphasis is placed on encouraging alignment across data views. The optimal value of ρ can be estimated using a validation set or through cross-validation, which we describe in more detail in Appendix Section A.

By varying the weight of the agreement penalty, CoCA has the property of encompassing PCA and CCA as special cases at the two extremes of the solution path:

- When $\rho = 0$, the solution $\hat{v}$ is proportional to the first principal component of the combined data view $\boldsymbol{X}$;

- As ρ approaches infinity, the appropriately scaled $\hat{v}$ converges to the leading canonical direction between the two views $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

We discuss this relationship in more detail in the next section.

2.2 Relation to PCA and CCA

We now show that CoCA generalizes both PCA and CCA: when $\rho = 0$, $\hat{v}$ is proportional to the first principal component of $\boldsymbol{X}$, and as $\rho \to \infty$, the appropriately scaled $\hat{v}$ converges to the leading canonical direction. We will find it convenient to transform $dv \mapsto v$ in Problem (1), in which case the problem can be reformulated as[*]

$$\min_{u,v} \frac{1}{2}\|\boldsymbol{X} - uv^\top\|_F^2 + \frac{\rho}{2}\|\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2\|_2^2 \quad \text{subject to } \|u\|_2^2 = 1. \tag{3}$$

Let $\hat{u}, \hat{v}$ be the solution to (3). Standard Lagrange calculus shows that $\hat{u}$ is the leading eigenvector of the matrix $\boldsymbol{X}(\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top$, while $\hat{v}$ is the leading eigenvector of the matrix $(\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top \boldsymbol{X}$, where $\boldsymbol{D} = \text{diag}(\boldsymbol{I}_{p_1}, -\boldsymbol{I}_{p_2})$. Let λ_1 be the leading eigenvalue of $(\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top \boldsymbol{X}$.

---

[*]Note that if $(\hat{u}, \hat{v})$ is the solution to (3), then $(\hat{u}, \hat{v}/\|\hat{v}\|_2, \|\hat{v}\|_2)$ is the solution to (1).

Theorem 1. *The solution $(\hat{u}, \hat{v})$ to* (3) *satisfies*

$$\boldsymbol{X}(\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top \hat{u} = \lambda_1 \hat{u}, \ \ \|\hat{u}\|_2^2 = 1. \tag{4}$$

and

$$(\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\hat{v} = \lambda_1 \hat{v}, \ \ \|\boldsymbol{X}\hat{v}\|_2 = \lambda_1. \tag{5}$$

Moreover, when $\rho = 0$, $\hat{v}$ is proportional to the first principal component of $\boldsymbol{X}$. As $\rho \to \infty$, if $\mathrm{rank}(\boldsymbol{X}) = p < n$ and $\boldsymbol{X}_1^\top \boldsymbol{X}_2 \neq 0$, then $(\hat{v}_1/\|\boldsymbol{X}_1\hat{v}_1\|_2, \hat{v}_2/\|\boldsymbol{X}_2\hat{v}_2\|_2)$ converges to the first canonical direction between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

The proof of Theorem 1 is given in Appendix B. An interesting consequence of Theorem 1 is that CoCA can be viewed from two perspectives:

1. The perspective we have been taking thus far is that CoCA is a kind of *penalized PCA*. The CoCA objective takes the traditional objective of SVD – minimizing rank-1 approximation error – and adds a penalty that encourages the solution to be highly correlated between views. If the first population PC is highly correlated between views, the agreement penalty exploits this structure so that the CoCA solution can be more accurate, and achieve lower reconstruction error on test data, than the usual first sample PC.

2. A new perspective, motivated by Theorem 1, is that CoCA is a kind of *penalized CCA*. Seen in this way, CoCA takes an objective – minimizing squared differences between the single-view scores $\boldsymbol{X}_1 v_1$ and $\boldsymbol{X}_2 v_2$ – that results in the first canonical directions, and adds a penalty that steers the solution towards higher variance-explained solutions. This penalty prevents overfitting spurious correlations in lower variance-explained directions. If $\boldsymbol{X}_1, \boldsymbol{X}_2$ have low effective rank, and the first population canonical directions are indeed in the higher variance-explained subspaces, then the CoCA solution can be more accurate than the usual first sample canonical directions.

These two perspectives on CoCA are visualized in Figure 2. The upshot is that CoCA can be viewed as an alternative to either PCA (for estimating the first population PC) or CCA (for estimating the first canonical direction), and we will see examples where CoCA is more accurate than either PCA or CCA shortly.

2.3   Illustrative simulation study

To show the potential benefits of CoCA, as compared to PCA and CCA, we conduct an illustrative simulation study with data drawn from a latent factor model. In this model, we suppose the data $\boldsymbol{X} = (x_1 \dots x_n)^\top$ consist of independent vectors $x_i \in \mathbb{R}^p$, with each $x_i = (x_{i,1}, x_{i,2})$ consisting of two views $x_{i,1} \in \mathbb{R}^{p_1}, x_{i,2} \in \mathbb{R}^{p_2}$ generated according to

$$
\begin{aligned}
x_1 &= \beta_1 z + \boldsymbol{W}_1 z_1 + \boldsymbol{B}_1 s + \epsilon_1, \quad z \sim N(0,1), z_1 \sim N_{k_1}(0, \boldsymbol{I}_{k_1}), s \sim N_l(0, \boldsymbol{I}_l), \epsilon_1 \sim N_{p_1}(0, \boldsymbol{\Omega}_1) \\
x_2 &= \beta_2 z + \boldsymbol{W}_2 z_2 + \boldsymbol{B}_2 s + \epsilon_2, \quad z_2 \sim N_{k_2}(0, \boldsymbol{I}_{k_2}), \epsilon_2 \sim N_{p_2}(0, \boldsymbol{\Omega}_2)
\end{aligned} \tag{6}
$$

with all random variables independent. Latent variable models similar to (6) have been used to study PCA [17, 18, 19, 20] and CCA [21, 22], and so (6) is a natural test bed for understanding CoCA. The covariance of a random vector $x \in \mathbb{R}^p$ distributed according to (6) is

$$\mathrm{Cov}(x) := \boldsymbol{\Sigma} = \beta\beta^\top + \begin{bmatrix} \boldsymbol{W}_1\boldsymbol{W}_1^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}_2\boldsymbol{W}_2^\top \end{bmatrix} + \boldsymbol{B}\boldsymbol{B}^\top + \begin{bmatrix} \boldsymbol{\Omega}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}_2 \end{bmatrix},$$

where $\beta = (\beta_1, \beta_2)$ and $\boldsymbol{B} = (\boldsymbol{B}_1, \boldsymbol{B}_2)$.

Consider (6), with $p_1 = p_2 = 4$, and parameters taken as follows:

$$\beta_1 = \beta_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \boldsymbol{W}_1 = \boldsymbol{W}_2 = (\|\beta\|_2 - 0.1) \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}, \ \boldsymbol{B}_1 = \boldsymbol{B}_2 = (\|\beta\|_2 - 1) \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \ \boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \mathrm{diag}(1, 1, 1, 0.09).$$

For these choices, we would expect an intermediate value of ρ to achieve the best performance:
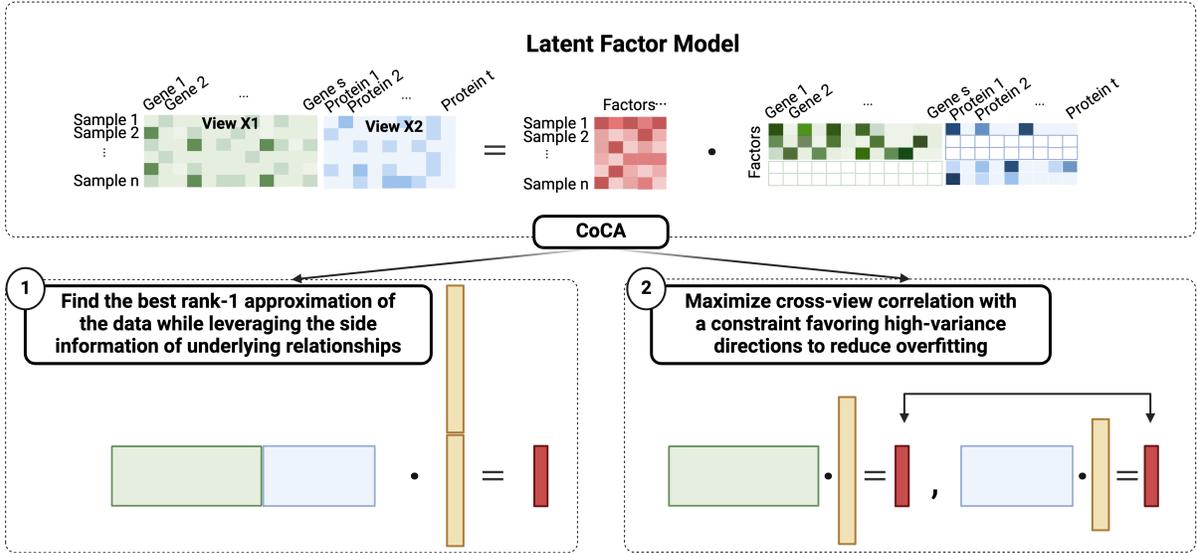
Figure 2: *Intuition behind CoCA.* Consider a scenario where an important biological pathway, such as the p53 pathway involved in cell cycle regulation and tumor suppression, contains the most signal and drives the correlation between two data views (e.g., transcriptomics and proteomics). This pathway can be considered a latent factor, with key genes and corresponding proteins having loadings on this factor in both views. While there could be other pathways that explain large variance within each view, they may not be consistently shared between the views. CoCA can be understood from two perspectives. In the subpanel (1) on the bottom left, CoCA finds the best rank-1 approximation of the data by leveraging side information, i.e., shared underlying relationships between views, to improve approximation error. The assumption is that the best rank-1 approximation is the information that is redundantly expressed across views, and the agreement penalty exploits this structure to improve the approximation. Alternatively, in the subpanel (2) on the bottom right, CoCA can be seen as identifying the direction that maximizes correlation across views while incorporating a constraint favoring high-variance directions to reduce overfitting.

