Convex Multi-view Subspace Learning

Martha White, Yaoliang Yu, Xinhua Zhang[*] and Dale Schuurmans
Department of Computing Science, University of Alberta, Edmonton AB T6G 2E8, Canada
{whitem,yaoliang,xinhua2,dale}@cs.ualberta.ca

Abstract

Subspace learning seeks a low dimensional representation of data that enables accurate reconstruction. However, in many applications, data is obtained from multiple sources rather than a single source (e.g. an object might be viewed by cameras at different angles, or a document might consist of text and images). The conditional independence of separate sources imposes constraints on their shared latent representation, which, if respected, can improve the quality of a learned low dimensional representation. In this paper, we present a convex formulation of multi-view subspace learning that enforces conditional independence while reducing dimensionality. For this formulation, we develop an efficient algorithm that recovers an optimal data reconstruction by exploiting an implicit convex regularizer, then recovers the corresponding latent representation and reconstruction model, jointly and optimally. Experiments illustrate that the proposed method produces high quality results.

1 Introduction

Dimensionality reduction is one of the most important forms of unsupervised learning, with roots dating to the origins of data analysis. Re-expressing high dimensional data in a low dimensional representation has been used to discover important latent information about individual data items, visualize entire data sets to uncover their global organization, and even improve subsequent clustering or supervised learning [1]. Modern data is increasingly complex, however, with descriptions of increasing size and heterogeneity. For example, multimedia data analysis considers data objects (e.g. documents or webpages) described by related text, image, video, and audio components. *Multi-view learning* focuses on the analysis of such multi-modal data by exploiting its implicit conditional independence structure. For example, given multiple camera views of a single object, the particular idiosyncrasies of each camera are generally independent, hence the images they capture will be *conditionally independent* given the scene. Similarly, the idiosyncrasies of text and images are generally conditionally independent given a topic. The goal of multi-view learning, therefore, is to use known conditional independence structure to improve the quality of learning results.

In this paper we focus on the problem of *multi-view subspace learning*: reducing dimensionality when data consists of multiple, conditionally independent sources. Classically, multi-view subspace learning has been achieved by an application of *canonical correlation analysis* (CCA) [2, 3]. In particular, many successes have been achieved in using CCA to recover meaningful latent representations in a multi-view setting [4–6]. Such work has been extended to probabilistic [7] and sparse formulations [8]. However, a key limitation of CCA-based approaches is that they only admit efficient global solutions when using the squared-error loss (i.e. Gaussian models), while extensions to robust models have had to settle for approximate solutions [9].

By contrast, in the *single-view* setting, recent work has developed new generalizations of subspace learning that can accommodate arbitrary convex losses [10–12]. These papers replace the hard bound on the dimension of the latent representation with a structured convex regularizer that still reduces rank, but in a relaxed manner that admits greater flexibility while retaining tractable formulations.

[*] Xinhua Zhang is now at the National ICT Australia (NICTA), Machine Learning Group.

Subspace learning can be achieved in this case by first recovering an optimal reduced rank response matrix and then extracting the latent representation and reconstruction model. Such formulations have recently been extended to the multi-view case [13, 14]. Unfortunately, the multi-view formulation of subspace learning does not have an obvious convex form, and current work has resorted to local training methods based on alternating descent minimization (or approximating intractable integrals). Consequently, there is no guarantee of recovering a *globally optimal* subspace.

In this paper we provide a formulation of multi-view subspace learning that can be solved *optimally* and *efficiently*. We achieve this by adapting the new single-view training methods of [11, 12] to the multi-view case. After deriving a new formulation of multi-view subspace learning that allows a *global* solution, we also derive efficient new algorithms. The outcome is an efficient approach to multi-view subspace discovery that can produce high quality repeatable results.

Notation: We use I_k for the $k \times k$ identity matrix, A' for the transpose of matrix A, $\|\cdot\|_2$ for the Euclidean norm, $\|X\|_F = \sqrt{\operatorname{tr}(X'X)}$ for the Frobenius norm and $\|X\|_{\operatorname{tr}} = \sum_i \sigma_i(X)$ for the trace norm, where $\sigma_i(X)$ is the ith singular value of X.

2 Background

Assume one is given t paired observations $\left\{ \left[\begin{smallmatrix} \mathbf{x}_j \\ \mathbf{y}_j \end{smallmatrix}\right] \right\}$ consisting of two views: an x-view and a y-view, of lengths m and n respectively. The goal of multi-view subspace learning is to infer, for each pair, a shared latent representation, $\mathbf{h}_j$, of dimension $k < \min(n, m)$, such that the original data can be accurately modeled. We first consider a linear formulation. Given paired observations the goal is to infer a set of latent representations, $\mathbf{h}_j$, and reconstruction models, A and B, such that $A\mathbf{h}_j \approx \mathbf{x}_j$ and $B\mathbf{h}_j \approx \mathbf{y}_j$ for all j. Let X denote the $n \times t$ matrix of x observations, Y the $m \times t$ matrix of y observations, and $Z = \left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$ the concatenated $(n+m) \times t$ data matrix. The problem can then be expressed as recovering a $(n+m) \times k$ matrix of model parameters, $C = \left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]$, and a $k \times t$ matrix of latent representations, H, such that $Z \approx CH$.

The key assumption of multi-view learning is that each of the two views, $\mathbf{x}_j$ and $\mathbf{y}_j$, is conditionally independent given the shared latent representation, $\mathbf{h}_j$. Although multi-view data can always be concatenated and treated as a single view, if the conditional independence assumption holds, explicitly representing multiple views enables more accurate identification of the latent representation (as we will see). The classical formulation of multi-view subspace learning is given by *canonical correlation analysis* (CCA), which is typically expressed as the problem of projecting two views so that the correlation between them is maximized [2]. Assuming the data is centered (i.e. $X\mathbf{1} = \mathbf{0}$ and $Y\mathbf{1} = \mathbf{0}$), the sample covariance of X and Y is given by XX'/t and YY'/t respectively. CCA can then be expressed as an optimization over matrix variables

$$\max_{U,V} \operatorname{tr}(U'XY'V) \;\; s.t. \;\; U'XX'U = V'YY'V = I \tag{1}$$

for $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{m \times k}$ [3]. Although this classical formulation (1) does not make the shared latent representation explicit, CCA can be expressed by a generative model: given a latent representation, $\mathbf{h}_j$, the observations $\mathbf{x}_j = A\mathbf{h}_j + \epsilon_j$ and $\mathbf{y}_j = B\mathbf{h}_j + \nu_j$ are generated by a linear mapping plus independent zero mean Gaussian noise, $\epsilon \sim N(\mathbf{0}, \Sigma_x)$, $\nu \sim N(\mathbf{0}, \Sigma_y)$ [7]. In fact, one can show that the classical CCA problem (1) is equivalent to the following multi-view subspace learning problem.

Proposition 1. *Fix k, let $\tilde{Z} = \begin{bmatrix} (XX')^{-1/2}X \\ (YY')^{-1/2}Y \end{bmatrix}$ and*

$$(C, H) \;\; = \;\; \arg\min_{C,H} \|\tilde{Z} - CH\|_F^2, \tag{2}$$

where $C = \left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]$. Then $U = (XX')^{-\frac{1}{2}}A$ and $V = (YY')^{-\frac{1}{2}}B$ provide an optimal solution to (1), implying that $A'A = B'B = I$ is satisfied in the solution to (2).

(The proof is given in Appendix A.) From Proposition 1, one can see how formulation (2) respects the conditional independence of the separate views: given a latent representation $\mathbf{h}_j$, the reconstruction losses on the two views, $\mathbf{x}_j$ and $\mathbf{y}_j$, cannot influence each other, since the reconstruction models A and B are *individually* constrained. By contrast, in single-view subspace learning (i.e. *principal* *components analysis*) A and B are concatenated in the larger variable C, where C as a whole is constrained but A and B are not. A and B must then compete against each other to acquire magnitude to explain their respective "views" given $\mathbf{h}_j$ (i.e. conditional independence is not enforced). Such sharing can be detrimental if the two views really are conditionally independent given $\mathbf{h}_j$.

