Appendix A for EigenGAN

In Sec. 1-8, we derive the analytical result of maximum likelihood estimation (MLE) for the linear case of the proposed EigenGAN. Sec. 8 discusses the MLE result and the relation among the linear EigenGAN, the Principal Component Analysis (PCA) [1], and the Probabilistic PCA [2].

1. The Likelihood

The linear EigenGAN relates a d-dimension observation vector $\mathbf{x}$ to a corresponding q-dimension ($q \leq d$) latent variables $\mathbf{z}$ by an affine transform $\mathbf{UL}$ and a translation $\boldsymbol{\mu}$, which is formulated as

$$\mathbf{x} = \mathbf{ULz} + \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}, \tag{1}$$

with constraints:

$$\mathbf{z} \sim \mathcal{N}_q\left(\mathbf{0}, \mathbf{I}\right), \tag{2}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{I}\right), \text{ independent of } \mathbf{z}, \tag{3}$$

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}, \mathbf{U} \text{ is of size } d \times q, \tag{4}$$

$$\mathbf{L} \text{ is a } q \times q \text{ diagonal matrix}. \tag{5}$$

The noise vector $\boldsymbol{\epsilon}$ in Eq. (1) is introduced to compensate the missing energy (missing rank) since the rank of the latent variables is no more than the rank of the observation ($q \leq d$). According to Eq. (1)-(3), the probability density function of $\mathbf{x}$ is

$$p_{\mathbf{x}}\left(\mathbf{x}\right) = \frac{1}{\sqrt{\left(2\pi\right)^d \left|\mathbf{C}\right|}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathrm{T}} \mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right), \tag{6}$$

where

$$\mathbf{C} = \mathbf{UL}^2\mathbf{U}^{\mathrm{T}} + \sigma^2\mathbf{I}, \tag{7}$$

$$\mathbf{C}^{-1} = \mathbf{UMU}^{\mathrm{T}} + \sigma^{-2}\mathbf{I}, \tag{8}$$

$$\mathbf{M} = \left(\mathbf{L}^2 + \sigma^2\mathbf{I}\right)^{-1} - \sigma^{-2}\mathbf{I}. \tag{9}$$

Then for n observations $\{\mathbf{x}\}_{i=1}^{n}$, the log-likelihood is

$$\mathcal{L}_1 = -\frac{n}{2}\left\{ d\log 2\pi + \log|\mathbf{C}| + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \right\}. \tag{10}$$

According to Eq. (7), only the square value of $\mathbf{L}$ affects the probability density function, therefore we can assume the elements of $\mathbf{L}$ to be non-negative. Further, for convenience of the following analysis, without loss of generality, we organize $\mathbf{L}$ by grouping and sorting it by the value of the diagonal elements:

$$\mathbf{L} = \mathrm{diag}\left(l_1\mathbf{I}_{d_1}, l_2\mathbf{I}_{d_2}, \cdots, l_p\mathbf{I}_{d_p}\right), \tag{11}$$

where $l_1 > l_2 > \cdots > l_p \geq 0$; $\mathbf{I}_{d_j}$ denotes a $d_j \times d_j$ identity matrix, $d_j \neq 0$, and $d_1 + d_2 + \cdots + d_p = q$. According to Eq. (9) and (11), $\mathbf{M}$ also has a grouped form:

$$\mathbf{M} = \mathrm{diag}\left(\left(\left(l_1^2+\sigma^2\right)^{-1} - \sigma^{-2}\right)\mathbf{I}_{d_1}, \cdots, \left(\left(l_p^2+\sigma^2\right)^{-1} - \sigma^{-2}\right)\mathbf{I}_{d_p}\right). \tag{12}$$

And we can also define a block form of $\mathbf{U}$ accordingly:

$$\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_p], \tag{13}$$

where $\mathbf{U}_i$ is of size $d \times d_i$.

2. Determination of $\boldsymbol{\mu}$

The partial derivative of the log-likelihood $\mathcal{L}_1$ (10) with respect to $\boldsymbol{\mu}$ is

$$\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n}\mathbf{C}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}). \tag{14}$$

Then the stationary point is

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \bar{\mathbf{x}}. \tag{15}$$

Since $\mathcal{L}_1$ is a concave function of $\boldsymbol{\mu}$, the above stationary point is also the global maximum point.

