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 $\begin{bmatrix} A \\ B \end{bmatrix}$ 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

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

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

We will use two diagonal matrices, $I^X = \mathrm{diag}([\mathbf{1}_n; \mathbf{0}_m])$ and $I^Y = \mathrm{diag}([\mathbf{0}_n; \mathbf{1}_m])$ 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

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

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

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

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

giving

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

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

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

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

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

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

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

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

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

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 $\|\!|\Gamma|\!\| = 1$ while achieving the optimal trace in (23) because $\mathrm{tr}(\tilde{\Gamma}' D_\rho^{-1}\hat{Z}) = \mathrm{tr}(\Sigma) = \|D_\rho^{-1}\hat{Z}\|_{\rm tr}$.

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

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

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

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

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 $C\,\mathrm{diag}(\boldsymbol{\mu})C'\Gamma = \hat{Z}$.

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

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

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

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

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

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

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

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.