Despite its elegance, a key limitation of CCA is its restriction to squared loss under a particular normalization. Recently, subspace learning algorithms have been greatly generalized in the single view case by relaxing the $\operatorname{rank}(H) = k$ constraint while imposing a structured regularizer that is a convex relaxation of rank [10–12]. Such a relaxation allows one to incorporate arbitrary convex losses, including robust losses [10], while maintaining tractability.

As mentioned, these relaxations of single-view subspace learning have only recently been proposed for the *multi-view* setting [13, 14]. An extension of these proposals can be achieved by reformulating (2) to first incorporate an arbitrary loss function L that is convex in its first argument, then relaxing the rank constraint by replacing it with a rank-reducing regularizer on H. In particular, we consider the following training problem that extends [14]:

$$\min_{A,B,H} L\left(\left[\begin{array}{c} A \\ B \end{array}\right] H; Z\right) + \alpha\|H\|_{2,1}, \qquad \text{s.t.} \left[\begin{array}{c} A_{:,i} \\ B_{:,i} \end{array}\right] \in \mathcal{C} \text{ for all } i,$$

$$\text{where} \quad \mathcal{C} := \left\{\left[\begin{array}{c} \mathbf{a} \\ \mathbf{b} \end{array}\right] : \|\mathbf{a}\|_2 \leq \beta, \|\mathbf{b}\|_2 \leq \gamma\right\}, \quad C = \left[\begin{array}{c} A \\ B \end{array}\right], \tag{3}$$

and $\|H\|_{2,1} = \sum_i \|H_{i,:}\|_2$ is a matrix block norm. The significance of using the $(2, 1)$-block norm as a regularizer is that it encourages *rows* of H to become sparse, hence reducing the dimensionality of the learned representation [15]. C must be constrained however, otherwise $\|H\|_{2,1}$ can be pushed arbitrarily close to zero simply by re-scaling H/s and Cs ($s > 0$) while preserving the same loss. Unfortunately, (3) is not jointly convex in A, B and H. Thus, the algorithmic approaches proposed by [13, 14] have been restricted to alternating block coordinate descent between components A, B and H, which cannot guarantee a global solution. Our main result is to show that (3) can in fact be solved globally and efficiently for A, B and H, improving on the previous local solutions [13, 14].

3 Reformulation

Our first main contribution is to derive an equivalent but tractable reformulation of (3), followed by an efficient optimization algorithm. Note that (3) can in principle be tackled by a boosting strategy; however, one would have to formulate a difficult weak learning oracle that considers both views simultaneously [16]. Instead, we find that a direct matrix factorization approach of the form developed in [11, 12] is more effective.

To derive our tractable reformulation, we first introduce the change of variable $\hat{Z} = CH$ which allows us to rewrite (3) equivalently as

$$\min_{\hat{Z}} \left\{L(\hat{Z}; Z) + \alpha \min_{\{C: C_{:,i} \in \mathcal{C}\}} \min_{\{H: CH = \hat{Z}\}} \|H\|_{2,1}\right\}. \tag{4}$$

A key step in the derivation is the following characterization of the inner minimization in (4).

Proposition 2. $\min_{\{C: C_{:,i} \in \mathcal{C}\}} \min_{\{H: CH = \hat{Z}\}} \|H\|_{2,1}$ *defines a norm* $|||\cdot|||^*$ *(on* $\hat{Z}$*) whose dual norm is*

$$|||\Gamma||| := \max_{\mathbf{c} \in \mathcal{C}, \|\mathbf{h}\|_2 \leq 1} \mathbf{c}'\Gamma\mathbf{h}.$$

Proof. Let $\lambda_i = \|H_{i,:}\|_2$ be the Euclidean norm of the i-th row of H. Then $H_{i,:} = \lambda_i \tilde{H}_{i,:}$ where $\tilde{H}_{i,:}$ has unit length (if $\lambda_i = 0$, then take $\tilde{H}_{i,:}$ to be any unit vector). Therefore

$$\min_{\{C: C_{:,i} \in \mathcal{C}\}} \min_{\{H: CH = \hat{Z}\}} \|H\|_{2,1} = \min_{\{C, \lambda_i: C_{:,i} \in \mathcal{C}, \lambda_i \geq 0, \hat{Z} = \sum_i \lambda_i C_{:,i} \tilde{H}_{i,:}\}} \sum_i \lambda_i = \min_{\{t \geq 0: \hat{Z} \in t\mathcal{K}\}} t, \tag{5}$$

where $\mathcal{K}$ is the convex hull of the set $\mathcal{G} := \{\mathbf{c}\mathbf{h}' : \mathbf{c} \in \mathcal{C}, \|\mathbf{h}\|_2 = 1\}$. In other words, we seek a rank-one decomposition of $\hat{Z}$, using only elements from $\mathcal{G}$. Since the set $\mathcal{K}$ is convex and symmetric,

(5) is known as a gauge function and defines a norm on $\hat{Z}$ (see *e.g.* [17, Proposition V.3.2.1]). This norm has a dual given by

$$\|\!|\Gamma|\!\| \quad := \quad \max_{Z\in\mathcal{K}} \operatorname{tr}(\Gamma' Z) \quad = \quad \max_{\mathbf{c}\in\mathcal{C}, \|\mathbf{h}\|_2\leq 1} \mathbf{c}'\Gamma\mathbf{h}, \tag{6}$$

where the last equality follows because maximizing any linear function over the convex hull $\mathcal{K}$ of a set $\mathcal{G}$ achieves the same value as maximizing over the set $\mathcal{G}$ itself. ■

Applying Proposition 2 to problem (3) leads to a simpler formulation of the optimization problem.

Lemma 3. (3) $= \min_{\hat{Z}} L(\hat{Z}; Z) + \alpha \max_{\rho\geq 0} \|D_\rho^{-1}\hat{Z}\|_{\mathrm{tr}}$, *where* $D_\rho = \begin{bmatrix} \sqrt{\beta^2+\gamma^2\rho}\, I_n & 0 \\ 0 & \sqrt{\gamma^2+\beta^2/\rho}\, I_m \end{bmatrix}$.

Proof. The lemma is proved by first deriving an explicit form of the norm $\|\!|\cdot|\!\|$ in (6), then deriving its dual norm. The details are given in Appendix B. ■

Unfortunately the inner maximization problem in Lemma 3 is not concave in ρ. However, it is possible to re-parameterize D_ρ to achieve a tractable formulation as follows. First, define a matrix

$$E_\eta \;:=\; D_{\frac{\beta^2(1-\eta)}{\gamma^2\eta}} = \begin{bmatrix} \beta/\sqrt{\eta}\, I_n & 0 \\ 0 & \gamma/\sqrt{1-\eta}\, I_m \end{bmatrix}, \quad \text{such that } D_\rho = E_{\frac{\beta^2}{\gamma^2\rho+\beta^2}}.$$

Note that $\max_{\rho\geq 0} \|D_\rho^{-1}\hat{Z}\| = \max_{0\leq\eta\leq 1} \|E_\eta^{-1}\hat{Z}\|$, with $\rho \geq 0$ corresponding to $0 \leq \eta \leq 1$. The following lemma proves that this re-parameterization yields an efficient computational approach.

Lemma 4. $h(\eta) := \|E_\eta^{-1}\hat{Z}\|_{\mathrm{tr}}$ *is concave in* η *over* $[0, 1]$.

Proof. Expand $h(\eta)$ into $\left\| \begin{bmatrix} \sqrt{\frac{\eta}{\beta^2}}\hat{Z}^X \\ \sqrt{\frac{1-\eta}{\gamma^2}}\hat{Z}^Y \end{bmatrix} \right\|_{\mathrm{tr}} = \operatorname{tr}\left(\sqrt{\frac{\eta}{\beta^2}(\hat{Z}^X)'\hat{Z}^X + \frac{1-\eta}{\gamma^2}(\hat{Z}^Y)'\hat{Z}^Y} \right)$, where $\operatorname{tr}(\sqrt{\cdot})$ means summing the square root of the eigenvalues (i.e. a spectral function). By [18], if a spectral function f is concave on $[0, \infty)$, then $\operatorname{tr}(f(M))$ must be concave on positive semidefinite matrices. The result follows since $\frac{\eta}{\beta^2}(\hat{Z}^X)'\hat{Z}^X + \frac{1-\eta}{\gamma^2}(\hat{Z}^Y)'\hat{Z}^Y$ is positive semi-definite for $\eta \in [0, 1]$ and $f = \sqrt{\cdot}$ is concave on $[0, \infty)$. ■

From Lemmas 3 and 4 we achieve the first main result.