3. Determination of U: Part (1)

Substituting Eq. (15) into the log-likelihood $\mathcal{L}_1$ (10), we obtain a new objective:

$$\mathcal{L}_2 = -\frac{n}{2}\left\{d\log 2\pi + \log|\mathbf{C}| + \mathrm{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)\right\}, \tag{16}$$

where

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{\mathrm{T}}, \tag{17}$$

i.e., the covariance matrix of the data. According to Eq. (4), the maximization of $\mathcal{L}_2$ (16) with respect to $\mathbf{U}$ is a constrained optimization as below:

$$\begin{aligned} \max_{\mathbf{U}} \quad & \mathcal{L}_2 \\ \text{subject to} \quad & \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} \end{aligned}$$

Introducing the Lagrange multiplier $\mathbf{H}$, the Lagrangian function is

$$\begin{aligned} \mathcal{L}_{\mathbf{U}} &= \mathcal{L}_2 + \mathrm{tr}\left(\mathbf{H}^{\mathrm{T}}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U} - \mathbf{I}\right)\right) \\ &= -\frac{n}{2}\left\{d\log 2\pi + \log|\mathbf{C}| + \mathrm{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)\right\} + \mathrm{tr}\left(\mathbf{H}^{\mathrm{T}}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U} - \mathbf{I}\right)\right). \end{aligned} \tag{18}$$

Then the partial derivative of $\mathcal{L}_{\mathbf{U}}$ with respect to $\mathbf{U}$ is

$$\frac{\partial \mathcal{L}_{\mathbf{U}}}{\partial \mathbf{U}} = -n\left\{\mathbf{U}\left(\left(\mathbf{L}^2 + \sigma^2\mathbf{I}\right)^{-1}\mathbf{L}^2 + \frac{\mathbf{H}+\mathbf{H}^{\mathrm{T}}}{2}\right) + \mathbf{SUM}\right\}. \tag{19}$$

At the stationary point,

$$\mathbf{SUM} = -\mathbf{U}\left(\left(\mathbf{L}^2 + \sigma^2\mathbf{I}\right)^{-1}\mathbf{L}^2 + \frac{\mathbf{H}+\mathbf{H}^{\mathrm{T}}}{2}\right). \tag{20}$$

Left multiplying the above equation by $\mathbf{U}^{\mathrm{T}}$ and using $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, we obtain

$$\mathbf{U}^{\mathrm{T}}\mathbf{SUM} = -\left(\mathbf{L}^2 + \sigma^2\mathbf{I}\right)^{-1}\mathbf{L}^2 - \frac{\mathbf{H}+\mathbf{H}^{\mathrm{T}}}{2}. \tag{21}$$

The right-hand side of above equation is a symmetric matrix, therefore the left-hand side $\mathbf{U}^{\mathrm{T}}\mathbf{SUM}$ is also symmetric. Furthermore, since both $\mathbf{U}^{\mathrm{T}}\mathbf{SU}$ and $\mathbf{M}$ are also symmetric, to satisfy the symmetry of $\mathbf{U}^{\mathrm{T}}\mathbf{SUM}$, according to the form of $\mathbf{M}$ in Eq. (12), $\mathbf{U}^{\mathrm{T}}\mathbf{SU}$ must have a similar block diagonal form:

$$\mathbf{U}^{\mathrm{T}}\mathbf{SU} = \mathrm{diag}\left(\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_p\right) \tag{22}$$

$$= \mathrm{diag}\left(\mathbf{Q}_1^{\mathrm{T}}\mathbf{\Lambda}_1\mathbf{Q}_1, \mathbf{Q}_2^{\mathrm{T}}\mathbf{\Lambda}_2\mathbf{Q}_2, \cdots, \mathbf{Q}_p^{\mathrm{T}}\mathbf{\Lambda}_p\mathbf{Q}_p\right), \tag{23}$$

where $\mathbf{A}_j$ is a $d_j \times d_j$ symmetric matrix, and $\mathbf{Q}_j^{\mathrm{T}}\mathbf{\Lambda}_j\mathbf{Q}_j$ is the eigendecomposition of $\mathbf{A}_j$. Using Eq. (20), (21), and (23), we can derive

$$\mathbf{SUM} = \mathbf{U} \cdot \mathrm{diag}\left(\mathbf{Q}_1^{\mathrm{T}}\mathbf{\Lambda}_1\mathbf{Q}_1, \mathbf{Q}_2^{\mathrm{T}}\mathbf{\Lambda}_2\mathbf{Q}_2, \cdots, \mathbf{Q}_p^{\mathrm{T}}\mathbf{\Lambda}_p\mathbf{Q}_p\right) \cdot \mathbf{M}. \tag{24}$$

