Augmenting Reconstruction Accuracy in β-VAE Model through Linear Gaussian Framework

1 Introduction

1.1 Main Objectives

The β-VAE model [8], an extension of the VAE framework [14], has gained prominence within the machine learning domain. Within the realm of β-VAE, three fundamental goals come to the fore: accurate data reconstruction, efficient data compression, and disentangled representation.

1. **Accurate Data Reconstruction**: At the core of the β-VAE framework lies the pursuit of precise data reconstruction. Post-training, it encodes input data into a compact latent space and subsequently decodes it to restore the original data. The quality of this reconstructed data is measured against the initial data. Lower reconstruction error signifies more faithful data restoration, indicating the β-VAE's proficiency in preserving crucial information while reducing data dimensionality.

2. **Efficient Data Compression**: The β-VAE aims to efficiently represent input data within the latent space. Its objective is to acquire a succinct yet informative latent representation that encapsulates significant information while minimizing storage and computational demands. The hyperparameter β governs this data compression. Heightened β values intensify information compression in the latent space, capturing salient features while mitigating redundancy. However, excessive compression should be avoided to maintain accurate data reconstruction. Striking the optimal β balance yields a concise, meaningful representation, enabling efficient storage and manipulation of high-dimensional data while retaining critical attributes.

3. **Disentangled Representation**: Disentanglement constitutes another pivotal goal of the β-VAE framework. The model endeavors to learn latent representations where distinct dimensions in the latent space correspond to independent factors of variation present in the input data [12]. This separation in the latent space facilitates the distinct control and management of individual factors, fostering interpretability and manipulation of representations. Effective disentanglement entails the β-VAE's capacity to capture diverse factors of variation autonomously, resulting in efficient data compression and a structured, meaningful data depiction.

By simultaneously exploring and optimizing these three objectives, the β-VAE framework seeks to delicately balance accurate data reconstruction, effective data compression, and meaningful disentanglement. This comprehensive approach not only enhances the model's ability to efficiently represent data but also augments its interpretability and applicability across a spectrum of applications. However, for the scope of this study, our attention remains exclusively on the primary objective of augmenting reconstruction accuracy. This decision stems from the recognition that pursuing other objectives without first achieving satisfactory reconstruction accuracy is impractical.

1.2 Methodology

The research paper is structured as follows:

- In Section 2, we begin with a concise introduction to the β-VAE framework, setting the foundation for our work. We then delve into the motivation behind and formulation of three novel β-VAE-based

problems: γ-VAE with arbitrary positive definite $\boldsymbol{\Sigma_Z}$, γ-VAE with diagonalized positive definite $\boldsymbol{\Sigma_Z}$, and $\gamma\lambda$-VAE with diagonalized positive definite $\boldsymbol{\Sigma_Z}$.

- In Section 3, we establish the linear Gaussian setting used throughout the paper, along with the introduction of relevant notations. Subsequently, we derive closed-form solutions for all three proposed problems utilizing the gradient approach.
- In Section 4, we explore the interplay between the rate-distortion theory and the β-VAE framework. By leveraging this relationship, we employ the alternating iteration Blahut-Arimoto algorithm to unveil optimal solutions for the three proposed β-based problems. We also provide analytical proof that the solutions attained through the gradient approach align with those obtained via the iterative algorithm, reinforcing the robustness of our results.
- In Section 5, we present comprehensive Python algorithms devised for conducting numerical experiments. These algorithms play a pivotal role in determining numerical solutions for the three proposed problems.
- In Sections 6 and 7, we undertake numerical investigations concentrated on the first two γ-VAE problems. Subsequently, we conduct an in-depth comparative analysis to underscore the significance of adopting a diagonalized positive definite $\boldsymbol{\Sigma_Z}$ instead of an arbitrary positive definite $\boldsymbol{\Sigma_Z}$. This strategic choice emerges as instrumental in producing more insightful numerical outcomes and enhancing our ability to control reconstruction accuracy. These findings are expounded upon extensively within Section 8.
- In Section 9, we address limitations inherent in the previous two γ-VAE problems. We introduce a new framework, the $\gamma\lambda$-VAE, characterized by an added hyperparameter. This framework presents a constrained optimization scenario marked by a predefined threshold for reconstruction accuracy. We explore how this additional hyperparameter λ effectively ameliorates current limitations, particularly augmenting our capacity to manipulate reconstruction accuracy.
- In Section 10, we wrap up our exploration by providing a comprehensive summary of the key findings throughout the paper. This section serves as a conclusion, encapsulating the essence of our research into the proposed β-VAE-based problems and the insights they offer.

2 Formulation of Three β-VAE-Based Problems

2.1 β-VAE Framework

Before delving into the investigation of β-VAE-based problems, it is crucial to establish a clear understanding of the foundational framework. In the context of the β-VAE framework, let $\mathbf{Y} \in \mathbb{R}^n$ denote the input data, and $\mathbf{X} \in \mathbb{R}^m$ represent the latent random variable. The β-VAE loss function is then expressed as follows:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})}[\log p_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})] - \beta D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]. \tag{1}$$

Here, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ serve as parameterizations for the encoder and decoder, respectively. The loss function (1) encompasses two pivotal components: the reconstruction term and the regularization term. The reconstruction term ensures accurate data reconstruction and is quantified by the expected log-likelihood of the data given the latent variables. Simultaneously, the regularization term, quantified by the Kullback-Leibler (KL) divergence between the approximate posterior distribution $q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})$ and the prior distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$, fosters the approximation of the prior distribution by the learned latent variables, thereby promoting the emergence of disentangled representations.

The hyperparameter β plays a critical role in the β-VAE framework as a trade-off factor, influencing the balance between disentanglement and reconstruction accuracy [8]. For $\beta > 1$, the regularization term within the loss function (1) gains prominence over the reconstruction term. This emphasis encourages the learned latent variables to closely align with a specific distribution, often chosen as a simple prior distribution such as a standard Gaussian. The primary objective is to constrain the latent space, thereby fostering the

disentanglement of underlying factors of variation within the data. The prevalence of the regularization term leads to a more compact and structured representation, steering the latent variables toward a distribution that facilitates the emergence of disentangled features. Consequently, higher values of β enhance the model's capacity to capture and distinguish distinct factors of variation, ultimately elevating the interpretability of the latent space. Careful tuning of β proves crucial to achieving the desired level of disentanglement while maintaining satisfactory data reconstruction accuracy.

Conversely, when $\beta < 1$, the model allocates greater emphasis to the reconstruction term compared to the regularization term. By diminishing the value of β, the model prioritizes accurate data reconstruction over the regularization of the latent space. The reduced weighting on the regularization term permits the latent variables to encode more information about the input data, potentially resulting in higher-dimensional or less compressed latent representations. The selection of an appropriate β value becomes crucial, as it strikes a delicate balance between data compression and disentanglement. This balance is pivotal for achieving optimal trade-offs and contributes to the β-VAE framework's effectiveness in capturing meaningful and interpretable latent representations aligned with the underlying factors of the input data.

2.2 Two γ-VAE Problems

Our primary goal is to enhance reconstruction accuracy within the β-VAE framework in the context of the linear Gaussian setting. To achieve this objective, we adopt a strategic noise incorporation approach. In this study, we represent the input data as $\boldsymbol{Y}$, the noise added to the reconstructed output $\hat{\boldsymbol{Y}}$ as $\boldsymbol{Z}$, and the noise introduced during encoding as $\boldsymbol{W}$. The introduction of noise into both the input and reconstructed data enhances their authenticity and relevance to real-world scenarios, thereby fortifying the model's robustness. Furthermore, we assume that the encoder and decoder noise $\boldsymbol{W}$ and $\boldsymbol{Z}$ follow Gaussian distributions of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$, respectively.

Acknowledging that a diagonal matrix could facilitate the learning of independent features, while a non-diagonal matrix might capture intricate feature relationships, we leverage these distinct matrix structures for $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ to investigate their impact on reconstruction accuracy. Consequently, the choice of structure for $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ also becomes a consideration. However, its influence on the encoding process is generally less pronounced, as encoders often learn an isotropic Gaussian distribution for latent variables. Thus, across all β-VAE-based problems examined in this study, we consistently opt for an arbitrary positive definite matrix for $\boldsymbol{\Sigma}_{\boldsymbol{W}}$.

Expanding on these noise incorporation strategies within the β-VAE framework, we present formulations for the first two β-VAE variations. These formulations are subsequently extended to a corresponding γ-VAE framework, where $\gamma = \frac{1}{\beta}$. The first variation corresponds to the γ-VAE with an arbitrary positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ accounting for noise added to the reconstructed output, in conjunction with an arbitrary positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ representing the noise introduced during encoding. The second variation corresponds to the γ-VAE employing a diagonalized positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ along with an arbitrary positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{W}}$.