Theorem 5.

$$(3) = \min_{\hat{Z}} L(\hat{Z}; Z) + \alpha \max_{0\leq\eta\leq 1} \|E_\eta^{-1}\hat{Z}\|_{\mathrm{tr}} = \max_{0\leq\eta\leq 1} \min_{\hat{Z}} L(\hat{Z}; Z) + \alpha\|E_\eta^{-1}\hat{Z}\|_{\mathrm{tr}}. \tag{7}$$

Hence (3) is equivalent to a concave-convex maxi-min problem with no local maxima nor minima.

Thus we have achieved a new formulation for multi-view subspace learning that respects conditional independence of the separate views (see discussion in Section 2) while allowing a globally solvable formulation. To the best of our knowledge, this has not previously been achieved in the literature.

4 Efficient Training Procedure

This new formulation for multi-view subspace learning also allows for an efficient algorithmic approach. Before conducting an experimental comparison to other methods, we first develop an efficient implementation. To do so we introduce a further transformation $\hat{Q} = E_\eta^{-1}\hat{Z}$ in (7), which leads to an equivalent but computationally more convenient formulation of (3):

$$(3) \;=\; \max_{0\leq\eta\leq 1} \min_{\hat{Q}} L(E_\eta\hat{Q}; Z) + \alpha\|\hat{Q}\|_{\mathrm{tr}}. \tag{8}$$

Denote $g(\eta) := \min_{\hat{Q}} L(E_\eta\hat{Q}; Z) + \alpha\|\hat{Q}\|_{\mathrm{tr}}$. The transformation does not affect the concavity of the problem with respect to η established in Lemma 4; therefore, (8) remains tractable. The training procedure then consists of two stages: first, solve (8) to recover η and $\hat{Q}$, which allows $\hat{Z} = E_\eta\hat{Q}$ to be computed; then, recover the optimal factors H and C (*i.e.* A and B) from $\hat{Z}$.

Recovering an optimal $\hat{Z}$: The key to efficiently recovering $\hat{Z}$ is to observe that (8) has a convenient form. The concave outer maximization is defined over a *scalar* variable η, hence simple line search can be used to solve the problem, normally requiring at most a dozen evaluations to achieve a small tolerance. Crucially, the inner minimization in $\hat{Q}$ is a standard trace-norm-regularized loss minimization problem, which has been extensively studied in the matrix completion literature [19–21]. By exploiting these algorithms, $g(\eta)$ and its subgradient can both be computed efficiently.

Recovering C and H from $\hat{Z}$: Once $\hat{Z}$ is obtained, we need to recover a C and H that satisfy

$$CH = \hat{Z}, \quad \|H\|_{2,1} = |\!|\!|\hat{Z}|\!|\!|^*, \quad \text{and } C_{:,i} \in \mathcal{C} \text{ for all } i. \tag{9}$$

We exploit recent sparse approximation methods [22, 23] to solve this problem. First, note from (5) that $|\!|\!|\hat{Z}|\!|\!|^* = \min_{\{C,\lambda_i : C_{:i} \in \mathcal{C}, \lambda_i \geq 0,\, \hat{Z} = \sum_i \lambda_i C_{:,i} \tilde{H}_{i,:}\}} \sum_i \lambda_i$, where $\|\tilde{H}_{i,:}\|_2 \leq 1$. Since we already have $|\!|\!|\hat{Z}|\!|\!|^* = \|E_\eta^{-1}\hat{Z}\|_{\mathrm{tr}}$ from the first stage, we can rescale the problem so that $|\!|\!|\hat{Z}|\!|\!|^* = 1$ without loss of generality. In such a case, $\hat{Z} = \sum_i \lambda_i C_{:,i} \tilde{H}_{i,:}$ where $\lambda \geq 0$ and $\sum_i \lambda_i = 1$ (we restore the proper scale to $\tilde{H}$ afterward). So now, $\hat{Z}$ lies in the convex hull of the set $\mathcal{G} := \{\mathbf{c}\mathbf{h}' : \mathbf{c} \in \mathcal{C}, \|\mathbf{h}\|_2 \leq 1\}$ and any expansion of $\hat{Z}$ as a convex combination of the elements in $\mathcal{G}$ is a valid recovery. From this connection, we can now exploit the recent greedy algorithms developed in [22, 23] to solve the recovery problem. In particular, the recovery just needs to solve

$$\min_{K \in \mathrm{conv}\mathcal{G}} f(K), \quad \text{where} \quad f(K) := \|\hat{Z} - K\|_F^2. \tag{10}$$

where conv denotes the convex hull. Note that the optimal value of (10) is 0. The greedy (boosting) algorithm provided by [22, 23] produces a factorization of $\hat{Z}$ into C and H and proceeds as follows:

1. Weak learning step: greedily pick $G_t = \mathbf{c}_t \mathbf{h}_t' \in \operatorname{argmin}_{G \in \mathcal{G}} \langle \nabla f(K_{t-1}), G \rangle$. Note that this step can be computed efficiently with a form of power method iteration (see Appendix C.2).

2. "Totally corrective" step: $\boldsymbol{\mu}^{(t)} = \operatorname*{argmin}_{\boldsymbol{\mu} \geq \mathbf{0}, \sum_i \mu_i = 1} f\left(\sum_{i=1}^t \mu_i G_i\right)$, then $K_t = \sum_{i=1}^t \mu_i^{(t)} G_i$.

This procedure will find a K_t satisfying $\|\hat{Z} - K_t\|_F^2 < \epsilon$ within $O(1/\epsilon)$ iterations [22, 23].

Acceleration: In practice, this procedure can be considerably accelerated via more refined analysis. Recall $\hat{Z}$ is penalized by the dual of the norm in (6). Given $\hat{Z}$, it is not hard to recover its dual variable Γ by exploiting the dual norm relationship: $\Gamma = \operatorname{argmax}_{\Gamma : |\!|\!|\Gamma|\!|\!| \leq 1} \mathrm{tr}(\Gamma' \hat{Z})$. Then given Γ, the following theorem guarantees many bases in $\mathcal{C}$ can be eliminated from the recovery problem (9).

Theorem 6. *(C, H) satisfying $\hat{Z} = CH$ is optimal iff $\|\Gamma' C_{:,i}\| = 1$ and $H_{i,:} = \|H_{i,:}\|_2 C_{:,i}' \Gamma$, $\forall i$.*

Theorem 6 prunes many elements from $\mathcal{G}$ and the weak learning step only needs to consider a proper subset. Interestingly this constrained search can be solved with no increase in the computational complexity. The accelerated boosting generates $\mathbf{c}_t$ in the weak learning step, giving the recovery $C = [\mathbf{c}_1, \ldots, \mathbf{c}_k]$ and $H = \mathrm{diag}(\boldsymbol{\mu}) C' \Gamma$. The rank, k, is implicitly determined by termination of the boosting algorithm. The detailed algorithm and proof of Theorem 6 are given in Appendix C.

5 Comparisons

Below we compare the proposed global learning method, **Multi-view Subspace Learning (MSL)**, against a few benchmark competitors.

Local Multi-view Subspace Learning (LSL) An obvious competitor is to solve (3) by alternating descent over the variables: optimize H with A and B fixed, optimize A with B and H fixed, etc. This is the computational strategy employed by [13, 14]. Since A and B are both constrained and H is regularized by the (2,1)-block norm which is not smooth, we optimized them using the proximal gradient method [24].