Substituting Eq. (12) and (13) into Eq. (24), we obtain

$$\left(\left(l_j^2+\sigma^2\right)^{-1} - \sigma^{-2}\right)\mathbf{S}\mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}} = \left(\left(l_j^2+\sigma^2\right)^{-1} - \sigma^{-2}\right)\mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}}\mathbf{\Lambda}_j \tag{25}$$

$$\Longrightarrow \qquad \mathbf{S}\mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}} = \mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}}\mathbf{\Lambda}_j, \quad j = 1, 2, \cdots, p', \tag{26}$$

where

$$p' = \begin{cases} p, & l_p > 0, \\ p-1, & l_p = 0. \end{cases} \tag{27}$$

Eq. (26) tells us that, the columns of $\mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}}$ are eigenvectors of $\mathbf{S}$, and the diagonal elements of $\mathbf{\Lambda}_j$ are the corresponding eigenvalues. Further, since $\left(\mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}}\right)^{\mathrm{T}}\mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}} = \mathbf{I}$, these eigenvectors are orthonormal. Let $\mathbf{V}_j = \mathbf{U}_j\mathbf{Q}_j^{\mathrm{T}}$, we obtain the stationary point:

$$\mathbf{U}_j = \mathbf{V}_j\mathbf{Q}_j, \quad j = 1, 2, \cdots, p', \tag{28}$$

where the columns of $\mathbf{V}_j$ are orthonormal eigenvectors of $\mathbf{S}$ with corresponding eigenvalues as $\mathbf{\Lambda}_j = \mathrm{diag}\left(\lambda_{j1}, \lambda_{j_2}, \cdots, \lambda_{j_{d_j}}\right)$, and $\mathbf{Q}_j$ is an arbitrary orthogonal matrix. Note that if $p' = p-1$, i.e., $l_p = 0$, $\mathbf{U}_p$ is an arbitrary matrix.

4. Determination of $\mathbf{L} = \mathbf{diag}\left(l_1\mathbf{I}_{d_1}, l_2\mathbf{I}_{d_2}, \cdots, l_p\mathbf{I}_{d_p}\right)$

Substituting Eq. (26), Eq. (7)-(8), and Eq. (11)-(12) into $\mathcal{L}_2$ (16) and after some manipulation, a new objective is derived:

$$\begin{aligned}\mathcal{L}_3 = -\frac{n}{2}\Bigg\{ & d\log 2\pi + \sum_{j=1}^{p'}\left[d_j\log\left(l_j^2+\sigma^2\right) + \left(l_j^2+\sigma^2\right)^{-1}\mathrm{tr}\left(\mathbf{\Lambda}_j\right)\right] \\ & + (d-q')\log\sigma^2 + \sigma^{-2}\left(\mathrm{tr}\left(\mathbf{S}\right) - \sum_{j=1}^{p'}\mathrm{tr}\left(\mathbf{\Lambda}_j\right)\right)\Bigg\},\end{aligned} \tag{29}$$

where

$$q' = \sum_{j=1}^{p'} d_j. \tag{30}$$

Then the partial derivative of $\mathcal{L}_3$ with respect to l_j is

$$\frac{\partial \mathcal{L}_3}{\partial l_j} = -n \left\{ \frac{d_j l_j}{l_j^2 + \sigma^2} - \frac{l_j \text{tr}\left(\mathbf{\Lambda}_j\right)}{(l_j^2 + \sigma^2)^2} \right\}, \qquad j = 1, 2, \cdots, p', \tag{31}$$

and the stationary point is

$$l_j^2 = \frac{\text{tr}\left(\mathbf{\Lambda}_j\right)}{d_j} - \sigma^2, \qquad j = 1, 2, \cdots, p'. \tag{32}$$

5. Determination of σ

Substituting Eq. (32) into $\mathcal{L}_3$ (29), we obtain a new objective:

$$\begin{aligned} \mathcal{L}_4 = -\frac{n}{2} \Bigg\{ & d \log 2\pi + \sum_{j=1}^{p'} \left[d_j \log \frac{\text{tr}\left(\mathbf{\Lambda}_j\right)}{d_j} + d_j \right] \\ & + (d - q') \log \sigma^2 + \sigma^{-2} \left(\text{tr}\left(\mathbf{S}\right) - \sum_{j=1}^{p'} \text{tr}\left(\mathbf{\Lambda}_j\right) \right) \Bigg\}. \end{aligned} \tag{33}$$