2.3 $\gamma\lambda$-VAE Problem

In a pursuit to further refine our ability to control reconstruction accuracy within the β-VAE framework, we introduce a new framework named the $\gamma\lambda$-VAE, constituting the third variant of the β-VAE. This framework builds upon the foundational principles established in the second problem of the γ-VAE, where $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ adopts a diagonalized structure. We extend this approach to formulate the $\gamma\lambda$-VAE, with the aim of achieving a more nuanced level of control over reconstruction error. This objective is pursued through the incorporation of an additional term involving the hyperparameter λ into the original γ-VAE loss function. This additional term serves as a mechanism to directly modulate the reconstruction error during the model's training process.

In contrast to the prior two scenarios, the formulation of the $\gamma\lambda$-VAE introduces a constrained optimization problem. Specifically, we impose a maximum threshold of 0.05 on the reconstruction error. This constraint ensures that the reconstruction error remains suitably small, enabling the model to generate meaningful and accurate results, while simultaneously affording us precise control over reconstruction accuracy.

3 Closed-Form Solutions using Gradient Approach

3.1 Linear Gaussian Setting

In this paper, we employ three β-VAE-based problems formulated within a linear Gaussian framework to explore reconstruction accuracy. While acknowledging that the linear Gaussian assumption may have limitations in capturing all complexities, it remains a reasonable approximation in many scenarios. The linearity assumption enables closed-form solutions, leading to simplified calculations and explicit formulas. Modeling variables as Gaussian distributions allows us to leverage Gaussian properties and exploit linear transformations, facilitating the development of efficient algorithms. The availability of closed-form solutions enhances our understanding, enabling thorough analysis and providing deeper insights into the model's properties.

Considering a given training dataset $\boldsymbol{Y} \in \mathbb{R}^n$, we assume that the N observations $\{\boldsymbol{Y}_i\}_{i=1}^N$ are drawn from a zero-mean multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_Y})$ with a positive definite covariance matrix $\boldsymbol{\Sigma_Y}$. We postulate the existence of a latent random variable $\boldsymbol{X} \in \mathbb{R}^m$, where $1 \leq m < n$, with a marginal distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$. The choice of a Gaussian distribution with zero mean and identity covariance matrix $\mathbf{I}_m$, denoted by $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, as the prior distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$, is a commonly favored approach in achieving dimensionality reduction. Within this linear Gaussian framework, the encoding process is given by:

$$\boldsymbol{X} = \mathbf{B}\boldsymbol{Y} + \boldsymbol{W}, \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{m \times n}$ is the encoding matrix and $\boldsymbol{W} \in \mathbb{R}^m$ represents the Gaussian noise added during the encoding process. This encoder noise follows a distribution of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_W})$, where $\boldsymbol{\Sigma_W}$ is a positive definite matrix. Furthermore, $\boldsymbol{W}$ is assumed to be independent of the input signal $\boldsymbol{Y}$. Similarly, the decoding process can be expressed as

$$\hat{\boldsymbol{Y}} = \mathbf{A}\hat{\boldsymbol{X}} + \boldsymbol{Z}, \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the decoding matrix, $\hat{\boldsymbol{X}} \in \mathbb{R}^m$ is a sample drawn from the latent variable distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, and $\boldsymbol{Z} \in \mathbb{R}^n$ is the Gaussian noise added during decoding with a distribution of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_Z})$, where $\boldsymbol{\Sigma_Z}$ is a positive definite matrix. Additionally, $\boldsymbol{Z}$ is assumed to be independent of $\hat{\boldsymbol{X}}$. The mean and covariance matrices of the encoder and decoder are described as follows:

$$\begin{aligned} (\boldsymbol{\mu_X}, \boldsymbol{\Sigma_X}) &= (\mathbf{0}, \mathbf{B}\boldsymbol{\Sigma_Y}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}) \\ (\boldsymbol{\mu}_{\hat{\boldsymbol{Y}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}}) &= (\mathbf{0}, \mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma_Z}). \end{aligned} \tag{4}$$

In the given setup, we establish the mutual information of the input data and the latent variable as the mutual information of the encoder, denoted by $I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$, and the mutual information of the latent variable and the reconstructed output as the mutual information of the decoder, denoted by $I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}}; \hat{\boldsymbol{Y}})$. These mutual information measures quantify the amount of information shared between the input data and the latent variable and between the latent variable and the reconstructed output. From the system defined

in (4), the mutual information of the encoder and the decoder can be computed as follows:

$$\begin{aligned}
I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X}) &= h(\boldsymbol{X}) - h(\boldsymbol{X}|\boldsymbol{Y}) \\
&= h(\boldsymbol{X}) - h(\boldsymbol{W}) \\
&= h(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})) - h(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}})) \\
&= \frac{1}{2}\log((2\pi e)^m |\boldsymbol{\Sigma}_{\boldsymbol{X}}|) - \frac{1}{2}\log((2\pi e)^m |\boldsymbol{\Sigma}_{\boldsymbol{W}}|) \\
&= \frac{1}{2}\log\frac{|\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}}|}{|\boldsymbol{\Sigma}_{\boldsymbol{W}}|} \\
I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}}) &= h(\hat{\boldsymbol{Y}}) - h(\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}}) \\
&= h(\hat{\boldsymbol{Y}}) - h(\boldsymbol{Z}) \\
&= h(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}})) - h(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})) \\
&= \frac{1}{2}\log((2\pi e)^n |\boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}}|) - \frac{1}{2}\log((2\pi e)^n |\boldsymbol{\Sigma}_{\boldsymbol{Z}}|) \\
&= \frac{1}{2}\log\frac{|\mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}|}{|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|}.
\end{aligned} \tag{5}$$

Moreover, the reconstruction error, denoted as $\mathcal{L}_{\text{rec}}$, serves as a metric that quantifies the dissimilarity between the input data and the reconstructed data generated by the decoder. It is calculated using the following formula:

$$\mathcal{L}_{\text{rec}} = \frac{\|\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}}\|}{\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\|}, \tag{6}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}}$ represents the covariance matrix of the reconstructed data, and $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ represents the covariance matrix of the input data. The formula calculates the l^2-norm of the difference between the two covariance matrices, normalized by the l^2-norm of the input data's covariance matrix. A smaller value of $\mathcal{L}_{\text{rec}}$ indicates a better reconstruction, indicating a closer match between the covariance matrices of the original data and the reconstructed data.

3.2 Optimal Solutions

The initial two problems utilize the γ-VAE, where $\gamma = \frac{1}{\beta}$, to establish the loss function as a combination of a regularization term and a reconstruction term.

$$\begin{aligned}
\mathcal{L}_{\gamma\text{-VAE}} = &\mathbb{E}_{\boldsymbol{Y}}[D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]] \\
&- \gamma\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})].
\end{aligned} \tag{7}$$

In addressing our third problem, we present the $\gamma\lambda$-VAE framework, designed to enhance our ability to control reconstruction accuracy. Building upon the assumptions outlined in Section 2.3, we introduce an additional term $\lambda\mathbb{E}_{\boldsymbol{Y}}[\|\boldsymbol{Y} - \mathbf{A}\boldsymbol{X}\|^2]$ into the loss function. The rationale behind the selection of this specific λ term is elucidated in Appendix B.5. This extension provides us with a direct means to influence the reconstruction error. The resulting formulation of the loss function is presented as follows:

$$\begin{aligned}
\mathcal{L}_{\gamma\lambda\text{-VAE}} = &\mathbb{E}_{\boldsymbol{Y}}[D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]] \\
&- \gamma\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})] \\
&+ \lambda\mathbb{E}_{\boldsymbol{Y}}[\|\boldsymbol{Y} - \mathbf{A}\boldsymbol{X}\|^2] \\
= &\mathbb{E}_{\boldsymbol{Y}}[D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]] \\
&+ \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[-\gamma\log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x}) + \lambda\|\boldsymbol{Y} - \mathbf{A}\boldsymbol{X}\|^2].
\end{aligned} \tag{8}$$

Within both the γ-VAE and $\gamma\lambda$-VAE frameworks, the loss function $\mathcal{L}$ quantifies the dissimilarity between the input data $\boldsymbol{Y}$ and its corresponding output $\hat{\boldsymbol{Y}}$. Our objective is to minimize this loss function with respect to the model parameters $\{\boldsymbol{\phi}, \boldsymbol{\theta}\}$, thereby obtaining the optimal parameter set $\{\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}\}$.

This section presents closed-form solutions for these frameworks, leveraging the linear Gaussian assumption. Specifically, we reformulate the loss function in terms of $\{\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}\}$ and determine optimal solutions by setting the gradient of the loss function to zero.