Single-view Subspace Learning (SSL) Single view learning can be cast as a relaxation of (3), where the columns of $C = \begin{bmatrix} A \\ B \end{bmatrix}$ are normalized *as a whole*, rather than individually for A and B:

$$\min_{\{H, C : \|C_{:,i}\|_2 \leq \sqrt{\beta^2 + \gamma^2}\}} L(CH; Z) + \alpha \|H\|_{2,1} = \min_{\{\hat{H}, \hat{C} : \|\hat{C}_{:,i}\|_2 \leq 1\}} L(\hat{C}\hat{H}; Z) + \alpha(\beta^2 + \gamma^2)^{-\frac{1}{2}} \|\hat{H}\|_{2,1} \tag{11}$$

$$= \min_{\hat{Z}} L(\hat{Z}; Z) + \alpha(\beta^2 + \gamma^2)^{-\frac{1}{2}} \|\hat{Z}\|_{\mathrm{tr}}. \tag{12}$$

Equation (12) matches the formulation given in [10]. The equality in (11) is by change of variable $C = \sqrt{\beta^2 + \gamma^2}\hat{C}$ and $\hat{H} = \sqrt{\beta^2 + \gamma^2}H$. Equation (12) can be established from the basic results of [11, 12] (or specializing Proposition 2 to the case where $\mathcal{C}$ is the unit Euclidean ball). To solve (12), we used a variant of the boosting algorithm [21] when α is large, due to its effectiveness when the solution has low rank. When α is small, we switch to the alternating direction augmented Lagrangian method (ADAL) [25] which does not enforce low-rank at all iterations. This hybrid choice of solver is also applied to the optimization of $\hat{Q}$ in (8) for MSL. Once an optimal $\hat{Z}$ is achieved, the corresponding C and H can be recovered by an SVD: for $\hat{Z} = U\Sigma V'$, set $C = (\beta^2+\gamma^2)^{\frac{1}{2}}U$ and $H = (\beta^2+\gamma^2)^{-\frac{1}{2}}\Sigma V'$ which satisfies $CH = \hat{Z}$ and $\|H\|_{2,1} = \|\hat{Z}\|_{\text{tr}}$, and so is an optimal solution to (11).

6 Experimental results

Datasets We provide experimental results on two datasets: a synthetic dataset and a face-image dataset. The synthetic dataset is generated as follows. First, we randomly generate a k-by-t_{tr} matrix H_{tr} for training, a k-by-t_{te} matrix H_{te} for testing, and two basis matrices, A (n-by-k) and B (m-by-k), by (*iid*) sampling from a zero-mean unit-variance Gaussian distribution. The columns of A and B are then normalized to ensure that the Euclidean norm of each is 1. Then we set

$$X_{\text{tr}} = AH_{\text{tr}}, \;\; Y_{\text{tr}} = BH_{\text{tr}}, \;\; X_{\text{te}} = AH_{\text{te}}, \;\; Y_{\text{te}} = BH_{\text{te}}.$$

Next, we add noise to these matrices, to obtain $\tilde{X}_{\text{tr}}$, $\tilde{Y}_{\text{tr}}$, $\tilde{X}_{\text{te}}$, $\tilde{Y}_{\text{te}}$. Following [10], we use sparse non-Gaussian noise: 5% of the matrix entries were selected randomly and replaced with a value drawn uniformly from $[-M, M]$, where M is 5 times the maximal absolute entry of the matrices.

The second dataset is based on the Extended Yale Face Database B [26]. It contains grey level face images of 28 human subjects, each with 9 poses and 64 lighting conditions. To construct the dataset, we set the x-view to a fixed lighting (+000E+00) and the y-view to a different fixed lighting (+000E+20). We obtain a pair of views by randomly drawing a subject and a pose (under the two fixed lightings). The underlying assumption is that each lighting has its own set of bases (A and B) and each (person, pose) pair has the same latent representation for the two lighting conditions. All images are down-sampled to 100-by-100, meaning $n = m = 10^4$. We kept one view (x-view) clean and added pixel errors to the second view (y-view). We randomly set 5% of the pixel values to 1, mimicking the noise in practice, *e.g.* occlusions and loss of pixel information from image transfer. The goal is to enable appropriate reconstruction of a noisy image using other views.

Model specification Due to the sparse noise model, we used $L_{1,1}$ loss for L:

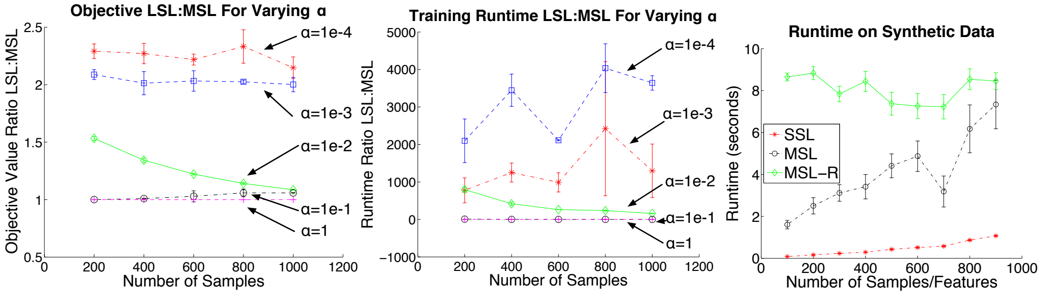
$$L\left(\begin{bmatrix} A \\ B \end{bmatrix} H, Z\right) = \underbrace{\|AH - X\|_{1,1}}_{:=L_1(AH,X)} + \underbrace{\|BH - Y\|_{1,1}}_{:=L_2(BH,Y)}. \tag{13}$$

For computational reasons, we worked on a smoothed version of the $L_{1,1}$ loss [25].

6.1 Comparing optimization quality

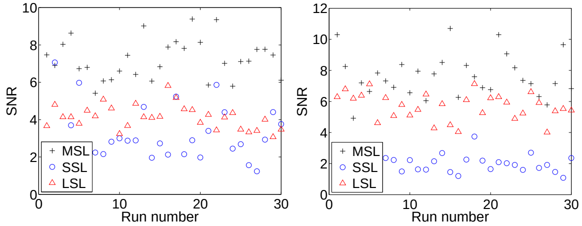
We first compare the optimization performance of MSL (global solver) versus LSL (local solver). Figure 1(a) indicates that MSL consistently obtains a lower objective value, sometimes by a large margin: more than two times lower for $\alpha = 10^{-4}$ and 10^{-3}. Interestingly, as α increases, the difference shrinks. This result suggests that more local minima occur in the higher rank case (a large α increases regularization and decreases the rank of the solution). In Section 6.2, we will see that the lower optimization quality of LSL and the fact that SSL optimizes a less constrained objective both lead to significantly worse denoising performance.

Second, we compare the runtimes of the three algorithms. Figure 1(b) presents runtimes for LSL and MSL for an increasing number of samples. Again, the runtime of LSL is significantly worse for smaller α, as much as 4000x slower; as α increases, the runtimes become similar. This result is likely due to the fact that for small α, the MSL inner optimization is much faster via the ADAL solver (the slowest part of the optimization), whereas LSL still has to slowly iterate over the three variables. They both appear to scale similarly with respect to the number of samples.

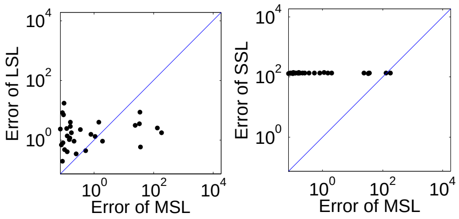


(a) Objectives for LSL:MSL (b) Runtimes for LSL:MSL (c) Runtimes for SSL and MSL

Figure 1: Comparison between LSL and MSL on synthetic datasets with changing α, $n = m = 20$ and 10 repeats. (a) LSL often gets stuck in local minima, with a significantly higher objective than MSL. (b) For small α, LSL is significantly slower than MSL. They scale similarly with the number of samples (c) Runtimes of SSL and MSL for training and recovery with $\alpha = 10^{-3}$. For growing sample size, $n = m = 20$. MSL-R stands for the recovery algorithm. The recovery time for SSL is almost 0, so it is not included.