- The first population PC and population canonical directions are both equal to $\beta = (\beta_1, \beta_2)$. Thus β corresponds to the direction that both explains the highest variance, and has the most cross-correlation between views. It is in fact also the solution to CoCA at the level of the population – i.e. the optimization problem (1) with $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}$ – for any value of ρ, as we demonstrate numerically in this simulation study.

- Both $(\boldsymbol{W}_1, 0)$ and $(0, \boldsymbol{W}_2)$ – i.e. $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, padded with zeros to make them vectors in $\mathbb{R}^p$ – explain a similar amount of variance to β, making them good candidates for PCA; but they are each supported in only one of the two views, and so have cross-correlation equal to zero.

- On the other hand, $\boldsymbol{B}$ has similar cross-correlation to β, making it a good candidate for CCA; but it explains much less of the overall variance.

Figure 3 visualizes the results when CoCA is computed on $n = 200$ samples drawn from this latent factor model. Two different criteria are used to evaluate performance of CoCA across different values of ρ. The left plot shows expected estimation error, which measures the difference between the estimated component $\hat{v}$ and the true component β:[†]

$$\mathbb{E}\Big[\min\{\|\hat{v} - \beta\|_2^2, \| - \hat{v} - \beta\|_2^2\}\Big]^{[‡]}.$$

The right plot shows the excess reconstruction error on a test set, where the test set $\boldsymbol{X}_{\text{test}}$ is generated with the same set of parameters with more number of data points $n_{\text{test}} = 5000$. The excess reconstruction error measures how well the estimated component $\hat{v}$ reconstructs the unseen test data as compared to the true component β, and is calculated as:

$$\mathbb{E}\Big[\frac{1}{n}\|\boldsymbol{X}_{\text{test}} - \boldsymbol{X}_{\text{test}}\hat{v}(\hat{v})^\top\|_F^2 - \frac{1}{n}\|\boldsymbol{X}_{\text{test}} - \boldsymbol{X}_{\text{test}}\beta(\beta)^\top\|_F^2\Big].$$

[†]We normalize $\beta \leftarrow \beta/\|\beta\|_2$ and $\hat{v} \leftarrow \hat{v}/\|\hat{v}\|_2$ before calculating the estimation and excess reconstruction error.

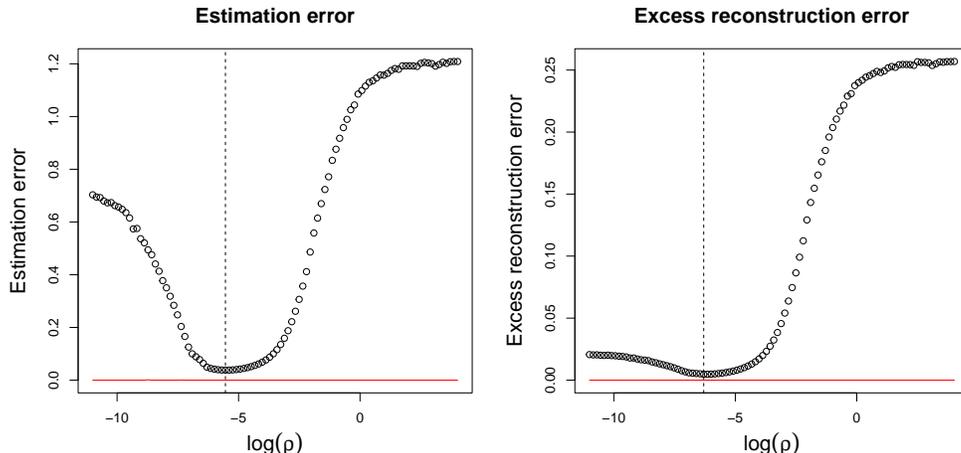[‡]The minimum is over $\hat{v}, -\hat{v}$ because the CoCA solution is defined only up to ± 1.

Figure 3: *An illustrative simulation study.* The plots show the estimation and excess reconstruction error of CoCA across different values of ρ under a latent factor model. The red line corresponds to the error of the population CoCA solution across different values of ρ. Each point is an average over 100 Monte Carlo runs.

Both the estimation error and excess reconstruction error plots reveal a clear U-shaped curve, with an optimal performance achieved at an intermediate value of ρ. The optimal point offers a marked improvement over the solution at either extreme of the solution path. The red line in the plots corresponds to the solution of population CoCA, which remains at β across different values of ρ. These results illustrate how CoCA effectively captures the shared signal and offers benefits over PCA and CCA in finite-sample scenarios.

In many applied problems with multiple data views, it can be reasonable to think of data as arising from a latent factor model in which there are (a few) shared and (many) individual latent factors. To give a concrete example in biomedicine, consider transcriptomics and proteomics data collected on a common set of cancer tissue samples (Figure 2). Suppose there is an important biological pathway, such as the p53 signaling pathway involved in cell cycle regulation and tumor suppression, that contains the most signal and makes the two views most correlated with each other. The pathway can be thought of as a latent factor: the key genes in the pathway (e.g. TP53, MDM2, CDKN1A) have expression levels in the transcriptomics data and corresponding protein abundances in the proteomics data that load strongly on this factor. Besides this shared factor, there could be other pathways that explain large variance but are not shared consistently between both gene and protein expression, therefore unlikely to be the consistent biological signals we aim to identify. Alternatively, there might be other shared pathways, such as the Wnt signaling pathway involved in embryonic development, that are less relevant to the cancer phenotype of interest. Our goal is to specifically identify the important shared patterns, which could be most informative about the underlying biology. The simulation study of this section indicates CoCA can be an effective method for this kind of problem.

3   Sparse cooperative component analysis (Sparse CoCA)

3.1   Sparse CoCA

To encourage sparsity in the solution, we incorporate an ℓ_1 penalty on v into the objective in Problem (3) and solve the following optimization problem:

$$\min_{u,v} \|\boldsymbol{X} - uv^\top\|_F^2 + \rho\|\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2\|_2^2 + \lambda\|v\|_1, \quad \text{subject to } \|u\|_2^2 = 1. \tag{7}$$

We build upon Problem (3) instead of Problem (1) because the non-convex constraint on v in the latter would make the optimization problem challenging to solve. In Problem (7), $\rho \geq 0$ is the hyperparameter that controls the relative importance of the agreement penalty $\|\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2\|_2^2$, and $\lambda \geq 0$ controls the sparsity of v through the ℓ_1 penalty. We refer to this as *sparse CoCA*.

We propose an alternating algorithm for optimizing the non-convex problem (7), inspired by the sparse PCA algorithm proposed by [23]. The algorithm iteratively optimizes over v and u, fixing one and optimizing over the other. Specifically, the updates are as follows. At step $k + 1$, with a fixed value of u^k, we solve v by solving the following minimization problem

$$\min_v \|\boldsymbol{X}^\top u^k - v\|_2^2 + \rho\|\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2\|_2^2 + \lambda\|v\|_1. \tag{8}$$

The solution can be computed as follows. Letting

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{I}_p \\ \sqrt{\rho}\boldsymbol{X}\boldsymbol{D} \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} \boldsymbol{X}^\top u^k \\ \boldsymbol{0}_p \end{bmatrix}, \quad \tilde{\beta} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

then Problem (8) can be rewritten as:

$$\min_v ||\tilde{y} - \tilde{\boldsymbol{X}}\tilde{\beta}||_2^2 + \lambda\|\tilde{\beta}\|_1. \tag{9}$$

This is a form of the Lasso, and can be computed, for example by the `glmnet` package [24].