3.2.1 Two γ-VAE Problems

Proposition 1. *The regularization term in the γ-VAE loss function (7) is described as follows:*

$$\mathbb{E}_{\boldsymbol{Y}}[D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]]$$

$$= \frac{1}{2}\bigg[\mathrm{Tr}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W}) - \log|\mathbf{\Sigma_W}| - m\bigg].$$

Proof. Appendix B.1.1. □

Proposition 2. *The reconstruction term in the γ-VAE loss function (7) is described as follows:*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})]$$

$$= \frac{1}{2}\bigg(\mathrm{Tr}\bigg[\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_Z}^{-1}\mathbf{AB}\mathbf{\Sigma_Y} - \mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y} - \mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W})\bigg] - n\log(2\pi) - \log|\mathbf{\Sigma_Z}|\bigg).$$

Proof. Appendix B.1.2. □

Corollary 1. *According to **Propositions 1** and **2**, we can represent the γ-VAE loss function (7) using the optimal parameter set $\{\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}\}$, denoted as $\mathbf{\Gamma}_1(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W})$, as follows:*

$$\mathbf{\Gamma}_1(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W})$$

$$\begin{aligned}
&= \frac{1}{2}\bigg[\mathrm{Tr}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W}) - \log|\mathbf{\Sigma_W}| - m\bigg] \\
&\quad - \frac{\gamma}{2}\bigg(\mathrm{Tr}\bigg[\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_Z}^{-1}\mathbf{AB}\mathbf{\Sigma_Y} - \mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y} \\
&\quad - \mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W})\bigg] - n\log(2\pi) - \log|\mathbf{\Sigma_Z}|\bigg).
\end{aligned}$$

Lemma 1. *By setting the gradient of the cost function $\mathbf{\Gamma}_1(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W})$ to zero, we obtain the optimal solution for the γ-VAE loss function (7) as follows:*

$$\begin{aligned}
\mathbf{A} &= (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})^{-1}\mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1} \\
\mathbf{B} &= (\mathbf{I}_m + \mathbf{A}^\mathsf{T}[\mathbf{\Sigma_Z}/\gamma]^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}[\mathbf{\Sigma_Z}/\gamma]^{-1} \\
\mathbf{\Sigma_Z} &= (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})^{-1} \\
\mathbf{\Sigma_W} &= (\mathbf{I}_m + \mathbf{A}^\mathsf{T}[\mathbf{\Sigma_Z}/\gamma]^{-1}\mathbf{A})^{-1}.
\end{aligned}$$

Proof. Appendix B.1.3. □

3.2.2 $\gamma\lambda$-VAE Problem

Notice that the additional term $\mathbb{E}_{\boldsymbol{Y}}[\|\boldsymbol{Y} - \mathbf{A}\boldsymbol{X}\|^2]$ incorporated into the γ-VAE model can be evaluated as

$$\mathbb{E}_{\boldsymbol{Y}}[\|\boldsymbol{Y} - \mathbf{A}\boldsymbol{X}\|^2]$$

$$\begin{aligned}
&= \mathbb{E}_{\boldsymbol{Y}}[\|(\mathbf{I}_n - \mathbf{AB})\boldsymbol{Y} - \mathbf{A}\boldsymbol{W}\|^2] \\
&= \mathbb{E}_{\boldsymbol{Y}}[[(\mathbf{I}_n - \mathbf{AB})\boldsymbol{Y} - \mathbf{A}\boldsymbol{W}]^\mathsf{T}[(\mathbf{I}_n - \mathbf{AB})\boldsymbol{Y} - \mathbf{A}\boldsymbol{W}]] \\
&= \mathbb{E}_{\boldsymbol{Y}}[\mathrm{Tr}[(\mathbf{I}_n - \mathbf{AB})\boldsymbol{Y}\boldsymbol{Y}^\mathsf{T}(\mathbf{I}_n - \mathbf{AB})^\mathsf{T} + \mathbf{A}\boldsymbol{W}\boldsymbol{W}^\mathsf{T}\mathbf{A}^\mathsf{T}]] \\
&= \mathrm{Tr}[(\mathbf{I}_n - \mathbf{AB})\mathbf{\Sigma_Y}(\mathbf{I}_n - \mathbf{AB})^\mathsf{T} + \mathbf{A}\mathbf{\Sigma_W}\mathbf{A}^\mathsf{T}].
\end{aligned}$$

Corollary 2. *Combining with **Corollary 1**, we can represent $\gamma\lambda$-VAE loss function (8) in terms of the optimal set of parameters $\{\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}\}$ using the notation $\boldsymbol{\Gamma}_2(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$. This loss function is expressed as follows:*

$$
\begin{aligned}
&\boldsymbol{\Gamma}_2(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) \\
&= \frac{1}{2}\Big[\operatorname{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{W}}) - \log|\boldsymbol{\Sigma}_{\boldsymbol{W}}| - m\Big] \\
&\quad - \frac{\gamma}{2}\Big(\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\
&\quad - \mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{W}})\Big] - n\log(2\pi) - \log|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|\Big) \\
&\quad + \lambda \operatorname{Tr}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\boldsymbol{\Sigma}_{\boldsymbol{Y}}(\mathbf{I}_n - \mathbf{A}\mathbf{B})^{\mathsf{T}} + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{W}}\mathbf{A}^{\mathsf{T}}].
\end{aligned}
$$

Lemma 2. *Setting the gradient of the cost function $\boldsymbol{\Gamma}_2(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$ to zero yields the optimal solution to $\gamma\lambda$-VAE loss function (8) as follows:*

$$
\begin{aligned}
\mathbf{A} &= (\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1} \\
\mathbf{B} &= [\mathbf{I}_m + \mathbf{A}^{\mathsf{T}}(\gamma\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}]^{-1}\mathbf{A}^{\mathsf{T}}(\gamma\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + 2\lambda\mathbf{I}_n) \\
\boldsymbol{\Sigma}_{\boldsymbol{Z}} &= (\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})^{-1} \\
\boldsymbol{\Sigma}_{\boldsymbol{W}} &= [\mathbf{I}_m + \mathbf{A}^{\mathsf{T}}(\gamma\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}]^{-1}.
\end{aligned}
$$

Proof. Appendix B.1.4. □

4 Closed-Form Solutions using Alternating Iteration Algorithm

4.1 Rate-Distortion Theory

Rate-distortion theory [5] offers valuable insights into the intricate interplay between regularization and reconstruction losses within the β-VAE model. The core of the rate-distortion problem [12] revolves around the pursuit of an encoder that minimizes the extraction of information from data while maintaining a confined reconstruction error. This optimization task is centered on the rate-distortion Lagrangian and a convex curve known as the rate-distortion curve. The β-VAE loss function is closely tied to this problem, requiring a careful balance between compression and reconstruction accuracy through the hyperparameter β. The rate-distortion curve encompasses multiple solutions, each representing distinct trade-offs between compression and reconstruction, thus providing the model with substantial flexibility.

To deepen our comprehension of the relationship between β-VAE and rate-distortion theory, we employ the Blahut-Arimoto algorithm [1, 2]. This algorithm is purpose-built to minimize the mutual information between input and output variables while adhering to a distortion constraint. It efficiently identifies the optimal trade-off between rate and distortion by iteratively adjusting probabilities assigned to output values. This process achieves a delicate balance between low rate and acceptable distortion. In this section, we utilize the Blahut-Arimoto algorithm to derive optimal solutions for the three proposed β-VAE-based problems. We anticipate that the numerical solutions will align with the analytical solutions obtained through the gradient approach described in Section 3. The congruence between the numerical and analytical solutions reinforces our understanding of the correlation between β-VAE and rate-distortion theory, thereby affirming the robustness of our findings.

4.2 Optimal Solutions

Let t represent the iteration step. Given a decoder $\boldsymbol{Y}$ in iteration t, we proceed to update an encoder $\boldsymbol{X}$ in the subsequent iteration $t+1$ using the following equation:

$$
\boldsymbol{X}^{(t+1)} = \mathbf{B}^{(t+1)}\boldsymbol{Y}^{(t)} + \boldsymbol{W}^{(t+1)},
$$

where $\mathbf{B}^{(t+1)} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{W}^{(t+1)} \in \mathbb{R}^m$ be such that $\boldsymbol{W}^{(t+1)} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$, where $\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}$ is a positive definite matrix. Furthermore, $\boldsymbol{W}^{(t+1)}$ and $\boldsymbol{Y}^{(t)}$ are independent. In addition, given the encoder $\hat{\boldsymbol{X}}^{(t+1)}$, the decoder $\hat{\boldsymbol{Y}}^{(t+1)}$ can be updated as follows:

$$\hat{\boldsymbol{Y}}^{(t+1)} = \mathbf{A}^{(t+1)}\hat{\boldsymbol{X}}^{(t+1)} + \boldsymbol{Z}^{(t+1)},$$

where $\hat{\boldsymbol{X}}^{(t+1)} \in \mathbb{R}^m$ satisfying $\hat{\boldsymbol{X}}^{(t+1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, $\mathbf{A}^{(t+1)} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{Z}^{(t+1)} \in \mathbb{R}^n$ be such that $\boldsymbol{Z}^{(t+1)} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right)$, where $\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}$ is a positive definite matrix; and $\boldsymbol{Z}^{(t+1)}$ and $\hat{\boldsymbol{X}}^{(t+1)}$ are independent.