Then the partial derivative of $\mathcal{L}_4$ with respect to σ is

$$\frac{\partial \mathcal{L}_4}{\partial \sigma} = -n \left\{ \frac{d - q'}{\sigma} - \frac{1}{\sigma^3} \left(\text{tr}\left(\mathbf{S}\right) - \sum_{j=1}^{p'} \text{tr}\left(\mathbf{\Lambda}_j\right) \right) \right\}, \tag{34}$$

and the stationary point is

$$\sigma^2 = \frac{\text{tr}\left(\mathbf{S}\right) - \sum_{j=1}^{p'} \text{tr}\left(\mathbf{\Lambda}_j\right)}{d - q'}. \tag{35}$$

6. Determination of $\mathbf{\Lambda}_j$

Substituting Eq. (35) into $\mathcal{L}_4$ (33), we obtain a new objective:

$$\mathcal{L}_5 = -\frac{n}{2} \left\{ d \log 2\pi + \sum_{j=1}^{p'} d_j \log \frac{\text{tr}\left(\mathbf{\Lambda}_j\right)}{d_j} + (d - q') \log \frac{\text{tr}\left(\mathbf{S}\right) - \sum_{j=1}^{p'} \text{tr}\left(\mathbf{\Lambda}_j\right)}{d - q'} + d \right\}. \tag{36}$$

According to Sec. 3, the diagonal elements of $\mathbf{\Lambda}_j = \text{diag}\left(\lambda_{j1}, \lambda_{j_2}, \cdots, \lambda_{j_{d_j}}\right), j = 1, \cdots, p'$ are the eigenvalues of $\mathbf{S}$, therefore the problem here is to select the suitable eigenvalues from $\mathbf{S}$ and separate them into different $\mathbf{\Lambda}_j$s to maximize $\mathcal{L}_5$ (36). Using Jensen's inequality:

$$\log \frac{\text{tr}\left(\mathbf{\Lambda}_j\right)}{d_j} = \log \frac{\lambda_{j_1} + \cdots + \lambda_{j_{d_j}}}{d_j} \geq \frac{\log \lambda_{j_1} + \cdots + \log \lambda_{j_{d_j}}}{d_j}, \tag{37}$$

and the equality holds if and only if $\lambda_{j_1} = \cdots = \lambda_{j_{d_j}}$. That means, no matter how we select the eigenvalues, only grouping them by the same values can maximize $\mathcal{L}_5$ (36). Therefore the optimal grouping is

$$\begin{aligned} \mathbf{\Lambda}_j &= \text{diag}\left(\lambda_{j1}, \lambda_{j_2}, \cdots, \lambda_{j_{d_j}}\right) \\ &= \text{diag}\left(\lambda_j, \lambda_j, \cdots, \lambda_j\right) \\ &= \lambda_j \mathbf{I}_{d_j}, \end{aligned} \tag{38}$$

where λ_j is an eigenvalue of $\mathbf{S}$ whose algebraic multiplicity $\geq d_j$.

Now, the left problem is to select the eigenvalues $\lambda_j, j = 1, \cdots, p'$. Substituting Eq. (38) into $\mathcal{L}_5$ (36), we obtain

$$\begin{aligned} \mathcal{L}_6 &= -\frac{n}{2}\left\{ d\log 2\pi + \sum_{j=1}^{p'} d_j \log \lambda_j + (d-q')\log \frac{\text{tr}\left(\mathbf{S}\right) - \sum_{j=1}^{p'} \text{tr}\left(\mathbf{\Lambda}_j\right)}{d-q'} + d \right\} \\ &= -\frac{n}{2}\left\{ d\log 2\pi + \text{tr}\left(\log \mathbf{S}\right) - \sum_{i=q'+1}^{d} \log \gamma_i + (d-q')\log \frac{\sum_{i=q'+1}^{d} \gamma_i}{d-q'} + d \right\} \\ &= -\frac{n}{2}\left\{ d\log 2\pi + \text{tr}\left(\log \mathbf{S}\right) - (d-q')\left(\frac{\sum_{i=q'+1}^{d} \log \gamma_i}{d-q'} - \log \frac{\sum_{i=q'+1}^{d} \gamma_i}{d-q'} \right) + d \right\}, \end{aligned} \tag{39}$$