With a fixed value of v^{k+1}, the u-update is simply

$$u^{k+1} = \frac{\boldsymbol{X}v^{k+1}}{\|\boldsymbol{X}v^{k+1}\|_2}.$$

Let $\mathrm{Lasso}(\boldsymbol{X}, \boldsymbol{y}, \lambda)$ denote the generic problem:

$$\min_\beta \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1. \tag{10}$$

We outline the alternating algorithm for sparse CoCA in Algorithm 1.

Algorithm 1 *Alternating algorithm for sparse CoCA.*

Input: Let $\boldsymbol{X}_1 \in \mathbb{R}^{n \times p_1}, \boldsymbol{X}_2 \in \mathbb{R}^{n \times p_2}$ — representing two data views — and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the concatenation of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. For fixed hyperparameters $\rho \geq 0$ and $\lambda \geq 0$.

for $k \leftarrow 0, 1, 2, \ldots$ until convergence **do**

At step $k + 1$,

1. *v-update:* At a fixed value of u^k, let

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{I}_p \\ \sqrt{\rho}\boldsymbol{X}\boldsymbol{D} \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} \boldsymbol{X}^\top u^k \\ \boldsymbol{0}_p \end{bmatrix}$$

Update v^{k+1} by solving $\mathrm{Lasso}(\tilde{\boldsymbol{X}}, \tilde{y}, \lambda)$.

2. *u-update:*

$$u^{k+1} = \frac{\boldsymbol{X}v^{k+1}}{\|\boldsymbol{X}v^{k+1}\|_2}.$$

end

The optimal value of ρ and λ can be determined using a validation set or cross-validation (CV) to estimate the test reconstruction error or other metrics based on the specific applications. We describe the detailed CV procedures for CoCA in Appendix Section B.

Remark A. Without sparsity, each v-update satisfies:

$$v^{k+1} = \frac{[\boldsymbol{I} + \rho\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}]^{-1}\boldsymbol{X}^\top\boldsymbol{X}v^k}{\|\boldsymbol{X}v^k\|_2}.$$

These are the iterates of the power method, up to normalization, and it can be seen that the true solution $\hat{v}$ as defined in (5) is a fixed point of the algorithm. Hence, convergence to the optimum without sparsity is

guaranteed. With sparsity, there is no longer a guarantee of convergence, but the algorithm behaves well in practice based on our experimental observations.

Remark B. We have explored alternative formulations of CoCA. They correspond to different approaches to sparse PCA [25]. Without sparsity, they are equivalent to our final formulation, which corresponds to the SVD-based approach to sparse PCA. The key requirements for finalizing our formulation are two-fold:

- First, it encompasses PCA and CCA as special cases at the two ends of the solution path.
- Second, it is computationally tractable when sparsity is introduced, allowing the use of efficient algorithms like `glmnet` in an alternating optimization scheme.

Our final formulation is the only one that satisfies these key requirements. We provide further details on the alternative formulations in Appendix Section C.

3.2 Relation to other existing approaches

We have mentioned in Section 2.2 the close connection of CoCA to PCA and CCA: setting $\rho = 0$ or $\rho \to \infty$ in CoCA gives PCA and CCA, respectively, as special cases at the two ends of the solution path.

For high-dimensional settings, researchers have proposed various sparse variants of PCA [23, 26, 27, 28, 11, 29, 30, 31] and CCA [32, 13, 33, 34, 35, 36], among others. For example, Jolliffe et al. [26] built upon the maximal variance characterization of PCA and proposed a method for sparse PCA called SCoTLASS to obtain sparse loadings. The first sparse principal component solves:

$$\max_v v^\top (\boldsymbol{X}^\top \boldsymbol{X}) v, \quad \text{subject to } \|v\|_2^2 \le 1, \|v\|_1 \le c, \tag{11}$$

where c is a tuning parameter controlling the level of sparsity. Subsequent components are obtained by solving the same problem with the additional constraint that they must be orthogonal to the previous components. However, this direct formulation leads to a non-convex optimization problem that is computationally costly to solve.

After SCoTLASS, Zou et al. [23] introduced a more computationally efficient SPCA algorithm for high-dimensional data. They showed that PCA can be formulated as a regression-type optimization problem, where sparse loadings are obtained by imposing the Lasso constraint on the regression coefficients β:

$$\min_{\alpha,\beta} \|\boldsymbol{X} - \boldsymbol{X}\beta\alpha^\top\|_F^2 + \lambda_0 \|\beta\|_2^2 + \lambda \|\beta\|_1, \quad \text{subject to } \|\alpha\|_2^2 = 1, \tag{12}$$

where $\lambda_0, \lambda \ge 0$, and $u \in \mathbb{R}^p, v \in \mathbb{R}^p$. They leveraged the regression/reconstruction error property of PCA to derive sparse principal components and proposed an efficient alternating algorithm for solving Problem (12).

Furthermore, as PCA can also be solved via the SVD of the data matrix, there have also been methods proposed based on the SVD [28, 11]. Specifically, Witten et al. [11] proposed a penalized matrix decomposition (PMD) framework as follows:

$$\min_{d,u,v} \|\boldsymbol{X} - duv^\top\|_F^2 \quad \text{subject to } \|u\|_2^2 = 1, \|u\|_1 \le c_1, \|v\|_2^2 = 1, \|v\|_1 \le c_2. \tag{13}$$

An alternating algorithm was used to solve u and v iteratively. In addition, they also showed that the PMD framework applied to a cross-product matrix solves penalized CCA. We note that the approximation error term in the CoCA objective, $\|\boldsymbol{X} - duv^\top\|_F^2$, is based on the best rank-1 approximation of the data matrix $\boldsymbol{X}$ using the SVD, similar to the PMD framework of [11]. Furthermore, the computation of sparse CoCA draws inspiration from the algorithm proposed by Zou et al. [23], where we iteratively optimize for u and v. Specifically, the update for v can be reformulated as a Lasso problem, which can be efficiently solved using existing optimization techniques.

Another relevant approach is the *joint and individual variation explained (JIVE)* method proposed by Lock et al. [37]. JIVE decomposes multiple data matrices into three terms: a low-rank approximation capturing joint structure between data views, low-rank approximations capturing patterns individual to each data view, and residual noise. This method allows for the simultaneous exploration of shared and view-specific patterns of variability in multi-view data. Tang and Allen introduced another method to capture both individual and

joint patterns across data views called *integrated principal components analysis (iPCA)* [38]. They employ a matrix-variate normal model and utilize penalized covariance estimators to extract these patterns.

CoCA also shares conceptual similarities with the supervised learning method *cooperative learning* proposed by Ding et al. [39] for multi-view data. The method combines the usual squared error loss of predictions with an "agreement" penalty to encourage the predictions from different data views to agree. By varying the weight of the agreement penalty, cooperative learning encompasses the commonly-used early and late fusion and blended versions of these methods. In the regularized setting, considering feature matrices $\boldsymbol{X} \in \mathbb{R}^{n \times p_x}$ and $\boldsymbol{Z} \in \mathbb{R}^{n \times p_z}$, and target vector $y \in \mathbb{R}^n$, cooperative learning seeks to solve the following problem:

$$\min_{\theta_x, \theta_z} \frac{1}{2} \|y - \boldsymbol{X}\theta_x - \boldsymbol{Z}\theta_z\|_2^2 + \frac{\rho}{2} \|(\boldsymbol{X}\theta_x - \boldsymbol{Z}\theta_z)\|_2^2 + \lambda_x P^x(\theta_x) + \lambda_z P^z(\theta_z), \tag{14}$$

where $\theta_x \in \mathbb{R}^{p_x}$ and $\theta_z \in \mathbb{R}^{p_z}$, ρ controls the importance of the agreement term $\|(\boldsymbol{X}\theta_x - \boldsymbol{Z}\theta_z)\|_2^2$, and P^x and P^z are penalty functions. Cooperative learning can be particularly effective when the different data views share some underlying relationship in their signals that can be exploited to boost the signal strength. Furthermore, cooperative learning has also been extended to semi-supervised settings [40], where the agreement penalty is utilized to leverage matched, unlabeled samples across data views to aid the learning process. CoCA further extends this concept of promoting agreement between views to unsupervised settings.

3.3 Simulation study with sparse CoCA

We evaluate sparse CoCA on simulated data drawn from the latent factor model with added noise as described in (6). We introduce sparsity by letting some columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be pure noise dimensions. For simplicity, we consider the two views to have the same dimension $p_1 = p_2$, but this easily extends to views of different dimensions.

Figure 4 presents the simulation results under the different settings. Panel A in Figure 4 corresponds to the scenario where β has 2 non-sparse dimensions. The training data consists of $n = 200$ data points with $p_1 = p_2 = 30$ features in each view. A large test dataset with 5000 data points is generated using the same set of parameters as the training set.