4.2.1 γ-VAE given Arbitrary $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$

Lemma 3. *By employing the Blahut-Arimoto algorithm with a fixed decoder, we can iteratively adjust the encoder until convergence is reached, thereby determining the optimal encoder that minimizes the γ-VAE loss function (7).*

$$q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}^{(t+1)}(\boldsymbol{x}|\boldsymbol{y}) \sim \mathcal{N}\left(\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right),$$

where

$$\mathbf{B}^{(t+1)} = \left[\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right]^{-1}\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)} = \left[\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right]^{-1}.$$

Proof. Appendix B.2.1 □

Lemma 4. *By employing the Blahut-Arimoto algorithm with a fixed encoder, we can iteratively adjust the decoder until convergence is reached, thereby determining the optimal decoder that minimizes the γ-VAE loss function (7).*

$$p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t+1)}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x}) \sim \mathcal{N}\left(\mathbf{A}^{(t+1)}\boldsymbol{x}^{(t+1)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right),$$

where

$$\mathbf{A}^{(t+1)} = \left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)}\right)^{-1}\left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)} = \left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)}\right)^{-1}.$$

Proof. Appendix B.2.2 □

Proposition 3. *The optimal solution for γ-VAE loss function (7) described in **Lemma 1**, which utilizes the gradient approach, can also be obtained through the iterative algorithm, as presented in **Lemmas 3** and **4**. Notably, both of these methods produce equivalent results.*

4.2.2 γ-VAE given Diagonalized $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$

The iterative algorithm for optimizing the γ-VAE loss function, with a diagonalized $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$, resembles the process of optimizing the loss function when $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ is arbitrary. We employ **Lemmas 3** and **4** to iteratively update the encoder and decoder. However, a distinction emerges after updating the decoder $\left(\mathbf{A}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ using **Lemma 4**: we enforce $\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}$ to be a diagonal matrix and scale it by $\frac{1}{\sqrt{\gamma}}$. The matrix $\mathbf{A}^{(t)}$ remains unaltered. Subsequently, employing this decoder $\left(\mathbf{A}^{(t)}$ along with a diagonalized $\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$, we iteratively update the encoder $\left(\mathbf{B}^{(t+1)}, \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$ using the algorithm outlined in **Lemma 3**. It is important to note that we only assume the covariance matrix of the decoder noise $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ to be diagonal and do not make the same assumption for that of the encoder noise $\boldsymbol{\Sigma}_{\boldsymbol{W}}$.

4.2.3 $\gamma\lambda$-VAE given Diagonalized $\mathbf{\Sigma}_{\boldsymbol{Z}}$

To determine the optimal solutions for the $\gamma\lambda$-VAE loss function (8), we will adopt the identical iterative framework detailed in Section 4.2.1. The sole distinction lies in the introduction of a novel decoder noise variable, denoted as $\hat{\boldsymbol{Z}}$. Specifically, we take into account the new decoder

$$\hat{\boldsymbol{Y}}^{(t)} = \mathbf{A}^{(t)}\hat{\boldsymbol{X}}^{(t)} + \hat{\boldsymbol{Z}}^{(t)}.$$

Here, $\hat{\boldsymbol{Z}}^{(t)} \in \mathbb{R}^n$ satisfying $\hat{\boldsymbol{Z}}^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\hat{\boldsymbol{Z}}}^{(t)}\right)$, where $\left[\mathbf{\Sigma}_{\hat{\boldsymbol{Z}}}^{(t)}\right]^{-1} = \gamma\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} + 2\lambda\mathbf{I}_n$. The covariance matrix $\mathbf{\Sigma}_{\hat{\boldsymbol{Z}}}^{(t)}$ is positive definite. Moreover, $\hat{\boldsymbol{Z}}^{(t)}$ and $\hat{\boldsymbol{X}}^{(t)}$ are independent.

We have established in **Proposition 3** that the optimal solution for the γ-VAE loss function, obtained through the gradient approach, can also be attained using the Blahut-Arimoto algorithm. As a consequence, we can formulate the following two lemmas to guide the iterative updates of the encoder and decoder, much akin to **Lemmas 3** and **4**, while adhering to the same proofs. The key alteration lies in substituting the original decoder noise $\boldsymbol{Z}$ with the newly introduced decoder noise $\hat{\boldsymbol{Z}}$. This method empowers us to efficiently optimize the $\gamma\lambda$-VAE loss function (8).

Lemma 5. *By employing the Blahut-Arimoto algorithm with a fixed decoder, we can iteratively adjust the encoder until convergence is reached, thereby determining the optimal encoder that minimizes the $\gamma\lambda$-VAE loss function (8).*

$$q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}^{(t+1)}(\boldsymbol{x}|\boldsymbol{y}) \sim \mathcal{N}\left(\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right),$$

where

$$\mathbf{B}^{(t+1)} = \left[\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left(\gamma\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} + 2\lambda\mathbf{I}_n\right)\mathbf{A}^{(t)}\right]^{-1}\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left(\gamma\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} + 2\lambda\mathbf{I}_n\right)$$

$$\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)} = \left[\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left(\gamma\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} + 2\lambda\mathbf{I}_n\right)\mathbf{A}^{(t)}\right]^{-1}.$$

Lemma 6. *The iterative algorithm for updating the decoder to minimize the $\gamma\lambda$-VAE loss function (8), under the premise of a fixed encoder, aligns precisely with the procedure delineated in **Lemma 4**.*

5 Alternating Iteration Algorithm

5.1 γ-VAE given Arbitrary $\mathbf{\Sigma}_{\boldsymbol{Z}}$

In this section, we present a comprehensive algorithm implemented in Python for conducting numerical experiments aimed at determining the optimal solution $(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})$ that minimizes the γ-VAE loss function. Our approach revolves around utilizing the objective function as the cost function $\mathbf{\Gamma}_1$, as outlined in **Corollary 1**. Additionally, we calculate the mutual information for both the encoder and decoder, as well as the reconstruction error. To accomplish these objectives, we follow these procedural steps:

1. Choose values for 4 inputs:
   - Maximum number of iterations: $maxits$
   - Dimensions of the matrix: n and m
   - Tolerable error: e_{tol}
   - A $n \times n$ positive definite matrix $\mathbf{\Sigma}_{\boldsymbol{Y}}$
2. Initialize variable $flag$ based on 2 following cases:
   (a) If we start with the encoder $\boldsymbol{X}$, then set $flag = 0$.
   (b) If we start with the decoder $\hat{\boldsymbol{Y}}$, then set $flag = 1$.
3. Generate initial inputs for the iteration step:

(a) If $flag = 0$, then do:

- create random initial encoder inputs, including
  - a random $m \times n$ matrix $\mathbf{B}$
  - a random $m \times m$ positive definite covariance matrix $\boldsymbol{\Sigma_W}$
- switch $flag$ to 1.

(b) If $flag = 1$, then do:

- create random initial decoder inputs, including
  - a random $n \times m$ matrix $\mathbf{A}$
  - a random $n \times n$ positive definite covariance matrix $\boldsymbol{\Sigma_Z}$
- switch $flag$ to 0.

4. Initialize iteration counter $t = 1$.

5. Iteration step:

(a) If $flag = 0$,

- first, update the encoder inputs $\boldsymbol{\phi}^{(t+1)} = \left(\mathbf{B}^{(t+1)}, \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$ given decoder $\boldsymbol{\theta}^{(t)} = \left(\mathbf{A}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ using equations in **Lemma 3**.
- second, set $flag = 1$.
- third, compute the resulting loss function of γ-VAE and mutual information of the encoder.
- fourth, check if the cost function $\boldsymbol{\Gamma}_1$ is NaN:
  - if it is, conclude that the algorithm fails to converge and skip to step 8.
- finally, check for convergence after the second iteration:
  i. compute the Frobenius norm of the difference between $\mathbf{B}$ and itself in the previous iteration to get B_norm_diff.
  ii. compute the Frobenius norm of the difference between $\boldsymbol{\Sigma_W}$ and itself in the previous iteration to get $Sigma\_W\_norm\_diff$.
  iii. compute the difference between the cost function $\boldsymbol{\Gamma}_1$ and itself in the previous iteration to get obj_diff.
  iv. check for convergence:
    - if both $B\_norm\_diff, Sigma\_W\_norm\_diff \leq e_{tol}$ and the difference obj_diff is close to 0, conclude that the algorithm converges and skip to step 8.
    - otherwise, move to step 6.