(a) $t_L = 100$ (b) $t_L = 300$



(a) MSL vs LSL (b) MSL vs SSL

Figure 2: Signal-to-noise ratio of denoising algorithms on synthetic data using recovered models on hold-out views. $n = m = 10$. In (a), we used $t_L = 100$ pairs of views for training A and B and tested on 100 hold-out pairs, with 30 repeated random draws of training and test data. In (b) we used $t_L = 300$. Parameters were set to optimize respective methods.

Figure 3: MSL versus SSL error in synthesizing y-view, over 30 random runs. We set n=m=200, t_L=20, and t_{test}=80. In (a), LSL error is generally above the diagonal line, indicating higher error than MSL. In (b), SSL error is considerably higher than MLS.

For SSL versus MSL, we expect SSL to be faster than MSL because it is a more straightforward optimization: in MSL, each inner optimization of (8) over $\hat{Q}$ (with a fixed η) has the same form as the SSL objective. Figure 1(c), however, illustrates that this difference is not substantial for increasing sample size. Interestingly, the recovery runtime seems independent of dataset size, and is instead likely proportional to the rank of the data. For an increasing number of features, MSL scales well, requiring only about a minute for 1000 features.

6.2 Comparing denoising quality

Next we compare the denoising capabilities of the algorithms on synthetic and face image datasets. There are two denoising approaches. The simplest is to run the algorithm on the noisy $\tilde{Y}_{\text{te}}$, giving the reconstructed $\hat{Y}_{\text{te}}$ as the denoised image. Another approach is to recover the models, A and B, in a training phase. Given a new set of instances, $\tilde{X}_{\text{te}} = \{\tilde{\mathbf{x}}_i\}_{i=1}^s$ and $\tilde{Y}_{\text{te}} = \{\tilde{\mathbf{y}}_i\}_{i=1}^s$, noise in $\tilde{X}_{\text{te}}$ and $\tilde{Y}_{\text{te}}$ can be removed using A and B, *without* re-training. This approach requires first recovering the latent representation, $\hat{H}_{\text{te}} = (\mathbf{h}_1, \ldots, \mathbf{h}_s)$, for $\tilde{X}_{\text{te}}$ and $\tilde{Y}_{\text{te}}$. We use a batch approach for inference:

$$\hat{H}_{\text{te}} = \operatorname*{argmin}_H L_1(AH, \tilde{X}_{\text{te}}) + L_2(BH, \tilde{Y}_{\text{te}}) + \alpha \|H\|_{2,1}. \tag{14}$$

The x-views and y-views are then reconstructed using $\hat{X}_{\text{te}} = A\hat{H}_{\text{te}}$ and $\hat{Y}_{\text{te}} = B\hat{H}_{\text{te}}$. We compared these reconstructions with the clean data, X_{te} and Y_{te}, in terms of the signal-to-noise ratio:

$$SNR(\hat{X}_{\text{te}}, \hat{Y}_{\text{te}}) = \left(\|X_{\text{te}}\|_F^2 + \|Y_{\text{te}}\|_F^2\right) \Big/ \left(\|X_{\text{te}} - \hat{X}_{\text{te}}\|_F^2 + \|Y_{\text{te}} - \hat{Y}_{\text{te}}\|_F^2\right). \tag{15}$$

We present the recovery approach on synthetic data and the direct reconstruction approach on the face dataset. We cross-validated over $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1\}$ according to the highest signal-to-noise ratio on the training data. We set $\gamma = \beta = 1$ because the data is in the $[0, 1]$ interval.

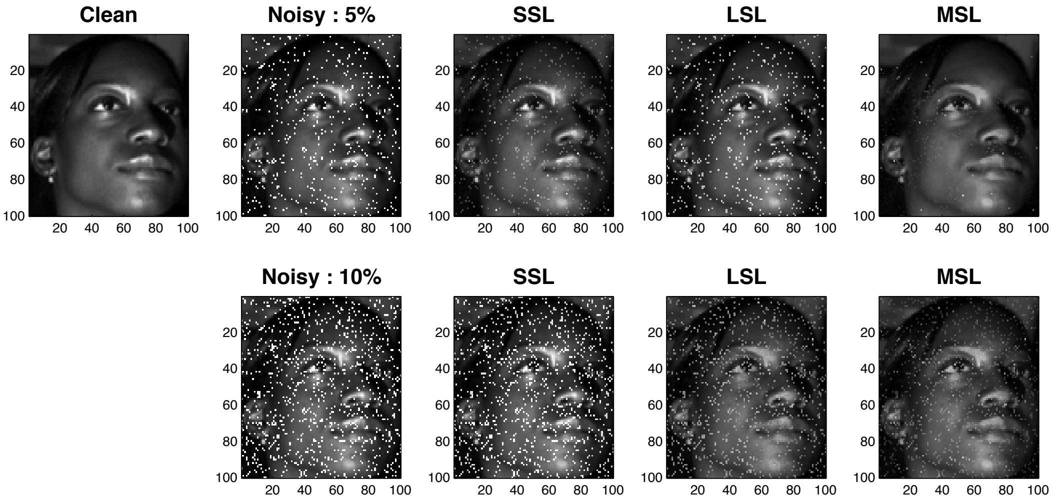


Figure 4: Reconstruction of a noisy image with 5% or 10% noise. LSL performs only slightly worse than MSL for larger noise values: a larger regularization parameter is needed for more noise, resulting in fewer local minima (as discussed in Figure 1). Conversely, SSL performs slightly worse than MSL for 5% noise, but as the noise increases, the advantages of the MSL objective are apparent.

6.2.1 Using Recovered Models for Denoising

Figure 2 presents the signal-to-noise ratio for recovery on synthetic data. MSL produced the highest value of signal-to-noise ratio. The performance of LSL is inferior to MSL, but still better than SSL, corroborating the importance of modelling the data as two views.

6.2.2 Image Denoising

In Figure 4, we can see that MSL outperforms both SSL and LSL on the face image dataset for two noise levels: 5% and 10%. Interestingly, in addition to having on average a 10x higher SNR than SSL for these results, MSL also had significantly different objective values. SSL had larger reconstruction error on the clean x-view (10x higher), lower reconstruction error on the noisy y-view (3x lower) and a higher representation norm (3x higher). Likely, the noisy y-view skewed the representation, due to the joint rather than separate constraint as in the MSL objective.

6.3 Comparing synthesis of views

In image synthesis, the latent representation is computed from only one view: $\hat{H}_{\text{te}} = \text{argmin}_H\, L_1(AH, \tilde{X}_{\text{te}}) + \alpha\|H\|_{2,1}$. The y-view is then synthesized: $\hat{Y}_{\text{te}} = B\hat{H}_{\text{te}}$.

Figure 3 shows the synthesis error, $||\hat{Y}_{\text{te}} - Y_{\text{te}}||_F^2$, of MSL, LSL, and SSL over 30 random runs: MSL generally incurs less error than LSL, and SSL incurs much higher error because it is not modelling the conditional independence between views.

7 Conclusion

We provided a convex reformulation of multi-view subspace learning that enables *global* learning, as opposed to previous local formulations. We also developed a new training procedure which reconstructs the data optimally and discovers the latent representations efficiently. Experimental results on synthetic data and image data confirm the effectiveness of our method, which consistently outperformed other approaches in denoising quality. For future work, we are investigating extensions to semi-supervised settings, such as global methods for co-training and co-regularization. It should also be possible to extend our approach to more than two views and incorporate kernels.

Acknowledgements

We thank the reviewers for their helpful comments, in particular, an anonymous reviewer whose suggestions greatly improved the presentation. Research supported by AICML and NSERC.

References

[1] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2010.

[2] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.

[3] T. De Bie, N. Cristianini, and R. Rosipal. Eigenproblems in pattern recognition. In *Handbook of Geometric Computing*, pages 129–170, 2005.

[4] P. Dhillon, D. Foster, and L. Ungar. Multi-view learning of word embeddings via CCA. In *NIPS*, 2011.

[5] C. Lampert and O. Krömer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *ECCV*, 2010.

[6] L. Sigal, R. Memisevic, and D. Fleet. Shared kernel information embedding for discriminative inference. In *CVPR*, 2009.