The first subpanel of Panel A shows the estimation error across different values of ρ, aggregated over 10 Monte Carlo runs. The estimation error measures the difference between the true component used to generate the data and the estimated component. The plot exhibits a U-shaped pattern, with the best performance achieved at an intermediate value of ρ. This optimal point demonstrates an estimation error over four orders of magnitude smaller than those seen at the two ends of the solution path.

Moreover, CoCA is compared with sparse PCA and CCA methods implemented using the `PMA` package [11].[§] The optimal parameters for each method are selected using a validation set. The second subpanel shows the reconstruction error on the unseen test set. Due to high variation across experiment runs, the results are also benchmarked against the sparse PCA method in each Monte Carlo run, with the difference shown in the third subpanel and a dotted horizontal line shown at zero. The last subpanel presents the difference with the oracle error in each run, where the oracle error is obtained by using the true component to reconstruct the test set.

CoCA outperforms PCA ($p = 0.054$) and CCA ($p = 0.007$) in reconstruction error, as determined by paired t-tests. In addition, Panel B in Figure 4 considers a more challenging setting with a denser component and a smaller training set of $n = 50$ data points. CoCA again demonstrates improved estimation error, and significantly outperforms PCA ($p = 0.020$) and CCA ($p = 0.002$) in reconstruction error.

[§]Note that although CoCA with $\rho = 0$ and $\rho \to \infty$ corresponds to PCA and CCA, respectively, the *sparse* PCA and CCA methods proposed in [11] are different than sparse CoCA with $\rho = 0$ and $\rho \to \infty$.
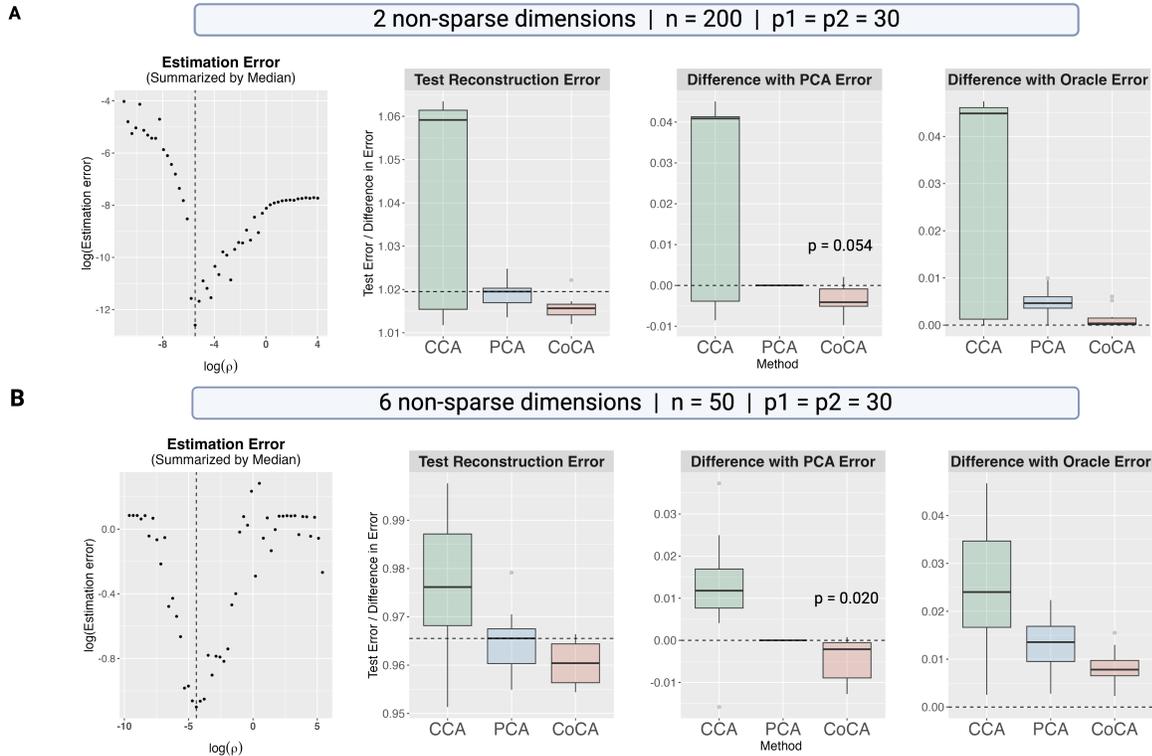
Figure 4: *Simulation results comparing CoCA with sparse PCA and CCA under different settings.* Panel A corresponds to the simulation setting with a simple true component with 2 non-sparse dimensions, $n = 200$, and $p = p_1 = p_2 = 30$. Panel B considers a denser component and a smaller training set with $n = 50$. In both settings, CoCA demonstrates benefits in reducing estimation and reconstruction error by leveraging shared underlying relationships between data views. The results are aggregated over 10 Monte Carlo runs.

4 Real data examples

4.1 Integration of CT scan-derived radiomics and laboratory measurements of COVID-19 patients

We applied CoCA to integrate CT scan-derived radiomics features and laboratory result features measured on a cohort of 127 COVID-19 patients [15]. The goal of the analysis is to identify the component that captures important signals within data views while exhibiting strong correlations across data views. In addition, we utilized the scores derived from the two data views for each patient under a linear discriminant analysis (LDA) model to predict their risk of severe disease progression requiring intensive care unit (ICU) admission.

The radiomics data consists of 1576 quantitative features extracted from CT scans, capturing various aspects of the imaged tissues, including density distribution and texture characteristics. These features provide an imaging-based characterization of the COVID-19-induced abnormalities. The clinical data, on the other hand, comprises 22 key laboratory measurements, including hemoglobin, white blood cell count, albumin, lactate dehydrogenase, D-dimer, C-reactive protein, and ferritin. These laboratory tests offer insights into the systemic effects of the disease and the patient's overall health status.

We split the dataset of 127 patients into training and test sets of 102 and 25 patients, respectively. CV was employed to select the optimal hyperparameters for each method: ρ and λ for CoCA, and sparsity levels for sparse PCA and sparse CCA. For sparse PCA and sparse CCA, we utilized the `PMA` package, which provides efficient implementations of these methods and allows for the specification of sparsity levels. The CV procedure involved deriving components using the training set and then estimating the test error by using the derived scores to predict ICU outcomes on the validation set.
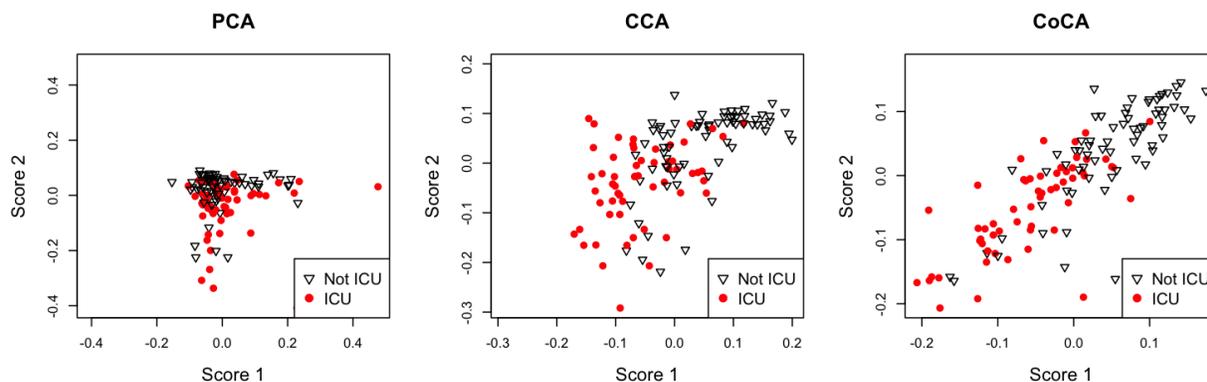
10