(b) If $flag = 1$,

- first, update the decoder inputs $\boldsymbol{\theta}^{(t)} = \left(\mathbf{A}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ given encoder $\boldsymbol{\phi}^{(t)} = \left(\mathbf{B}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right)$ using equations in **Lemma 4**.
- second, set $flag = 0$.
- third, compute the resulting loss function of γ-VAE, mutual information of the decoder, and reconstruction error.
- fourth, check if the cost function $\boldsymbol{\Gamma}_1$ is NaN:
  - if it is, conclude that the algorithm fails to converge and skip to step 8.
- finally, check for convergence after the second iteration:
  i. compute the Frobenius norm of the difference between $\mathbf{A}$ and itself in the previous iteration to get A_norm_diff.
  ii. compute the Frobenius norm of the difference between $\boldsymbol{\Sigma_Z}$ and itself in the previous iteration to get $Sigma\_Z\_norm\_diff$.
  iii. compute the difference between the cost function $\boldsymbol{\Gamma}_1$ and itself in the previous iteration to get obj_diff.

iv. check for convergence:
   - if both $A\_norm\_diff, Sigma\_Z\_norm\_diff \leq e_{tol}$ and the difference obj_diff is close to 0, conclude that the algorithm converges and skip to step 8.
   - otherwise, move to step 6.

6. Increment iteration counter t by 1.
7. If the iteration counter $t \leq maxits$, then move back to step 5. Otherwise, move to step 8.
8. Compute the values of $\boldsymbol{\Sigma_X}$ and $\boldsymbol{\Sigma_{\hat{Y}}}$.
9. Display results
   - display the optimal solution $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma_Z}, \boldsymbol{\Sigma_W})$.
   - display the minimum value of γ-VAE loss function.
   - display the mutual information of both encoder and decoder.
   - display the values of $\boldsymbol{\Sigma_X}$ and $\boldsymbol{\Sigma_{\hat{Y}}}$.
   - display the value of reconstruction error.
   - move to step 10.
10. Stop.

5.2 γ-VAE given Diagonalized $\boldsymbol{\Sigma_Z}$

The algorithm described in Section 5.1 for obtaining the optimal solution $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma_Z}, \boldsymbol{\Sigma_W})$ for the γ-VAE with arbitrary $\boldsymbol{\Sigma_Z}$ can be applied with a slight modification. In this case, we utilize the diagonalized $\boldsymbol{\Sigma_Z}$ as elaborated in Section 4.2.2.

5.3 $\gamma\lambda$-VAE given Diagonalized $\boldsymbol{\Sigma_Z}$

We can revise the algorithm for determining the optimal solution $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma_Z}, \boldsymbol{\Sigma_W})$ for the $\gamma\lambda$-VAE by building upon the algorithm outlined in Section 5.1 and incorporating a few adjustments. To begin, we adopt the diagonalized $\boldsymbol{\Sigma_Z}$ as explained in Section 4.2.2. Diverging from the preceding scenarios, we now tackle a constrained optimization problem where a maximum reconstruction error threshold of 0.05 is imposed to ensure the meaningfulness of the reconstruction error. The iterative algorithm pursues the optimal solution $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma_Z}, \boldsymbol{\Sigma_W})$ for the $\gamma\lambda$-VAE loss function while satisfying the constraint of a maximum allowable reconstruction error of 5%. This iterative process involves updating the encoder and decoder using **Lemmas 5** and **6**. Subsequently, we compute the associated cost function $\boldsymbol{\Gamma}_2$ as outlined in **Corollary 2** along with the mutual information for both the encoder and decoder, as well as the reconstruction error.

Finally, we carry out a comparative analysis of the reconstruction errors for the three β-VAE-based problems introduced in this study. This analysis aims to assess the influence of framework modifications and underscores the roles played by the hyperparameters γ and λ in governing reconstruction accuracy.

6 Numerical Analysis of γ-VAE given Arbitrary $\boldsymbol{\Sigma_Z}$

In our numerical investigations, our primary objective is to determine the optimal solution for γ-VAE, considering an arbitrary positive definite $\boldsymbol{\Sigma_Z}$. To facilitate this analysis, we execute a small-scale numerical experiment where the input data $\boldsymbol{Y}$ is of dimension $n = 3$, and the latent variable $\boldsymbol{X}$ is of dimension $m = 2$. Within the γ-VAE framework, the process of dimensionality reduction and compression of the input data is realized through the encoding process, which effectively maps the higher-dimensional input data to a lower-dimensional latent space. The size of this latent space directly governs the degree of achieved compression. It is generally understood that a smaller latent space leads to more pronounced compression and greater dimensionality reduction. However, a disproportionately reduced latent space can result in information loss and diminished reconstruction accuracy. In order to strike an optimal balance between compression and

reconstruction accuracy, we select a latent variable dimension of 2. This choice is made to ensure that the dimension is sufficiently small to enable compression, while simultaneously preventing significant reduction in reconstruction accuracy. We then select the covariance matrix $\boldsymbol{\Sigma_Y}$ as follows:

$$\boldsymbol{\Sigma_Y} = \begin{bmatrix} 1 & 0.18 & 0.12 \\ 0.18 & 1 & 0.06 \\ 0.12 & 0.06 & 1 \end{bmatrix}.$$

Next, we consider three cases for γ:

- Case 1: $0 < \gamma < 1$, for which we choose $\gamma = 0.98$.
- Case 2: $\gamma = 1$.
- Case 3: $\gamma > 1$, for which we choose $\gamma = 1.02$.

For each case, we apply the alternating iteration algorithm described in Section 5.1 to find the optimal solution.

6.1 Numerical Solutions

6.1.1 Case 1: $\gamma = 0.98$

The iterative algorithm converges to the trivial solution described in Table 1.

Given $\gamma = 0.98$	Optimal solution
$(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma_Z}, \boldsymbol{\Sigma_W})$	$(\mathbf{0}, \mathbf{0}, \boldsymbol{\Sigma_Y}, \mathbf{I})$
$\boldsymbol{\Sigma_X}$	$\mathbf{I}$
$\boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}}$	$\boldsymbol{\Sigma_Y}$
$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	0
$I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}}; \hat{\boldsymbol{Y}})$	0
Minimum value of $\boldsymbol{\Gamma}_1$	4.1477
Reconstruction error	0

Table 1: Optimal solution for $\gamma = 0.98$ given arbitrary positive definite $\boldsymbol{\Sigma_Z}$.

6.1.2 Case 2: $\gamma = 1$

Notice that multiple optimal solutions exist, and we present two examples of such solutions in Table 2.

6.1.3 Case 3: $\gamma = 1.02$

The algorithm fails to converge due to errors related to NaN or singular matrices. Two examples of invalid optimal solutions are presented in Table 3.

6.2 Analysis of Numerical Solutions

To explore the convergence and uniqueness of optimal solutions for the γ-VAE loss function across varying γ values, we introduce the concept of an *auxiliary decoder*. This concept significantly contributes to comprehending the behavior of decoder noise and its impact on optimal solution convergence under different γ values. Let us begin by considering the decoder of the γ-VAE at iteration t, represented as:

$$\hat{\boldsymbol{Y}}^{(t)} = \mathbf{A}^{(t)} \hat{\boldsymbol{X}}^{(t)} + \boldsymbol{Z}^{(t)},$$

Given $\gamma = 1$	Optimal solution 1	Optimal solution 2
$\mathbf{A}$	$\begin{bmatrix} 0.1137 & 0.1372 \\ 0.3528 & 0.1174 \\ 0.3178 & -0.1509 \end{bmatrix}$	$\begin{bmatrix} 0.2753 & -0.1125 \\ 0.1798 & 0.619 \\ 0.722 & -0.0284 \end{bmatrix}$
$\mathbf{B}$	$\begin{bmatrix} 0.0185 & 0.3317 & 0.2957 \\ 0.1396 & 0.1027 & -0.1738 \end{bmatrix}$	$\begin{bmatrix} 0.1726 & 0.107 & 0.6948 \\ -0.2267 & 0.6623 & -0.041 \end{bmatrix}$
$\boldsymbol{\Sigma_Z}$	$\begin{bmatrix} 0.9682 & 0.1238 & 0.1046 \\ 0.1238 & 0.8618 & -0.0344 \\ 0.1046 & -0.0344 & 0.8762 \end{bmatrix}$	$\begin{bmatrix} 0.9116 & 0.2001 & -0.0819 \\ 0.2001 & 0.5845 & -0.0522 \\ -0.0819 & -0.0522 & 0.478 \end{bmatrix}$
$\boldsymbol{\Sigma_W}$	$\begin{bmatrix} 0.7869 & 0.0031 \\ 0.0031 & 0.9426 \end{bmatrix}$	$\begin{bmatrix} 0.4316 & -0.027 \\ -0.027 & 0.5634 \end{bmatrix}$
$\boldsymbol{\Sigma_X}$	$\mathbf{I}$	$\mathbf{I}$
$\boldsymbol{\Sigma_{\hat{Y}}}$	$\boldsymbol{\Sigma_Y}$	$\boldsymbol{\Sigma_Y}$
$I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X})$	0.1494	0.7085
$I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}})$	0.1494	0.7085
Minimum value of $\boldsymbol{\Gamma}_1$	4.2323	4.2323
Reconstruction error	0	0

Table 2: Optimal solutions for $\gamma = 1$ given arbitrary positive definite $\boldsymbol{\Sigma_Z}$.