where $q' = d_1 + d_2 + \cdots + d_{p'}$ and $\gamma_i, i = q'+1, \cdots, d$ are the rest eigenvalues not

been selected. Maximizing $\mathcal{L}_6$ (39) requires maximizing

$$\mathcal{F} = \frac{\sum_{i=q'+1}^{d} \log \gamma_i}{d-q'} - \log \frac{\sum_{i=q'+1}^{d} \gamma_i}{d-q'}, \tag{40}$$

which only requires $\gamma_i, i = q'+1, \cdots, d$ to be adjacent in ordered eigenvalues. However according to Eq. (32), we need $\lambda_j > \sigma^2, j = 1, \cdots, p'$, and then from Eq. (35), the only choice to maximize $\mathcal{F}$ is to let $\gamma_i, i = q'+1, \cdots, d$ be the $d-q'$ smallest eigenvalues. Meanwhile, lager q' leads to larger $\mathcal{F}$, therefore,

$$p' = p \tag{41}$$

$$q' = q = \sum_{j=1}^{p} d_j \tag{42}$$

7. Determination of U: Part (2)

According to Eq. (28) and Eq. (38), the columns of $\mathbf{V}_j$ are orthonormal eigenvectors of $\mathbf{S}$ corresponding to a same eigenvalue λ_j. Since $\mathbf{Q}_j$ is an arbitrary orthogonal matrix, the column of $\mathbf{U}_j = \mathbf{V}_j\mathbf{Q}_j$ are still orthonormal eigenvectors corresponding to the eigenvalue λ_j.

8. Summary and Discussion

Summarizing the above analysis (Eq. (15), (32), (35), and Sec. 7), the global maximum of the likelihood with respect to the model parameters is

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \tag{43}$$

$$\sigma^2 = \frac{\mathrm{tr}\,(\mathbf{S}) - \mathrm{tr}\,(\boldsymbol{\Lambda})}{d-q}, \tag{44}$$

$$\mathbf{L}^2 = \boldsymbol{\Lambda} - \sigma^2\mathbf{I}, \tag{45}$$

$$\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_q], \tag{46}$$

where the elements of the diagonal matrix $\boldsymbol{\Lambda}$ is the q largest eigenvalues of the data covariance $\mathbf{S}$, and $\mathbf{u}_1, \cdots, \mathbf{u}_q$ are the principal q eigenvectors corresponding to $\boldsymbol{\Lambda}$. As

can be seen, under maximum likelihood estimation, the basis vectors $\mathbf{U}$ of our linear model are exactly the same as that learned by PCA [1]. Moreover, diagonal elements of $\mathbf{L}$ represent the "importance" or "energy" of the corresponding basis vectors, and from Eq. (45), when $\sigma \to 0$, the elements of $\mathbf{L}^2$ approach the q largest eigenvalues. Besides, as shown in Eq. (44), energy (σ^2) of the noise is the average of the discard eigenvalues, which exactly compensates the energy missed by the subspace model.

Our model can be viewed as constrained case of Probabilistic PCA (PPCA) [2]:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon}, \tag{47}$$

whose maximum likelihood estimation is

$$\mathbf{W} = \mathbf{V}\left(\boldsymbol{\Lambda} - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}\mathbf{Q}, \tag{48}$$

where the columns of $\mathbf{V}$ are the principal eigenvectors of the data covariance, $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements are the corresponding eigenvalues, $\mathbf{Q}$ is an arbitrary orthogonal matrix. Therefore, MLE result of PPCA is nondeterministic due to the arbitrary $\mathbf{Q}$. Although $\mathbf{W}$ contains information of the principal eigenvectors, the columns of $\mathbf{W}$ itself do not show explicit property of the orthogonality. Our model (1) restricts $\mathbf{W}$ of PPCA (47) to the special form of $\mathbf{UL}$ where $\mathbf{U}$ has orthonormal columns and $\mathbf{L}$ is diagonal matrix. In consequence, MLE result of our model is deterministic (Eq. (43)-(46)). Moreover, our model can build a linear subspace with the principal eigenvectors as the basis vectors explicitly, which is very suitable for extension to the nonlinear case to learn layer-wise interpretable dimensions, as introduced in the main text.

References

[1] Ian T Jolliffe. *Principal component analysis*. 1986. 1, 8

[2] 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. 1, 8