A Proof of Lemma 1

According to the Lemma 4 in Mai & Zhang (2019), we present the following Lemma 1 to solve Eqs. (9) and (10).

Lemma A.1. *Consider the following minimization problem*

$$\min_{\widehat{\mathbf{u}}} \frac{1}{2N}\|\widehat{\mathbf{u}}^{\mathrm{T}}\mathbf{X}-\mathbf{y}\|_2^2+\lambda_u\|\widehat{\mathbf{u}}\|_1$$
$$s.t.\ Var(\widehat{\mathbf{u}}^{\mathrm{T}}\mathbf{X})=\frac{1}{N}\widehat{\mathbf{u}}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{u}}=1, \tag{20}$$

where $\mathbf{X}$ *and* $\mathbf{y}$ *are fixed, and* λ_u *is a hyperparameter that controls sparsity. The solution to Eq. (20) is* $\widehat{\mathbf{u}}^*=(Var(\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}))^{-1/2}\widetilde{\mathbf{u}}^*$*, where* $\widetilde{\mathbf{u}}^*$ *is obtained via*

$$\widetilde{\mathbf{u}}^*=\arg\min_{\widetilde{\mathbf{u}}} \frac{1}{2N}\|\widetilde{\mathbf{u}}^{\mathrm{T}}\mathbf{X}-\mathbf{y}\|_2^2+\lambda_u\|\widetilde{\mathbf{u}}\|_1 \tag{21}$$

Proof. To prove that $\widehat{\mathbf{u}}^*$ is a solution of Eq. (11), it is equivalent to show that $\widehat{\mathbf{u}}^*$ satisfies the unit-variance constraint in Eq. (11). Specifically, we have

$$Var(\widehat{\mathbf{u}}^{*\mathrm{T}}\mathbf{X})=\frac{1}{N}\widehat{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{u}}^* \tag{22}$$

Substituting $\widehat{\mathbf{u}}^*=(Var(\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}))^{-1/2}\widetilde{\mathbf{u}}^*$ into Eq. (22), then yields

$$\begin{aligned}
&Var(\widehat{\mathbf{u}}^{*\mathrm{T}}\mathbf{X})=\frac{1}{N}\widehat{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{u}}^*\\
=&\frac{1}{N}(Var(\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}))^{-1/2}\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{u}}^*(Var(\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}))^{-1/2}\\
=&\frac{1}{N}(Var(\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}))^{-1}\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{u}}^*\\
=&\frac{1}{N}(Var(\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}))^{-1}(N\cdot Var(\widetilde{\mathbf{u}}^{*\mathrm{T}}\mathbf{X}))\\
=&1
\end{aligned}$$

□

B Data Preprocessing

Handwritten Digit (MNIST). MNIST database (LeCun et al., 1998) contains 60000 training and 10000 testing images of size 28 × 28 with labels from '0' to '9'. We choose a subset consists of 1000 images with labels '0', '1', and '2' and follow the same settings as Chen et al. (2021) to conduct our experiments. The goal is to learn correlated representations between the upper and lower halves (two views) of the original images.

Human Face (Yale). The face images are collected from the Yale database (Cai et al., 2007), which contains 165 images of size 32 × 32. For each image, we apply wavelet transformation to generate the corresponding encoded feature image using the *dw2* function in MATLAB. The original and encoded feature tensors are used as two different views.

Brain Network (BP). Brain networks play an important role in understanding brain functions. Bipolar disorder (BP) dataset (Whitfield-Gabrieli & Nieto-Castanon, 2012) is collected from two modalities, *e.g.*, functional magnetic resonance imaging (fMRI), and diffusion tensor imaging (DTI). We follow Liu et al. (2018) to preprocess the imaging data, including realignment, co-registration, normalization and smoothing, and then construct the brain networks from fMRI and DTI based on the Brodmann template, which are treated as two views.

Facial Expression (JAFFE). The JAFFE database (Lyons et al., 2020) contains female facial expressions of seven categories (neutral, happiness, sadness, surprise, anger, disgust and fear), and the number of images for each category is almost the same. We first crop each image to the size of 200 × 180, and then construct dataset of different sizes and orders. Specifically, we use the cropped image as the first view, and its 3D Gabor features as the second view.

Gait Sequence (Gait32). The Gait32 dataset (Lu et al., 2008) contains 731 video sequences with 71 subjects designed for human identification. The size of each gait video is 32 × 22 × 10, which can be naturally represented by a third-order tensor with the column, row, and time mode. We calculate optical flow and obtain two 3rd-order tensors, which are used as two views (Wang et al., 2016).