[7] F. Bach and M. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2006.

[8] C. Archambeau and F. Bach. Sparse probabilistic projections. In *NIPS*, 2008.

[9] J. Viinikanoja, A. Klami, and S. Kaski. Variational Bayesian mixture of robust CCA. In *ECML*, 2010.

[10] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J.ACM*, 58(1): 1–37, 2011.

[11] X. Zhang, Y. Yu, M. White, R. Huang, and D. Schuurmans. Convex sparse coding, subspace learning, and semi-supervised extensions. In *AAAI*, 2011.

[12] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. arXiv:0812.1869v1, 2008.

[13] N. Quadrinto and C. Lampert. Learning multi-view neighborhood preserving projections. In *ICML*, 2011.

[14] Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. In *NIPS*, pages 982–990, 2010.

[15] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[16] D. Bradley and A. Bagnell. Convex coding. In *UAI*, 2009.

[17] J-B Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms, I and II*, volume 305 and 306. Springer-Verlag, 1993.

[18] D. Petz. A survey of trace inequalities. In *Functional Analysis and Operator Theory*, pages 287–298. Banach Center, 2004.

[19] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128:321–353, 2011.

[20] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[21] X. Zhang, Y. Yu, and D. Schuurmans. Accelerated training for matrix-norm regularization: A boosting approach. In *NIPS*, 2012.

[22] A. Tewari, P. Ravikumar, and I. S. Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In *NIPS*, 2011.

[23] X. Yuan and S. Yan. Forward basis selection for sparse approximation over dictionary. In *AISTATS*, 2012.

[24] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[25] D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, to appear.

[26] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, 23:643–660, 2001.

Supplementary Material

A Proof of Proposition 1

To show that (1) and (2) have equivalent solutions we exploit some developments from [27]. Let $N = (XX')^{-\frac{1}{2}}$ and $M = (YY')^{-\frac{1}{2}}$, hence

$$\tilde{Z}\tilde{Z}' = \begin{bmatrix} I & NXY'M \\ MYX'N & I \end{bmatrix}.$$

First consider (1). Its solution can be characterized by the maximal solutions to the generalized eigenvalue problem [3]:

$$\begin{bmatrix} 0 & XY' \\ YX' & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \lambda \begin{bmatrix} XX' & 0 \\ 0 & YY' \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix},$$

which, under the change of variables $\mathbf{u} = N\mathbf{a}$ and $\mathbf{v} = M\mathbf{b}$ and then shifting the eigenvalues by 1, is equivalent to

$$\begin{aligned}
\equiv \quad & \begin{bmatrix} 0 & XY'M \\ YX'N & 0 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \lambda \begin{bmatrix} N^{-1} & 0 \\ 0 & M^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \\
\equiv \quad & \begin{bmatrix} 0 & NXY'M \\ MYX'N & 0 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \lambda \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \\
\equiv \quad & \tilde{Z}\tilde{Z}' \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = (\lambda + 1) \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}
\end{aligned}$$

By setting $\left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]$ to the top k eigenvectors of $\tilde{Z}\tilde{Z}'$ one can show that $U = NA$ and $V = MB$ provides an optimal solution to (1) [3].

By comparison, for (2), an optimal H is given by $H = C^\dagger \tilde{Z}$, where $C^\dagger$ denotes pseudo-inverse. Hence

$$\begin{aligned}
& \min_{C,H} \|\tilde{Z} - CH\|_F^2 = \min_{C} \|(I - CC^\dagger)\tilde{Z}\|_F^2 \\
= & \operatorname{tr}(\tilde{Z}\tilde{Z}') - \max_{\{C:C'C=I\}} \operatorname{tr}(C'\tilde{Z}\tilde{Z}'C).
\end{aligned}$$

Here again the solution is given by the top k eigenvectors of $\tilde{Z}\tilde{Z}'$ [28].[1]

B Proof for Lemma 3

First, observe that

$$\begin{aligned}
(3) = & \min_{\{C:C_{:,i}\in\mathcal{C}\}} \min_{H} L(CH; Z) + \alpha\|H\|_{2,1} = \min_{\hat{Z}} L(\hat{Z}; Z) + \alpha \min_{\{C:C_{:,i}\in\mathcal{C}\}} \min_{\{H:CH=\hat{Z}\}} \|H\|_{2,1} \\
= & \min_{\hat{Z}} L(\hat{Z}; Z) + \alpha |\!|\!|\hat{Z}|\!|\!|^*,
\end{aligned}$$

where the last step follows from Proposition 2.

It only remains to show $|\!|\!|\hat{Z}|\!|\!|^* = \max_{\rho\geq 0} \|D_\rho^{-1}\hat{Z}\|_{\text{tr}}$, which was established in [11]. We reproduce the proof in [11] for the convenience of the reader.

We will use two diagonal matrices, $I^X = \operatorname{diag}([\mathbf{1}_n; \mathbf{0}_m])$ and $I^Y = \operatorname{diag}([\mathbf{0}_n; \mathbf{1}_m])$ such that $I^X + I^Y = I_{m+n}$. Similarly, for $c \in \mathbb{R}^{m+n}$, we use c^X (respectively c^Y) to denote $c_{1:m}$ (respectively $c_{m+1:m+n}$).

The first stage is to prove that the dual norm is characterized by

$$|\!|\!|\Gamma|\!|\!| = \min_{\rho\geq 0} \|D_\rho \Gamma\|_{\text{sp}}. \tag{16}$$

[1] [29] gave a similar but not equivalent formulation to (2), due to the lack of normalization.

where the spectral norm $\|X\|_{\mathrm{sp}} = \sigma_{\max}(X)$ is the dual of the trace norm, $\|X\|_{\mathrm{tr}}$. To this end, recall that

$$\vert\!\vert\!\vert\Gamma\vert\!\vert\!\vert = \max_{\mathbf{c}\in\mathcal{C}, \|\mathbf{h}\|_2\leq 1} \mathbf{c}'\Gamma h = \max_{\mathbf{c}\in\mathcal{C}} \|\mathbf{c}'\Gamma\|_2 = \max_{\{\mathbf{c}:\|\mathbf{c}^X\|_2=\beta,\, \|\mathbf{c}^Y\|_2=\gamma\}} \|\mathbf{c}'\Gamma\|_2$$

giving

$$\vert\!\vert\!\vert\Gamma\vert\!\vert\!\vert^2 = \max_{\{\mathbf{c}:\|\mathbf{c}^X\|_2=\beta,\, \|\mathbf{c}^Y\|_2=\gamma\}} \mathbf{c}'\Gamma\Gamma'\mathbf{c} = \max_{\{\Phi:\Phi\succeq 0,\, \mathrm{tr}(\Phi I^X)\leq\beta^2,\, \mathrm{tr}(\Phi I^Y)\leq\gamma^2\}} \mathrm{tr}(\Phi\Gamma\Gamma'), \tag{17}$$

using the fact that when maximizing a convex function, one of the extreme points in the constraint set $\{\Phi : \Phi\succeq 0, \mathrm{tr}(\Phi I_n)\leq\beta^2, \mathrm{tr}(\Phi I_m)\leq\gamma^2\}$ must be optimal. Furthermore, since the extreme points have rank at most one in this case [30], the rank constraint $\mathrm{rank}(\Phi) = 1$ can be dropped.