To update the encoder $\left(\mathbf{B}^{(t+1)}, \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$ in the γ-VAE, a transformation involving the factor γ is necessary. To enable this transformation, we introduce an auxiliary decoder that incorporates a scaling factor of $\frac{1}{\sqrt{\gamma}}$ in the noise term, resulting in the following expression:

$$\hat{\boldsymbol{Y}}^{(t)} = \mathbf{A}^{(t)}\hat{\boldsymbol{X}}^{(t)} + \frac{\boldsymbol{Z}^{(t)}}{\sqrt{\gamma}}.$$

As a result of this normalization, the covariance matrix of the decoder noise $\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}$ undergoes rescaling by $\frac{1}{\sqrt{\gamma}}$. Hence, during the decoder update process, the following observations can be made:

- For $0 < \gamma < 1$, the introduction of the scaling factor $\frac{1}{\sqrt{\gamma}}$ in the noise term $\boldsymbol{Z}^{(t)}$ within the auxiliary decoder progressively amplifies the influence of the noise $\boldsymbol{Z}$ as iterations proceed. As a consequence, the original relationship between $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$ is eventually diminished, leading to the convergence of $\mathbf{A}$ toward $\mathbf{0}$. This outcome yields an optimal solution characterized by $\mathbf{A} = \mathbf{0}$, $\mathbf{B} = \mathbf{0}$, and $\boldsymbol{\Sigma_W} = \mathbf{I}$. Consequently, the resulting $\boldsymbol{X}$ manifests as an additive Gaussian noise with zero mean and an identity covariance matrix. A detailed analytical proof establishing the unique trivial solution for $0 < \gamma < 1$ will be presented in Appendix B.3.1.

  Given the aforementioned trivial solution, the resulting cost function $\boldsymbol{\Gamma}_1$ can be computed as:

$$\boldsymbol{\Gamma}_1(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma_Z}, \boldsymbol{\Sigma_W}) = \frac{\gamma}{2}\left[\log|\boldsymbol{\Sigma_Y}| + n + n\log(2\pi)\right].$$

Given $\gamma = 1.02$	Optimal solution 1	Optimal solution 2
$\mathbf{A}$	$\begin{bmatrix} 0.0134 & -1.9537 \\ 0.2718 & 2.3634 \\ 0.2009 & 0.8163 \end{bmatrix}$	$\begin{bmatrix} -17.6809 & 26.4601 \\ -141.5379 & -107.5214 \\ -16.8229 & 11.3015 \end{bmatrix}$
$\mathbf{B}$	$\begin{bmatrix} -0.5106 & -1.3895 & 1.6202 \\ 3.4959 & 3.8611 & -0.3946 \end{bmatrix}$	$\begin{bmatrix} -14.8989 & -0.0476 & 9.702 \\ -5.569 & -0.0305 & 3.8112 \end{bmatrix}$
$\boldsymbol{\Sigma_Z}$	$\begin{bmatrix} -1.0331 & 0.9888 & 0.5224 \\ 0.9888 & -0.9463 & -0.5 \\ 0.5224 & -0.5 & -0.2642 \end{bmatrix}$	$\begin{bmatrix} -0.0032 & -0.0756 & -0.0052 \\ -0.0756 & -1.8127 & -0.125 \\ -0.0052 & -0.125 & -0.0086 \end{bmatrix}$
$\boldsymbol{\Sigma_W}$	$\mathbf{0}$	$\mathbf{0}$
$\boldsymbol{\Sigma_X}$	$\begin{bmatrix} 4.6032 & -7.9063 \\ -7.9063 & 31.6311 \end{bmatrix}$	$\begin{bmatrix} 281.6172 & 106.7541 \\ 106.7541 & 40.4939 \end{bmatrix}$
$\boldsymbol{\Sigma_{\hat{Y}}}$	$\begin{bmatrix} 2.7842 & -3.6251 & -1.0698 \\ -3.6251 & 4.7132 & 1.4839 \\ -1.0698 & 1.4839 & 0.4426 \end{bmatrix}$	$\begin{bmatrix} 1012.7491 & -342.5759 & 596.4788 \\ -342.5759 & 31592.0117 & 1165.8005 \\ 596.4788 & 1165.8005 & 410.7258 \end{bmatrix}$
$I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X})$	36.2986	39.4027
$I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}})$	34.0688	38.3848
Minimum value of $\boldsymbol{\Gamma}_1$	NaN	NaN
Reconstruction error	0.916	1553.1746

Table 3: Optimal solutions for $\gamma = 1.02$ given arbitrary positive definite $\boldsymbol{\Sigma_Z}$.

- For $\gamma > 1$, the covariance matrix of the noise progressively diminishes, ultimately tending towards zero. This situation yields an unstable algorithm and results in an invalid numerical solution due to the requirement that both the covariance matrices of the encoder and decoder noise must be positive definite. An analytical proof detailing the manifestation of the singular matrix error arising from the algorithm's failure to converge for $\gamma > 1$ will be presented in Appendix B.3.2.
- For $\gamma = 1$, multiple optimal solutions exist. Nevertheless, the resulting cost function $\boldsymbol{\Gamma}_1$ remains consistent. Additionally, we have conducted numerical experiments that confirm the alignment of the entropy of the input data $H(\boldsymbol{Y})$ with the minimum value of the cost function $\boldsymbol{\Gamma}_1$, as demonstrated in Table 2.

In the considered linear model, the best approximation of the second-order statistics is achieved when $\gamma = 1$, likely due to the simplicity of the model. This finding highlights the importance of considering the model's complexity and architecture when interpreting the impact of the parameter γ.

7 Numerical Analysis of γ-VAE given Diagonalized $\boldsymbol{\Sigma_Z}$

Using the same experimental framework outlined in Section 6, our goal remains to determine the optimal solution for γ-VAE, now with a diagonal covariance matrix $\boldsymbol{\Sigma_Z}$.

7.1 Numerical Solutions

7.1.1 Case 1: $\gamma = 0.98$

In comparison to the situation discussed in Section 6.1.1, the optimal solution for the case where a diagonalized positive definite $\boldsymbol{\Sigma_Z}$ is involved and $\gamma < 1$ is not unique. Two illustrative examples of such optimal solutions are provided in Table 4.

Given $\gamma = 0.98$	Optimal solution 1	Optimal solution 2
$\mathbf{A}$	$\begin{bmatrix} 0.0259 & 0.4801 \\ 0.0183 & 0.3402 \\ 0.0117 & 0.2167 \end{bmatrix}$	$\begin{bmatrix} 0.4462 & 0.1789 \\ 0.3162 & 0.1268 \\ 0.2014 & 0.0807 \end{bmatrix}$
$\mathbf{B}$	$\begin{bmatrix} 0.0224 & 0.0138 & 0.0082 \\ 0.4158 & 0.2563 & 0.1514 \end{bmatrix}$	$\begin{bmatrix} 0.3865 & 0.2382 & 0.1407 \\ 0.1549 & 0.0955 & 0.0564 \end{bmatrix}$
$\boldsymbol{\Sigma_Z}$	$\begin{bmatrix} 0.7689 & 0 & 0 \\ 0 & 0.8839 & 0 \\ 0 & 0 & 0.9529 \end{bmatrix}$	$\begin{bmatrix} 0.7689 & 0 & 0 \\ 0 & 0.8839 & 0 \\ 0 & 0 & 0.9529 \end{bmatrix}$
$\boldsymbol{\Sigma_W}$	$\begin{bmatrix} 0.9991 & -0.0172 \\ -0.0172 & 0.6804 \end{bmatrix}$	$\begin{bmatrix} 0.7238 & -0.1107 \\ -0.1107 & 0.9556 \end{bmatrix}$
$\boldsymbol{\Sigma_X}$	$\mathbf{I}$	$\mathbf{I}$
$\boldsymbol{\Sigma_{\hat{Y}}}$	$\begin{bmatrix} 1 & 0.1638 & 0.1043 \\ 0.1638 & 1 & 0.0739 \\ 0.1043 & 0.0739 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.1638 & 0.1043 \\ 0.1638 & 1 & 0.0739 \\ 0.1043 & 0.0739 & 1 \end{bmatrix}$
$I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X})$	0.1932	0.1932
$I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}})$	0.1965	0.1965
Minimum value of $\boldsymbol{\Gamma}_1$	4.152	4.152
Reconstruction error	0.0245	0.0245

Table 4: Optimal solutions for $\gamma = 0.98$ given diagonalized positive definite $\boldsymbol{\Sigma_Z}$.