Figure 5: *Comparison of multi-view scores derived from radiomics and clinical data views for a cohort of COVID-19 patients, using PCA, CCA, and CoCA, respectively.* We have seen this figure as a motivating example in the introduction. Here each point represents a patient, colored by ICU admission outcome. CoCA achieves clearer separation between ICU and non-ICU patients and better alignment between the two data views.
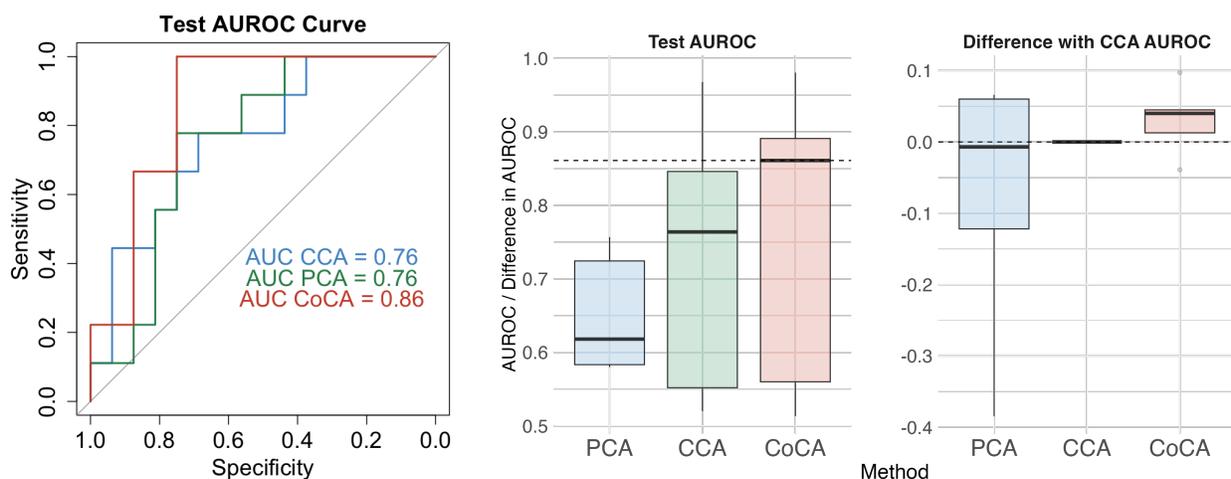

Figure 6: *Test performance of using the multi-view scores from radiomics and clinical data for predicting ICU outcomes based on AUROC.* The left panel shows the AUROC curve on an unseen test set: CoCA outperforms both PCA and CCA, improving the AUROC from 0.76 to 0.86. The middle panel shows the AUROC performance across five experiment runs with different random splits of the training and test sets: the median AUROC for CoCA is higher than the upper quantile AUROC of both PCA and CCA. The right panel plots the difference in AUROC between each method and CCA across experiment runs, with positive values indicating that the method outperformes CCA.

Figure 5 visualizes the scores for each patient obtained from sparse PCA, sparse CCA, and sparse CoCA. Each point represents a patient, and the color indicates their ICU admission outcome. The plot provides a visual assessment of how well the derived components can separate patients with different disease progression outcomes. From the CoCA plot, we observe a clearer separation between patients who required ICU admission and those who did not, compared to the PCA and CCA plots. This suggests that the component learned by CoCA capture more informative and discriminative patterns that are associated with the severity of COVID-19 progression. Moreover, the CoCA plot also exhibits better alignment of the scores derived from the two data modalities. This indicates that CoCA effectively identifies components that are consistent and strongly correlated across the different data types, reflecting a shared underlying biological signal.

In Figure 6, we evaluate the test performance of using the scores for predicting ICU outcomes based on the area under the receiver operating characteristic curve (AUROC) metric. The left panel shows the AUROC curve on the test set: CoCA outperforms both PCA and CCA, improving the AUROC from 0.76 to 0.86. To assess the robustness of these findings, we conducted the experiment for five times with different

11

random splits of the training and test sets. The middle panel shows the AUROC performance across different experiment runs: the median AUROC for CoCA (represented by the center line) is higher than that of the upper quantile AUROC of both PCA and CCA. With the inherent variability in performance across different runs, we further benchmarked the methods against CCA in each experiment run. The right panel plots the difference in AUROC between each method and CCA across experiment runs, with positive values indicating that the method outperformes CCA.

Figure 7: *Test performance of using the multi-view scores from radiomics and clinical data for predicting ICU outcomes based on AUPRC.* The left panel shows the AUPRC curve on an unseen test set: CoCA outperforms both PCA and CCA, improving the AUPRC from 0.86 (CCA) and 0.88 (PCA) to 0.94 (CoCA). The middle panel shows the AUPRC performance across five experiment runs with different random splits of the training and test sets: the median AUPRC for CoCA is higher than the upper quantile AUPRC of both PCA and CCA. The right panel plots the difference in AUPRC between each method and CCA across experiment runs, with positive values indicating that the method outperformed CCA.
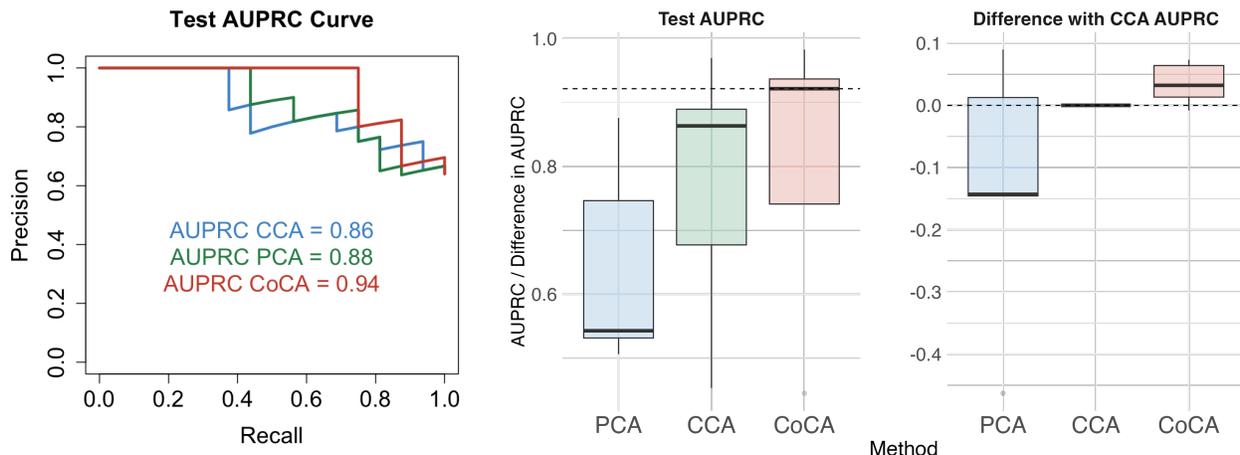
Moreover, to account for the class imbalance in the data, where 57 out of 127 patients required ICU admission, we also evaluated the performance using the area under the precision-recall curve (AUPRC). As before, the left panel of Figure 7 shows the AUPRC curve on the test set for one split of the training and test sets. The middle panel shows the AUPRC performance for each method across five different random splits of the training and test sets. The right panel plots the difference in AUPRC between each method and CCA. These results highlight the effectiveness of CoCA. By leveraging the shared information across the radiomics and clinical data views, CoCA is able to extract more informative components that are consistently present across views and more predictive of disease progression.

We also performed model interpretation by examining the top selected features from the two data modalities. The selected radiomics features include the first-order statistics such as total energy in the LLL (low-low-low) wavelet decomposed image, the gray level dependence matrix (GLDM) dependence variance in the LLH (low-low-high) wavelet decomposed image, and the gray level size zone matrix (GLSZM) small area emphasis in the LLL (low-low-low) wavelet decomposed image, among others. They capture the intensity and textural characteristics of the imaged tissues, reflecting the heterogeneity and complexity of COVID-19-induced abnormalities in the lungs. On the other hand, the top selected clinical features, albumin (ALB) and lactate dehydrogenase (LDH), are associated with disease severity and tissue damage. Specifically, low ALB levels indicate a severe systemic inflammatory response, potentially leading to pulmonary edema, while high LDH levels mark the extent of lung injury. The correlation between these radiomics and clinical features can be explained by the underlying pathophysiology of COVID-19, where the systemic inflammatory response and tissue damage manifest as alterations in the intensity and texture of the CT images. CoCA's integration of these features allows for a more comprehensive characterization of disease severity and progression, potentially improving risk stratification and guiding clinical decision-making.
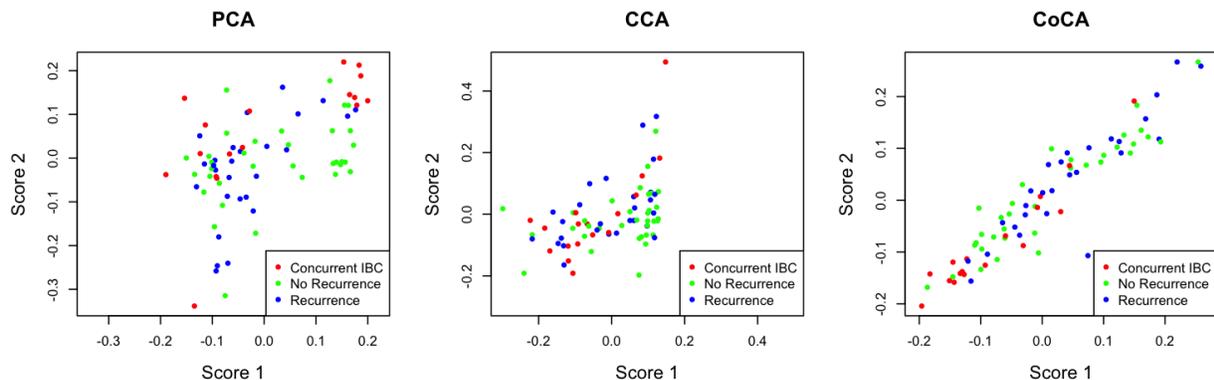
Figure 8: *Comparison of multi-view scores derived from epithelial and stromal gene expression data for a cohort of DCIS patients with different disease outcomes, using PCA, CCA, and CoCA, respectively.* Here each point represents a patient, colored by differnt outcome groups, i.e. concurrent contralateral IBC, disease recurrence and no-recurrence. CoCA achieves clearer separation between patients with different outcomes and better alignment between the two molecular data views.