Next, form the Lagrangian $L(\Phi; \lambda, \nu, \Lambda) = \mathrm{tr}(\Phi\Gamma\Gamma') + \mathrm{tr}(\Phi\Lambda) + \lambda(\beta^2 - \mathrm{tr}(\Phi I^X)) + \nu(\gamma^2 - \mathrm{tr}(\Phi I^Y))$ where $\lambda \geq 0$, $\nu \geq 0$ and $\Lambda \succeq 0$. Note that the primal variable Φ can be eliminated by formulating the equilibrium condition $\partial L/\partial\Phi = \Gamma\Gamma' + \Lambda - \lambda I^X - \nu I^Y = 0$, which implies $\Gamma\Gamma' - \lambda I^X - \nu I^Y \preceq 0$. Therefore, we achieve the equivalent dual formulation

$$(17) = \min_{\{\lambda,\nu:\lambda\geq 0,\, \nu\geq 0,\, \lambda I^X+\nu I^Y\succeq\Gamma\Gamma'\}} \beta^2\lambda + \gamma^2\nu. \tag{18}$$

Now observe that for $\lambda \geq 0$ and $\nu \geq 0$, the relation $\Gamma\Gamma' \preceq \lambda I^X + \nu I^Y$ holds if and only if $D_{\nu/\lambda}\Gamma\Gamma' D_{\nu/\lambda} \preceq D_{\nu/\lambda}(\lambda I^X + \nu I^Y)D_{\nu/\lambda} = (\beta^2\lambda + \gamma^2\nu)I_{n+m}$, hence

$$(18) = \min_{\{\lambda,\nu:\lambda\geq 0,\, \nu\geq 0,\, \|D_{\nu/\lambda}\Gamma\|_{sp}^2\leq\beta^2\lambda+\gamma^2\nu\}} \beta^2\lambda + \gamma^2\nu \tag{19}$$

The third constraint must be met with equality at the optimum due to continuity, for otherwise we would be able to further decrease the objective, a contradiction to optimality. Note that a standard compactness argument would establish the existence of minimizers. So

$$(19) = \min_{\lambda\geq 0,\nu\geq 0} \|D_{\nu/\lambda}\Gamma\|_{\mathrm{sp}}^2 = \min_{\rho\geq 0} \|D_\rho\Gamma\|_{\mathrm{sp}}^2.$$

Finally, for the second stage, we characterize the target norm by observing that

$$\begin{aligned}
\vert\!\vert\!\vert\hat{Z}\vert\!\vert\!\vert^* &= \max_{\Gamma:\vert\!\vert\!\vert\Gamma\vert\!\vert\!\vert\leq 1} \mathrm{tr}(\Gamma'\hat{Z}) \\
&= \max_{\rho\geq 0} \max_{\Gamma:\|D_\rho\Gamma\|_{\mathrm{sp}}\leq 1} \mathrm{tr}(\Gamma'\hat{Z}) \qquad (20) \\
&= \max_{\rho\geq 0} \max_{\tilde{\Gamma}:\|\tilde{\Gamma}\|_{\mathrm{sp}}\leq 1} \mathrm{tr}(\tilde{\Gamma}' D_\rho^{-1}\hat{Z}) \\
&= \max_{\rho\geq 0} \|D_\rho^{-1}\hat{Z}\|_{\mathrm{tr}}. \qquad (21)
\end{aligned}$$

where (20) uses (16), and (21) exploits the conjugacy of the spectral and trace norms. The lemma follows.

C Proof for Theorem 6 and Details of Recovery

Once an optimal reconstruction $\hat{Z}$ is obtained, we need to recover the optimal factors C and H that satisfy

$$CH = \hat{Z}, \ , \quad \|H\|_{2,1} = \vert\!\vert\!\vert\hat{Z}\vert\!\vert\!\vert^*, \quad \text{and } C_{:,i} \in \mathcal{C} \text{ for all } i. \tag{22}$$

Note that by Proposition 2 and Lemma 3, the recovery problem (22) can be re-expressed as

$$\min_{\{C,H:C_{:,i}\in\mathcal{C}\,\forall i,\, CH=\hat{Z}\}} \|H\|_{2,1} \quad = \quad \max_{\{\Gamma:\vert\!\vert\!\vert\Gamma\vert\!\vert\!\vert\leq 1\}} \mathrm{tr}(\Gamma'\hat{Z}). \tag{23}$$

Our strategy will be to first recover the optimal dual solution Γ given $\hat{Z}$, then use Γ to recover H and C.

First, to recover Γ one can simply trace back from (21) to (20). Let $U\Sigma V'$ be the SVD of $D_\rho^{-1}\hat{Z}$. Then $\tilde{\Gamma} = UV'$ and $\Gamma = D_\rho^{-1}UV'$ automatically satisfies $\vert\!\vert\!\vert\Gamma\vert\!\vert\!\vert = 1$ while achieving the optimal trace in (23) because $\mathrm{tr}(\tilde{\Gamma}' D_\rho^{-1}\hat{Z}) = \mathrm{tr}(\Sigma) = \|D_\rho^{-1}\hat{Z}\|_{\mathrm{tr}}$.

Given such an optimal Γ, we are then able to characterize an optimal solution (C, H). Introduce the set

$$\mathbf{C}(\Gamma) := \arg\max_{\mathbf{c}\in\mathcal{C}} \|\Gamma'\mathbf{c}\| = \left\{ \mathbf{c} = \left[\begin{array}{c} \mathbf{a} \\ \mathbf{b} \end{array} \right] : \|\mathbf{a}\| = \beta, \|\mathbf{b}\| = \gamma, \|\Gamma'\mathbf{c}\| = 1 \right\}. \tag{24}$$

Theorem 6. *For a dual optimal* Γ, (C, H) *solves recovery problem* (22) *if and only if* $C_{:,i} \in \mathbf{C}(\Gamma)$ *and* $H_{i,:} = \|H_{i,:}\|_2 C'_{:,i}\Gamma$, *such that* $CH = \hat{Z}$.

Proof. By (23), if $\hat{Z} = CH$, then

$$|\!|\!|\hat{Z}|\!|\!|^* = \operatorname{tr}(\Gamma'\hat{Z}) = \operatorname{tr}(\Gamma' CH) = \sum_i H_{i,:}\Gamma' C_{:,i}. \tag{25}$$

Note that $\forall C_{:,i} \in \mathcal{C}, \|\Gamma' C_{:,i}\|_2 \leq 1$ since $|\!|\!|\Gamma|\!|\!| \leq 1$ and $H_{i,:}\Gamma' C_{:,i} = \|H_{i,:}\Gamma' C_{:,i}\|_2 \leq \|H_{i,:}\|_2 \|\Gamma' C_{:,i}\|_2 \leq \|H_{i,:}\|_2$. If (C, H) is optimal, then (25) $= \sum_i \|H_{i,:}\|_2$, hence implying $\|\Gamma' C_{:,i}\|_2 = 1$ and $H_{i,:} = \|H_{i,:}\|_2 C'_{:,i}\Gamma$.

On the other hand, if $\|\Gamma' C_{:,i}\|_2 = 1$ and $H_{i,:} = \|H_{i,:}\|_2 C'_{:,i}\Gamma$, then we have $|\!|\!|\hat{Z}|\!|\!|^* = \sum_i \|H_{i,:}\|_2$, implying the optimality of (C, H). ■

Therefore, given Γ, the recovery problem (22) has been reduced to finding a vector $\boldsymbol{\mu}$ and matrix C such that $\boldsymbol{\mu} \geq 0$, $C_{:,i} \in \mathbf{C}(\Gamma)$ for all i, and $C \operatorname{diag}(\boldsymbol{\mu}) C'\Gamma = \hat{Z}$.

Next we demonstrate how to incrementally recover $\boldsymbol{\mu}$ and C. Denote the range of $C \operatorname{diag}(\boldsymbol{\mu}) C'$ by the set

$$\mathbf{S} := \{\textstyle\sum_i \mu_i \mathbf{c}_i \mathbf{c}_i' : \mathbf{c}_i \in \mathbf{C}(\Gamma), \boldsymbol{\mu} \geq 0\}.$$

Note that $\mathbf{S}$ is the conic hull of (possibly infinitely many) rank one matrices $\{\mathbf{c}\mathbf{c}' : \mathbf{c} \in \mathbf{C}(\Gamma)\}$. However, by Carathéodory's theorem [31, §17], any matrix $K \in \mathbf{S}$ can be written as the conic combination of finitely many rank one matrices of the form $\{\mathbf{c}\mathbf{c}' : \mathbf{c} \in \mathbf{C}(\Gamma)\}$. Therefore, conceptually, the recovery problem has been reduced to finding a sparse set of non-negative weights, $\boldsymbol{\mu}$, over the set of feasible basis vectors, $\mathbf{c} \in \mathbf{C}(\Gamma)$.