7.1.2 Case 2: $\gamma = 1$

Similarly to the scenario involving an arbitrary positive definite $\boldsymbol{\Sigma_Z}$ discussed in Section 6.1.2, the optimal solution for $\gamma = 1$ with a diagonal covariance matrix $\boldsymbol{\Sigma_Z}$ is also not unique. We present two examples of such optimal solutions in Table 5.

Given $\gamma = 1$	Optimal solution 1	Optimal solution 2
$\mathbf{A}$	$\begin{bmatrix} 0.0012 & 0.6816 \\ 0.0759 & 0.2639 \\ 0.1795 & 0.1757 \end{bmatrix}$	$\begin{bmatrix} 0.0894 & 0.6184 \\ 0.0493 & 0.284 \\ 0.5958 & 0.1079 \end{bmatrix}$
$\mathbf{B}$	$\begin{bmatrix} -0.0331 & 0.0711 & 0.1792 \\ 0.6452 & 0.1424 & 0.0898 \end{bmatrix}$	$\begin{bmatrix} 0.0163 & 0.0107 & 0.5932 \\ 0.5832 & 0.1773 & 0.0273 \end{bmatrix}$
$\boldsymbol{\Sigma_Z}$	$\begin{bmatrix} 0.5354 & 0 & 0 \\ 0 & 0.9246 & 0 \\ 0 & 0 & 0.9369 \end{bmatrix}$	$\begin{bmatrix} 0.6096 & 0 & 0 \\ 0 & 0.9169 & 0 \\ 0 & 0 & 0.6334 \end{bmatrix}$
$\boldsymbol{\Sigma_W}$	$\begin{bmatrix} 0.9625 & -0.0277 \\ -0.0277 & 0.5068 \end{bmatrix}$	$\begin{bmatrix} 0.6446 & -0.0771 \\ -0.0771 & 0.5861 \end{bmatrix}$
$\boldsymbol{\Sigma_X}$	$\mathbf{I}$	$\mathbf{I}$
$\boldsymbol{\Sigma_{\hat{Y}}}$	$\boldsymbol{\Sigma_Y}$	$\boldsymbol{\Sigma_Y}$
$I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X})$	0.3597	0.4947
$I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}})$	0.3597	0.4947
Minimum value of $\boldsymbol{\Gamma}_1$	4.2323	4.2323
Reconstruction error	0	0

Table 5: Optimal solutions for $\gamma = 1$ given diagonalized positive definite $\boldsymbol{\Sigma_Z}$.

7.1.3 Case 3: $\gamma = 1.02$

Similar to the case of arbitrary positive definite $\boldsymbol{\Sigma_Z}$ as discussed in Section 6.1.3, the algorithm encounters convergence issues attributed to NaN or singular matrices when a diagonalized positive definite $\boldsymbol{\Sigma_Z}$ is considered, and the corresponding optimal solutions are exhibited in Table 6. An analytical proof for this scenario is presented in Appendix B.4.1.

8 γ-VAE: Arbitrary vs. Diagonalized $\boldsymbol{\Sigma_Z}$

8.1 Plots of Numerical Solutions

The following three plots provide a comparison of numerical solutions for the γ-VAE loss function using both arbitrary and diagonalized positive definite $\boldsymbol{\Sigma_Z}$. The initial plot, depicted in Figure 1, showcases the mutual information in relation to varying γ values, ranging from 0.9 to 1.1 with a step size of 0.01. The second plot, outlined in Figure 2, illustrates the reconstruction error as a function of γ, utilizing the same range and step size. Lastly, the third plot displayed in Figure 3 exhibits the mutual information after 100 algorithm iterations with $\gamma = 1$, considering both arbitrary and diagonalized positive definite $\boldsymbol{\Sigma_Z}$.

8.2 Comparing Numerical Solutions for γ-VAE Problems

As illustrated in Figure 1, a distinct disparity emerges in the mutual information between the intervals $0 < \gamma < 1$ and $\gamma > 1$, irrespective of the configuration of the decoder noise covariance. Notice that

Given $\gamma = 1.02$	Optimal solution 1	Optimal solution 2
$\mathbf{A}$	$\begin{bmatrix} 0.207 & 0.0327 \\ 0.9205 & -0.378 \\ 0.4326 & 0.8961 \end{bmatrix}$	$\begin{bmatrix} 0.5192 & 0.8489 \\ 0.1263 & 0.1327 \\ 0.9051 & -0.4136 \end{bmatrix}$
$\mathbf{B}$	$\begin{bmatrix} 0 & 0.9067 & 0.3825 \\ 0 & -0.4377 & 0.9313 \end{bmatrix}$	$\begin{bmatrix} 0.4207 & 0 & 0.8636 \\ 0.9207 & 0 & -0.5282 \end{bmatrix}$
$\boldsymbol{\Sigma_Z}$	$\begin{bmatrix} 0.9556 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.9661 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$\boldsymbol{\Sigma_W}$	$\mathbf{0}$	$\mathbf{0}$
$\boldsymbol{\Sigma_X}$	$\begin{bmatrix} 1.0099 & 0 \\ 0 & 1.0099 \end{bmatrix}$	$\begin{bmatrix} 1.0099 & 0 \\ 0 & 1.01 \end{bmatrix}$
$\boldsymbol{\Sigma_{\hat{Y}}}$	$\begin{bmatrix} 0.9996 & 0.1782 & 0.1188 \\ 0.1782 & 0.9902 & 0.0594 \\ 0.1188 & 0.0594 & 0.9902 \end{bmatrix}$	$\begin{bmatrix} 0.9902 & 0.1782 & 0.1188 \\ 0.1782 & 0.9997 & 0.0594 \\ 0.1188 & 0.0594 & 0.9901 \end{bmatrix}$
$I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X})$	67.7313	68.52
$I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}})$	67.7214	68.5101
Minimum value of $\boldsymbol{\Gamma}$	-143552238122433.16	-143552238122433.12
Reconstruction error	0.0087	0.009

Table 6: Optimal solutions for $\gamma = 1.02$ given diagonalized positive definite $\boldsymbol{\Sigma_Z}$.

the mutual information exhibits a high sensitivity to even minor fluctuations in the parameter γ. This sensitivity manifests in the sudden surge of mutual information values shortly after γ surpasses a value of 1. Within the domain of $0 < \gamma < 1$, the mutual information remains substantially lower. Conversely, for $\gamma > 1$, the mutual information experiences a significant increase. Notably, the mutual information displays a heightened responsiveness to subtle changes in the parameter γ, evident through the swift escalation of mutual information values during the transition from $\gamma < 1$ to $\gamma > 1$. This phenomenon underscores the necessity for cautious parameter adjustment to prevent unintended oscillations in mutual information outcomes. Furthermore, we can see that in the specific scenario of $\gamma = 1$, as illustrated in Figure 3, the mutual information exhibits significant variability depending on the particular optimal solutions that are achieved.

The adoption of a diagonalized positive definite $\boldsymbol{\Sigma_Z}$ in lieu of an arbitrary positive definite $\boldsymbol{\Sigma_Z}$ yields significant benefits in terms of achieving more informative numerical results and fine-tuning reconstruction accuracy control, as demonstrated in Table 1 and Figure 2.

- Initially, when $\boldsymbol{\Sigma_Z}$ assumes an arbitrary form, as proved analytically in Appendix B.3.1, the solution becomes trivial for $0 < \gamma < 1$, leading to complete data recovery alongside zero mutual information and reconstruction error. This outcome is uninformative.