4.2 Integrating epithelial and stromal gene expression data of breast ductal carcinoma in situ (DCIS) patients

We also applied CoCA to integrate epithelial and stromal gene expression data from a study on ductal carcinoma in situ (DCIS), the most common precursor of invasive breast cancer (IBC) [41]. In the dataset, the Resource of Archival Breast Tissue (RAHBT) cohort contained 78 patients, including 17 patients with concurrent contralateral IBC, 27 patients with disease recurrence, and 34 controls without disease recurrence. The data consisted of paired epithelial and stromal gene expression profiles obtained through laser capture microdissection, allowing for the separate analysis of these two key tissue compartments. The goal of the analysis was to identify the component that captures important signals within and across different molecular data views, providing insights into the spectrum of molecular changes in DCIS and potential predictors of disease progression. In addition, we utilized the scores derived from the two data views for each patient under a LDA model to classify DCIS patients into concurrent IBC, disease recurrence and no-recurrence.

Both epithelial and stromal gene expression data consist of 60,662 gene expression features, which were screened by their variance across the subjects. We split the dataset of 78 patients into training and test sets of 70 and 8 patients, respectively. To ensure robustness of our results, we conducted the same set of experiments across 10 different random splits of the training and test sets. CV was employed to select the optimal hyperparameters for each method: ρ and λ for CoCA, and sparsity levels for sparse PCA and sparse CCA. For sparse PCA and sparse CCA, we utilized the `PMA` package. The CV procedure involved deriving components using the training set and then estimating the test error by using the scores to predict disease outcomes on the validation set.

Figure 8 visualizes the multi-view scores for each patient obtained from sparse PCA, sparse CCA, and sparse CoCA. Each point represents a patient, and the color indicates their disease outcome, i.e. concurrent contralateral IBC, future recurrence, or no recurrence. The plot provides a visual assessment of how well the derived components can separate patients with different disease outcomes. The CoCA plot demonstrates a more distinct separation of patients with differnt outcomes, as well as enhanced alignment between the two sets of scores derived from epithelial and stromal data views. This suggests that CoCA more effectively captures shared biological signals across tissue compartments that may be crucial in distinguishing DCIS progression patterns and risk of invasive cancer development.

Table 1 compares the performance of Sparse PCA, Sparse CCA, and Sparse CoCA in differentiating DCIS patients into outcome groups. The multi-class AUROC metric [42] is used for evaluation. The table shows the mean and standard deviation (SD) of both the absolute AUROC values and the performance relative to CCA across different splits of the training and test sets. Sparse CoCA achieves the highest mean AUROC (0.753) on unseen test sets and shows the best performance relative to CCA (+0.060), suggesting that it

offers superior discriminative power for classifying DCIS outcomes.

Methods	Test AUROC		Relative to CCA	
	Mean	SD	Mean	SD
Sparse PCA	0.628	0.141	-0.066	0.010
Sparse CCA	0.694	0.110	0.000	0.000
Sparse CoCA	**0.753**	**0.089**	**0.060**	**0.043**

Table 1: *Test performance of Sparse PCA, Sparse CCA, and Sparse CoCA in differentiating DCIS patients.* The first two columns in the table show the mean and standard deviation (SD) of the multi-class AUROC on the test set across different splits of the training and test sets. The third and fourth column show the AUROC difference relative to sparse CCA, with positive values indicating better performance than sparse CCA.

We also performed model interpretation by examining the top selected features from the paired epithelial and stromal gene expression profiles. Several genes were consistently selected as important features in both epithelial and stromal compartments, including IGKC, TFF1, MUCL1, SLC39A6, and ANKRD30A. IGKC, a key component of the humoral immune system, has been associated with improved distant metastasis-free survival in early breast cancer, suggesting a potential protective role of tumor-infiltrating immune cells [43, 44]. TFF1, an estrogen-regulated protein, has been shown to stimulate migration of human breast cancer cells, potentially contributing to disease recurrence [45]. These genes are worth further investigation for their roles in DCIS progression.

5   Discussion

In this paper, we introduce Cooperative Component Analysis (CoCA), a novel unsupervised learning method for integrating multiple data views, by identifying the component that simultaneously captures significant within-view variance and exhibits strong cross-view correlation. CoCA encourages alignment across data views through an agreement penalty. By varying the weight of the agreement penalty in the objective, CoCA encompasses the commonly used principal component analysis (PCA) and canonical correlation analysis (CCA) as special cases at the two ends of the solution path. This allows CoCA to choose the degree of agreement in a data-adaptive manner.

Additionally, this approach has promises to be extended to scenarios where explicit grouping information is unavailable or suboptimal. In such cases, we adaptively identify the optimal grouping that best captures the underlying shared structure across variables, with the potential to improve upon PCA without prior knowledge of the grouping. Moreover, it is also possible to define a rank-K generalization of CoCA, by replacing u and v with rank-K orthogonal matrices in (1). However, it is more difficult to define a computationally tractable rank-K sparse formulation of CoCA, due to challenges introduced by the orthogonality constraints, and so we leave this for future work.

Furthermore, to enhance interpretability, CoCA incorporates the Lasso penalty to yield a sparse component. This facilitates the identification of key features driving the observed patterns. The effectiveness of CoCA has implications for improving our understanding of complex systems and uncovering novel insights in an era of increasingly multi-view data.

Acknowledgements

References

[1] 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.

[2] David Reich, Alkes L Price, and Nick Patterson. Principal component analysis of genetic data. *Nature Genetics*, 40(5):491–492, 2008.

[3] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[4] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

[5] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, 2008.

[6] Kristian D. Allee, Chuong Do, and Fellipe G. Raymundo. Principal component analysis and factor analysis in accounting research. *Journal of Financial Reporting*, 7:1–39, September 2022.

[7] Cristiane Guellis, Daniele C. Valério, Guilherme G. Bessegato, Marcela Boroski, Josiane C. Dragunski, and Cleber A. Lindino. Non-targeted method to detect honey adulteration: Combination of electrochemical and spectrophotometric responses with principal component analysis. *Journal of Food Composition and Analysis*, 89:103466, 2020.

[8] Mahsa Ghorbani and Edwin K. P. Chong. Stock price prediction using principal components. *Plos One*, 15:e0230124, March 2020.

[9] John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[10] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 162–190. Springer, 1992.

[11] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[12] Chao Gao, Zongming Ma, and Harrison H Zhou. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, pages 2074–2101, 2017.

[13] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

[14] Roemer J Janse, Tiny Hoekstra, Kitty J Jager, Carmine Zoccali, Giovanni Tripepi, Friedo W Dekker, and Merel Van Diepen. Conducting correlation analysis: important limitations and pitfalls. *Clinical Kidney Journal*, 14(11):2332–2337, 2021.

[15] Ahmet Gorkem Er, Daisy Yi Ding, Berrin Er, Mertcan Uzun, Mehmet Cakmak, Christoph Sadee, Gamze Durhan, Mustafa Nasuh Ozmen, Mine Durusu Tanriover, Arzu Topeli, et al. Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a covid-19 cohort study. *NPJ Digital Medicine*, 7(1):117, 2024.

[16] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[17] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[18] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

[19] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[20] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in

high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[21] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[22] Mengjie Chen, Chao Gao, Zhao Ren, and Harrison H Zhou. Sparse cca via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*, 2013.

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

[24] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

[25] Hui Zou and Lingzhou Xue. A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, 106(8):1311–1320, 2018.

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

[27] Alexandre d'Aspremont, Laurent Ghaoui, Michael Jordan, and Gert Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Advances in Neural Information Processing Systems*, 17, 2004.

[28] Haipeng Shen and Jianhua Z Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

[29] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2), 2010.

[30] Zhaosong Lu and Yong Zhang. An augmented lagrangian approach for sparse principal component analysis. *Mathematical Programming*, 135:149–193, 2012.

[31] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(4), 2013.

[32] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. In *BMC proceedings*, volume 1, page S119. Springer, 2007.

[33] David R Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83:331–353, 2011.

[34] Qing Mai and Xin Zhang. An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics*, 75(3):734–744, 2019.