To find these weights, we use a totally corrective "boosting" procedure [21] that is guaranteed to converge to a feasible solution. Consider the following objective function for boosting

$$f(K) = \|K\Gamma - \hat{Z}\|_F^2, \text{ where } K \in \mathbf{S}.$$

Note that f is clearly a convex function in K with a Lipschitz continuous gradient. Theorem 6 implies that an optimal solution of (22) corresponds precisely to those $K \in \mathbf{S}$ such that $f(K) = 0$. The idea behind totally corrective boosting [21] is to find a minimizer of f (hence optimal solution of (22)) incrementally. After initializing $K_0 = 0$, we iterate between two steps:

1. Weak learning step: find

$$\mathbf{c}_t \in \operatorname*{argmin}_{\mathbf{c}\in\mathbf{C}(\Gamma)} \langle \nabla f(K_{t-1}), \mathbf{c}\mathbf{c}' \rangle = \operatorname*{argmax}_{\mathbf{c}\in\mathbf{C}(\Gamma)} \mathbf{c}' Q \mathbf{c}, \tag{26}$$

where $Q = -\nabla f(K_{t-1}) = 2(\hat{Z} - K_{t-1}\Gamma)\Gamma'$.

2. "Totally corrective" step:

$$\begin{aligned} \boldsymbol{\mu}^{(t)} &= \operatorname*{argmin}_{\boldsymbol{\mu}:\mu_i \geq 0} f\left(\textstyle\sum_{i=1}^t \mu_i \mathbf{c}_i \mathbf{c}_i'\right), \\ K_t &= \textstyle\sum_{i=1}^t \mu_i^{(t)} \mathbf{c}_i \mathbf{c}_i'. \end{aligned} \tag{27}$$

Three key facts can be established about this boosting procedure: (i) each weak learning step can be solved efficiently; (ii) each totally corrective weight update can be solved efficiently; and (iii) $f(K_t) \searrow 0$, hence a feasible solution can be arbitrarily well approximated. (iii) has been proved in [21], while (ii) is immediate because (27) is a standard quadratic program. Only (i) deserves some explanation. We show in the next subsection that $\mathbf{C}(\Gamma)$, defined in (24), can be much simplified, and consequently we give in the last subsection an efficient algorithm for the oracle problem (26) (the idea is similar to the one inherent in the proof of Lemma 3).

C.1 Simplification of $\mathbf{C}(\Gamma)$

Since $\mathbf{C}(\Gamma)$ is the set of optimal solutions to

$$\max_{\mathbf{c}\in\mathcal{C}} \|\Gamma'\mathbf{c}\|, \tag{28}$$

our idea is to first obtain an optimal solution to its dual problem, and then use it to recover the optimal $\mathbf{c}$ via the KKT conditions. In fact, its dual problem has been stated in (18). Once we obtain the optimal ρ in (21) by solving (8), it is straightforward to backtrack and recover the optimal λ and ν in (18). Then by KKT condition [31, §28], $\mathbf{c}$ is an optimal solution to (28) if and only if

$$\left\|\mathbf{c}^X\right\| = \beta, \quad \left\|\mathbf{c}^Y\right\| = \gamma, \tag{29}$$

$$\langle R, \mathbf{cc}'\rangle = \mathbf{0}, \quad \text{where } R = \lambda I^X + \nu I^Y - \Gamma\Gamma' \succeq 0. \tag{30}$$

Since (30) holds iff $\mathbf{c}$ is in the null space of R, we find an *orthonormal* basis $\{\mathbf{n}_1,\ldots,\mathbf{n}_k\}$ for this null space. Assume

$$\mathbf{c} = N\boldsymbol{\alpha}, \quad \text{where} \quad N = [\mathbf{n}_1,\ldots,\mathbf{n}_k] = \begin{bmatrix} N^X \\ N^Y \end{bmatrix}, \ \boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)'. \tag{31}$$

By (29), we have

$$0 = \gamma^2\left\|\mathbf{c}^X\right\|^2 - \beta^2\left\|\mathbf{c}^Y\right\|^2 = \boldsymbol{\alpha}'\left(\gamma^2(N^X)'N^X - \beta^2(N^Y)'N^Y\right)\boldsymbol{\alpha}. \tag{32}$$

The idea is to go through some linear transformations for simplification. Perform eigen-decomposition $U\Sigma U' = \gamma^2(N^X)'N^X - \beta^2(N^Y)'N^Y$, where $\Sigma = \operatorname{diag}(\sigma_1,\ldots,\sigma_k)$, and $U \in \mathbb{R}^{k\times k}$ is orthonormal. Let $\mathbf{v} = U'\boldsymbol{\alpha}$. Then by (31),

$$\mathbf{c} = NU\mathbf{v}, \tag{33}$$

and (32) is satisfied if and only if

$$\mathbf{v}'\Sigma\mathbf{v} = \sum_i \sigma_i v_i^2 = 0. \tag{34}$$

Finally, (29) implies

$$\beta^2 + \gamma^2 = \|\mathbf{c}\|^2 = \mathbf{v}'U'N'NU\mathbf{v} = \mathbf{v}'\mathbf{v}. \tag{35}$$

In summary, by (33) we have

$$\begin{aligned}\mathbf{C}(\Gamma) &= \{NU\mathbf{v} : \mathbf{v} \text{ satisfies (34) and (35)}\} \\ &= \left\{NU\mathbf{v} : \mathbf{v}'\Sigma\mathbf{v} = 0, \ \|\mathbf{v}\|^2 = \beta^2+\gamma^2\right\}. \end{aligned}\tag{36}$$

C.2 Solving the weak oracle problem (26)

The weak oracle needs to solve

$$\max_{\mathbf{c}\in\mathbf{C}(\Gamma)} \mathbf{c}'Q\mathbf{c},$$

where $Q = -\nabla f(K_{t-1}) = 2(\hat{Z} - K_{t-1}\Gamma)\Gamma'$. By (36), this optimization is equivalent to

$$\max_{\mathbf{v}:\mathbf{v}'\Sigma\mathbf{v}=0,\ \|\mathbf{v}\|^2=\beta^2+\gamma^2} \mathbf{v}'T\mathbf{v},$$

where $T = U'N'QNU$. Using the same technique as in the proof of Lemma 3, we have

$$\begin{aligned} & \max_{\mathbf{v}:\mathbf{v}'\mathbf{v}=1,\mathbf{v}'\Sigma\mathbf{v}=0} \mathbf{v}'T\mathbf{v} \\ (\text{let } H = \mathbf{vv}') \ &= \max_{H\succeq\mathbf{0},\operatorname{tr}(H)=1,\operatorname{tr}(\Sigma H)=0} \operatorname{tr}(TH) \\ (\text{Lagrange dual}) \ &= \min_{\tau,\omega:\tau\Sigma+\omega I-T\succeq\mathbf{0}} \omega \\ &= \min_{\tau\in\mathbb{R}} \lambda_{\max}(T-\tau\Sigma), \end{aligned}$$

where $\lambda_{\max}$ stands for the maximum eigenvalue. Since $\lambda_{\max}$ is a convex function over real symmetric matrices, the last line search problem is convex in τ, hence can be solved globally and efficiently.

Given the optimal τ and the optimal objective value ω, the optimal $\mathbf{v}$ can be recovered using a similar trick as in Appendix C.1. Let the null space of $\omega I + \tau\Sigma - T$ be spanned by $\hat{N} = \{\hat{\mathbf{n}}_1,\ldots,\hat{\mathbf{n}}_s\}$. Then find any $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^s$ such that $\mathbf{v} := \hat{N}\hat{\boldsymbol{\alpha}}$ satisfies $\|\mathbf{v}\|^2 = \beta^2+\gamma^2$ and $\mathbf{v}'\Sigma\mathbf{v} = 0$.

Auxiliary References

[27] L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE TPAMI*, 33(1):194–200, 2011.

[28] M. Overton and R. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62:321–357, 1993.

[29] B. Long, P. Yu, and Z. Zhang. A general model for multiple view unsupervised learning. In *ICDM*, 2008.

[30] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23(2):339–358, 1998.

[31] R. Rockafellar. *Convex Analysis*. Princeton U. Press, 1970.