[35] Shixiang Chen, Shiqian Ma, Lingzhou Xue, and Hui Zou. An alternating manifold proximal gradient method for sparse pca and sparse cca. *arXiv preprint arXiv:1903.11576*, 2019.

[36] Ofir Lindenbaum, Moshe Salhov, Amir Averbuch, and Yuval Kluger. L0-sparse canonical correlation analysis. In *International Conference on Learning Representations*, 2021.

[37] Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013.

[38] Tiffany M Tang and Genevera I Allen. Integrated principal components analysis. *Journal of Machine Learning Research*, 22(198):1–71, 2021.

[39] 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.

[40] Daisy Yi Ding, Xiaotao Shen, Michael Snyder, and Robert Tibshirani. Semi-supervised cooperative learning for multiomics data fusion. *Workshop on Machine Learning for Multimodal Healthcare Data, The Fortieth International Conference on Machine Learning (ICML)*, 2023.

[41] Siri H Strand, Belén Rivero-Gutiérrez, Kathleen E Houlahan, Jose A Seoane, Lorraine M King, Tyler

Risom, Lunden A Simpson, Sujay Vennam, Aziz Khan, Luis Cisneros, et al. Molecular classification and biomarkers of clinical outcome in breast ductal carcinoma in situ: Analysis of tbcrc 038 and rahbt cohorts. *Cancer Cell*, 40(12):1521–1536, 2022.

[42] David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

[43] Marcus Schmidt, Karolina Edlund, Jan G Hengstler, Anne-Sophie Heimes, Katrin Almstedt, Antje Lebrecht, Slavomir Krajnak, Marco J Battista, Walburgis Brenner, Annette Hasenburg, et al. Prognostic impact of immunoglobulin kappa c (igkc) in early breast cancer. *Cancers*, 13(14):3626, 2021.

[44] Marcus Schmidt, Patrick Micke, Mathias Gehrmann, and Jan G Hengstler. Immunoglobulin kappa chain as an immunologic biomarker of prognosis and chemotherapy response in solid tumors. *Oncoimmunology*, 1(7):1156–1158, 2012.

[45] Sara J Prest, Felicity EB May, and Bruce R Westley. The estrogen-regulated protein, tff1, stimulates migration of human breast cancer cells. *The FASEB Journal*, 16(6):592–594, 2002.

[46] Patrick O Perry. *Cross-validation for unsupervised learning.* Stanford University, 2009.

A Cross-validation procedure for CoCA

In this section, we describe three options for the cross-validation (CV) procedure to determine the optimal values of hyperparameters ρ and λ in CoCA: K-fold CV and "speckled CV" for unsupervised settings, and K-fold CV for supervised settings.

For K-fold CV in unsupervised settings, the training set is divided into K folds. For each combination of ρ and λ in a pre-defined grid, we iterate through the folds. In each iteration, we train CoCA on $K-1$ folds, excluding the current fold, and then calculate the reconstruction error on the held-out fold using the estimated component. This process is repeated K times, and the average reconstruction error across all K folds is computed for each hyperparameter combination. The ρ and λ values that minimize this average reconstruction error across folds are selected as optimal.

An alternative approach is "speckled CV", where a proportion of values in the data matrix are randomly masked. These masked values serve as a validation set for hyperparameter selection, while the unmasked values are used for training. In this approach, we evaluate how well the component derived from the unmasked data can reconstruct the masked elements. Specifically, for each combination of hyperparameters, we compute the components using the unmasked data, reconstruct the entire matrix, and then calculate the reconstruction error for the masked elements. The hyperparameters that minimize this reconstruction error are selected as optimal. This procedure is described in more detail in [46].

When an outcome of interest is available and is desired to be used for hyperparameter selection we follow a similar K-fold CV procedure, but with a focus on predicting the outcome of interest rather than reconstruction error. For each hyperparameter combination, we apply CoCA on the training data and derive multi-view scores. We then use K-1 folds of these scores to predict the responses in the K-th fold. This process is repeated K times, with each fold serving as the validation set once. We calculate a prediction error metric (such as misclassification rate or mean squared error) for each fold. The average prediction error across all folds is computed for each hyperparameter combination, and the ρ and λ values that minimize this average prediction error are selected as optimal.

B Proof of Theorem 1

We begin by establishing that the solutions $\hat{u}, \hat{v}$ to Problem (3) satisfy (4) and (5), using standard SVD arguments. Define

$$\mathcal{L}(u,v) := \frac{1}{2}v^\top v - u^\top \boldsymbol{X}v + \frac{\rho}{2}\|\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2\|_2^2.$$

This is simply the criterion of Problem (3) minus $\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})$, and it follows that Problem (3) is equivalent to

$$\min_{u,v} \mathcal{L}(u,v), \quad \text{subject to } \|u\|_2^2 = 1. \tag{15}$$

Let $\hat{v}(u)$ be the solution to $\min_{v}\mathcal{L}(u,v)$, and observe that

$$\boldsymbol{X}^\top u = (\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})\hat{v}(u). \tag{16}$$

Substituting this expression for v in (15), it follows that $\hat{u}$ is the solution to

$$\min_{u} -u^\top \boldsymbol{X}(\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top u, \quad \text{subject to } \|u\|_2^2 = 1.$$

Therefore $\hat{u}$ satisfies (4) as claimed:

$$\boldsymbol{X}(\boldsymbol{I} + \rho \boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top \hat{u} = \lambda_1 \hat{u}, \quad \|\hat{u}\|_2 = 1. \tag{17}$$

Now, let $\hat{u}(v)$ solve $\min_u \mathcal{L}(u,v)$ subject to $\|u\|_2^2 = 1$. Observe that

$$\hat{u}(v) = \frac{\boldsymbol{X}v}{\|\boldsymbol{X}v\|_2}. \tag{18}$$

Plugging this in to (4) and applying $\boldsymbol{X}^\top$ to each side gives:

$$\frac{1}{\|\boldsymbol{X}\hat{v}\|_2}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{I}+\rho\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\hat{v} = \frac{1}{\|\boldsymbol{X}\hat{v}\|_2}\lambda_1\boldsymbol{X}^\top\boldsymbol{X}\hat{v}, \Longleftrightarrow$$
$$(\boldsymbol{I}+\rho\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\hat{v} = \lambda_1\hat{v}.$$

We have now solved for $\hat{u}$ and $\hat{v}$, the latter up to a constant of proportionality. To determine what this constant is, note that from (16), (18) and the previous display we have that

$$\hat{v} = (\boldsymbol{I}+\rho\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\hat{u} = \frac{1}{\|\boldsymbol{X}\hat{v}\|_2}(\boldsymbol{I}+\rho\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\hat{v} = \frac{\lambda_1}{\|\boldsymbol{X}\hat{v}\|_2}\hat{v}.$$

In other words, $\hat{v}$ is proportional to the leading eigenvector of the matrix $(\boldsymbol{I}+\rho\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}$, and has variance equal to the square of the leading eigenvalue of this matrix:

$$(\boldsymbol{I}+\rho\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\hat{v} = \lambda_1\hat{v}, \quad \|\boldsymbol{X}\hat{v}\|_2 = \lambda_1. \tag{19}$$

Now we establish the statements in Theorem 1 regarding the equivalence of CoCA to PCA and CCA at the two ends of its solution path. The equivalence to PCA is true by definition – that is, when $\rho = 0$ by definition $\hat{v}$ is proportional to the first principal component of $\boldsymbol{X}$ – and so we concentrate on deriving the equivalence to CCA as $\rho \to \infty$.

Since we have assumed $\mathrm{rank}(\boldsymbol{X}) = p$, $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible. In that case, note that, as $\rho \to \infty$, $\hat{v}/\|\hat{v}\|_2$ converges to the leading eigenvector of the matrix $(\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}$. The eigenpairs of this matrix are the solutions $(v_k, \lambda_k), k = 1, \ldots, p$ to the eigenproblem

$$(\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}v = \lambda v, \quad \|v\|^2 = 1. \tag{20}$$

We begin by showing that each pair of canonical directions between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ defines an eigenvector of $(\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top\boldsymbol{X}v$. Recall that there are a total of $p^* := \min\{p_1, p_2\}$ canonical pairs $(\tilde{v}_1, \tilde{v}_2, \gamma)$ between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, each of which satisfy the stationary conditions

$$\begin{aligned}(\boldsymbol{X}_1^\top\boldsymbol{X}_2)\tilde{v}_2 &= \sqrt{\gamma}(\boldsymbol{X}_1^\top\boldsymbol{X}_1)\tilde{v}_1, \quad \tilde{v}_1^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1\tilde{v}_1 = 1 \\ (\boldsymbol{X}_2^\top\boldsymbol{X}_1)\tilde{v}_1 &= \sqrt{\gamma}(\boldsymbol{X}_2^\top\boldsymbol{X}_2)\tilde{v}_2, \quad \tilde{v}_2^\top\boldsymbol{X}_2^\top\boldsymbol{X}_2\tilde{v}_2 = 1,\end{aligned} \tag{21}$$

where γ defines the squared cross-correlation between $\tilde{v}_1$ and $\tilde{v}_2$. We will assume for ease of exposition that the cross-correlations have multiplicity 1, so that (21) defines $(\tilde{v}_1, \tilde{v}_2, \gamma)$ up to the sign of $\tilde{v}_1, \tilde{v}_2$. To see the relationship between the eigenvectors v_k and canonical directions $\tilde{v}_1, \tilde{v}_2$, notice that the eigenvector equation (20) can be rearranged to

$$\boldsymbol{D}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}v = \frac{1}{\lambda}\boldsymbol{X}^\top\boldsymbol{X}v.$$

Written in block-matrix form, this is

$$\boldsymbol{X}_1^\top\boldsymbol{X}_1 v_1 - \boldsymbol{X}_1^\top\boldsymbol{X}_2 v_2 = \frac{1}{\lambda}(\boldsymbol{X}_1^\top\boldsymbol{X}_1 v_1 + \boldsymbol{X}_1^\top\boldsymbol{X}_2 v_2)$$
$$-\boldsymbol{X}_2^\top\boldsymbol{X}_1 v_1 + \boldsymbol{X}_2^\top\boldsymbol{X}_2 v_2 = \frac{1}{\lambda}(\boldsymbol{X}_2^\top\boldsymbol{X}_1 v_1 + \boldsymbol{X}_2^\top\boldsymbol{X}_2 v_2),$$

which can be rearranged to read

$$\begin{aligned}(\boldsymbol{X}_1^\top\boldsymbol{X}_2)v_2 &= \frac{(1-\lambda^{-1})}{(1+\lambda^{-1})}(\boldsymbol{X}_1^\top\boldsymbol{X}_1)v_1 \\ (\boldsymbol{X}_2^\top\boldsymbol{X}_1)v_1 &= \frac{(1-\lambda^{-1})}{(1+\lambda^{-1})}(\boldsymbol{X}_2^\top\boldsymbol{X}_2)v_2,\end{aligned} \tag{22}$$

This is simply (21) with $\sqrt{\gamma} = \frac{(1-\lambda^{-1})}{(1+\lambda^{-1})}$. It follows that each canonical pair corresponds to an eigenpair of (20), meaning precisely that if $(\tilde{v}_1, \tilde{v}_2, \gamma)$ is a canonical pair, then (letting $\tilde{v} = (\tilde{v}_1, \tilde{v}_2)$)

$$\left(\frac{\tilde{v}}{\|\tilde{v}\|_2}, \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}\right)$$

19

is an eigenpair of (20). Additionally, notice that if (v, λ) is an eigenpair satisfying (20), then $(\boldsymbol{D}v, 1/\lambda)$ is also an eigenpair satisfying (20). As a result,

$$\left(\frac{\boldsymbol{D}\tilde{v}}{\|\tilde{v}\|_2}, \frac{1 - \sqrt{\gamma}}{1 + \sqrt{\gamma}} \right)$$

is also an eigenpair of (20).

This characterizes $2p^*$ eigenpairs of (20). The remaining $p - 2p^*$ eigenpairs correspond to the $p_1 - p_2$-dimensional subspace of $\mathrm{col}(\boldsymbol{X}_1)$ that lies in the null space of $\boldsymbol{X}_2^\top$ (assuming without loss of generality that $p_2 < p_1$); for any such eigenpair (v, λ) it is the case that $v_2 = 0$ and $\lambda = 1$. None of these are the leading eigenpair, since $\boldsymbol{X}_1^\top \boldsymbol{X}_2 \neq 0$ and therefore $\lambda_1 > 1$. We conclude that the leading canonical pair $(\tilde{v}_{11}, \tilde{v}_{12}, \gamma_1)$ corresponds to leading eigenpair of $(\boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D})^{-1}\boldsymbol{X}^\top \boldsymbol{X}$. This is the desired claim.

C   Alternative formulations of CoCA

As mentioned, we have experimented with alternative formulations of CoCA.

Alternative formulation 1   One alternative formulation is to minimize the following objective:

$$\min_{u,v} \frac{1}{2}||\boldsymbol{X} - \theta \cdot \boldsymbol{X}vu^\top||_F^2 + \frac{\rho}{2}||\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2||_2^2 + \lambda||v||_1, \text{ subject to } ||u||_2^2 = 1, ||v||_2^2 = 1. \tag{23}$$

Here $\rho \geq 0$ controls the relative importance of the agreement penalty and $\lambda \geq 0$ controls the level of sparsity. $v \in \mathbb{R}^p$ and $u \in \mathbb{R}^n$ are vectors, and θ is a scalar. v is partitioned as $v = (v_1, v_2)$, where $v_1 \in \mathbb{R}^{p_1}$ corresponds to $\boldsymbol{X}_1$ and $v_2 \in \mathbb{R}^{p_2}$ to $\boldsymbol{X}_2$. However, this formulation has an important drawback: as ρ approaches infinity, the solution does not converge to the CCA solution.

Alternative formulation 2   The limitation of (23) prompted us to explore another alternative formulation:

$$\min_{\theta,u,v} \frac{1}{2}||\boldsymbol{X} - \theta \cdot \boldsymbol{X}vu^\top||_F^2 + \frac{\rho}{2}||\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2||_2^2 + \lambda||v||_1, \text{ subject to } ||u||_2^2 \leq 1, ||\boldsymbol{X}v||_2^2 = 1. \tag{24}$$

Setting $\lambda = 0$, it can be shown that this formulation corresponds to PCA when $\rho = 0$ and CCA as $\rho \to \infty$; we omit the derivation as is it similar to the proof of Theorem 1.

The difficulty with this formulation is that constraint $||\boldsymbol{X}v||_2^2 = 1$ makes the problem challenging to solve. This can be finessed in the following way. We begin from (24), without sparsity:

$$\min_{\theta,u,v} \frac{1}{2}||\boldsymbol{X} - \theta \cdot \boldsymbol{X}vu^\top||_F^2 + \frac{\rho}{2}||\boldsymbol{X}_1 v_1 - \boldsymbol{X}_2 v_2||_2^2, \text{ subject to } ||u||_2^2 \leq 1, ||\boldsymbol{X}v||_2^2 = 1. \tag{25}$$

First, solving explicitly for θ, u, we see that (25) is equivalent to the following optimization problem:

$$\max_v v^\top \boldsymbol{\Sigma}^2 v - \rho v^\top \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}v, \quad \text{subject to} \quad v^\top \boldsymbol{\Sigma}v = 1.$$

By adding $c\boldsymbol{\Sigma}$ for a sufficiently large value of c–for instance taking c to be the maximum eigenvalue of $\rho\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}$ will suffice–we can reformulate this as the generalized eigenproblem

$$\max_v v^\top \boldsymbol{M}_\rho v, \quad \text{subject to} \quad v^\top \boldsymbol{\Sigma}v = 1, \tag{26}$$

where $\boldsymbol{M}_\rho = \boldsymbol{\Sigma}^2 - \rho v^\top \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D} + c\boldsymbol{I}$ is PSD. Problem (26) is equivalent to the following convex relaxation,

$$\min_{\boldsymbol{V} \in \mathbb{R}^{p \times p}} -\mathrm{tr}(\boldsymbol{M}_\rho \boldsymbol{V}), \quad \text{subject to} \quad \mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{V}) = 1, \boldsymbol{V} \succeq 0,$$

in the sense that the solution will be the rank-1 matrix $\hat{\boldsymbol{V}} = v_1 v_1^\top$ where v_1 is the leading eigenvector of $\boldsymbol{M}_\rho$.

Now, to incorporate sparsity, we constrain the ℓ^1 norm of $\boldsymbol{V}$ to be at most k^2:

$$\min_{\boldsymbol{V} \in \mathbb{R}^{p \times p}} -\text{tr}(\boldsymbol{M}_\rho \boldsymbol{V}), \quad \text{subject to } \text{tr}(\boldsymbol{\Sigma V}) = 1, \boldsymbol{V} \succeq 0, \|\boldsymbol{V}\|_1 \leq k^2. \tag{27}$$

The problem (27) is a semidefinite program (SDP), and can either be solved exactly using standard SDP solvers, or approximately using a first-order method. However, it should be noted that the solution is no longer guaranteed to be rank-1. Additionally, solving semidefinite programs becomes practically difficult for large-scale problems. Instead, we opt for the CoCA formulation (7) presented earlier. This formulation satisfies our key requirements: it encompasses PCA and CCA as special cases at the extremes of the regularization path, and it remains computationally tractable when incorporating sparsity, allowing for efficient optimization using algorithms like `glmnet` in an alternating scheme.