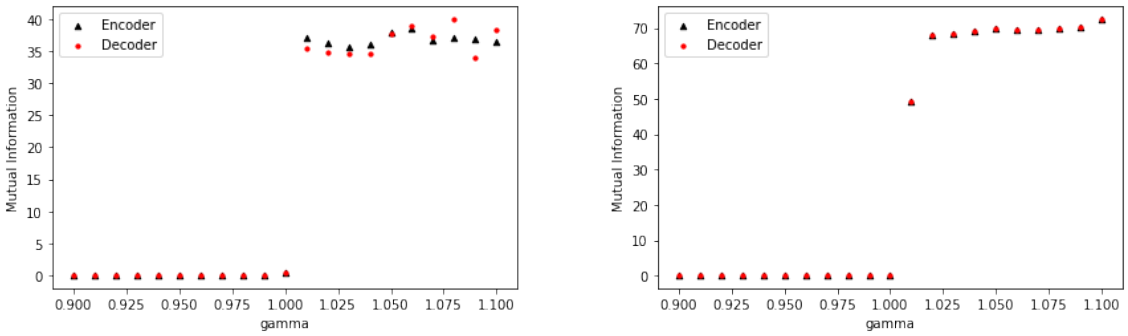

Figure 1: Plot of mutual information w.r.t. γ, given arbitrary $\mathbf{\Sigma_Z}$ (left) and diagonalized $\mathbf{\Sigma_Z}$ (right).

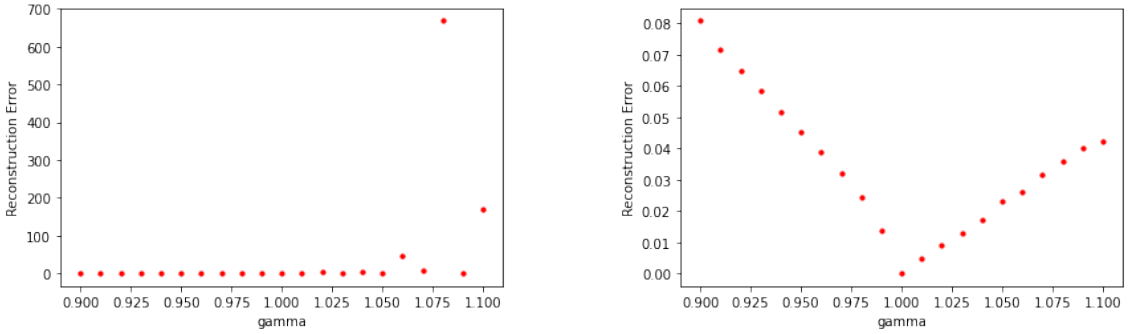

Figure 2: Plot of reconstruction error w.r.t. γ, given arbitrary $\mathbf{\Sigma_Z}$ (left) and diagonalized $\mathbf{\Sigma_Z}$ (right).

- Subsequently, optimizing the γ-VAE model with $\gamma > 1$ using an arbitrary positive definite $\mathbf{\Sigma_Z}$ presents challenges due to intricate correlations introduced among noise dimensions. This complexity can lead to suboptimal reconstruction performance, as depicted in Figure 2. Specifically, the controllability of reconstruction accuracy within the γ-VAE framework, given an arbitrary $\mathbf{\Sigma_Z}$, is compromised. Moreover, the phenomenon of a blow-up effect is numerically observed in the reconstruction error as γ exceeds 1. In contrast, the use of a diagonalized $\mathbf{\Sigma_Z}$ leads to enhanced reconstruction accuracy. The reconstruction error progressively diminishes and ultimately converges to zero as γ ascends from a value less than one to 1 (e.g., from 0.9 to 1). Conversely, as γ transitions from 1 to a value greater than one, such as 1.1 in this specific numerical experiment, the reconstruction error increases, albeit at a relatively slower rate compared to its decrease when $\gamma < 1$.

  Incorporating a diagonalized covariance matrix for the decoder noise results in improved reconstruction accuracy compared to using an arbitrary positive definite matrix for the decoder noise's covariance.

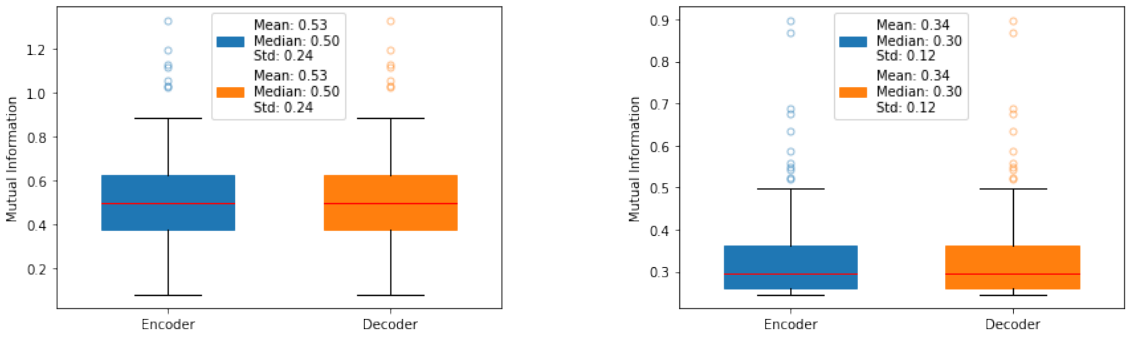

Figure 3: Plot of mutual information at $\gamma = 1$, given arbitrary $\mathbf{\Sigma_Z}$ (left) and diagonalized $\mathbf{\Sigma_Z}$ (right).

This advantage stems from the characteristics and simplifications introduced during mathematical operations within the γ-VAE framework. Specifically, the diagonalized covariance matrix ensures the independence of noise vector elements, simplifying computations and enhancing convergence efficiency in optimization algorithms. Consequently, this positively impacts the model's overall reconstruction accuracy.

9 Numerical Analysis of $\gamma\lambda$-VAE given Diagonalized $\mathbf{\Sigma_Z}$

Despite the transition from an arbitrary covariance matrix $\mathbf{\Sigma_Z}$ to a diagonalized form, the challenge of the blow-up phenomenon persists when $\gamma > 1$ in our pursuit of maximizing mutual information, as illustrated in Figure 1. Moreover, the reconstruction error continues to remain elevated when γ deviates from 1. So, our current objective shifts toward enhancing the controllability of both reconstruction accuracy and mutual information. We aim to achieve this by introducing modifications to the γ-VAE loss function. Our intention is to explore the potential of the new proposed $\gamma\lambda$-VAE framework. The primary aim is to determine whether this framework can yield improved accuracy in reconstructing data. This endeavor reflects our intent to delve into the interplay between the hyperparameters γ and λ within this revised formulation and the associated factors of reconstruction error and mutual information. Through this exploration, we seek to deepen our understanding of how these parameters collectively influence the balance between achieving accurate data reconstruction while preserving meaningful mutual information relationships.

We utilize an alternating iteration algorithm, as outlined in Section 5.3, to determine the optimal solution for the $\gamma\lambda$-VAE loss function (8). Our primary objective throughout this iterative procedure is to minimize the reconstruction error while maintaining a strict maximum threshold of 0.05 for the reconstruction error. We initialize the matrix dimensions to $n = 3$ and $m = 2$, using the same covariance matrix $\mathbf{\Sigma_Y}$ defined as:

$$\mathbf{\Sigma_Y} = \begin{bmatrix} 1 & 0.18 & 0.12 \\ 0.18 & 1 & 0.06 \\ 0.12 & 0.06 & 1 \end{bmatrix}.$$

We consider two arrays for the hyperparameters: one for γ containing values $\{0.98, 0.99, 1, 1.01, 1.02\}$ and the other for λ containing values $\{-0.02, -0.01, 0, 0.01, 0.02\}$. Both arrays have a step size of 0.01. For every pair of (γ, λ) within these arrays, we determine the optimal solution $(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W})$ for the $\gamma\lambda$-VAE function. This optimization is performed while adhering to two key constraints: a diagonalized positive definite $\mathbf{\Sigma_Z}$ and a reconstruction error tolerance of 5%. Our approach for optimization involves utilizing **Lemmas 5** and **6**. Subsequently, we compute the relevant cost function $\mathbf{\Gamma}_2$, as well as the reconstruction error and mutual information for both the encoder and decoder.

Given that each of the arrays has a length of 5, we examine a total of 25 unique combinations of (γ, λ). To determine the uniqueness of the optimal solution, we execute the algorithm 10 times for each specific combination of (γ, λ). It is important to recognize that there could potentially exist multiple solutions for any given (γ, λ) combination. Consequently, we select and retain only the solution that yields the smallest reconstruction error among these possibilities for each respective combination.

9.1 Plots of Numerical Solutions

With the optimal solutions obtained while adhering to the defined constraints, we move forward to present and discuss the results using three different types of plots.

1. The first type of plot, as depicted in Figure 4, employs a three-dimensional format to illustrate the relationship between mutual information and the parameter space (γ, λ). As the mutual information values of the encoder and decoder for each (γ, λ) pair are nearly identical, we exclusively display the mutual information value of the decoder on the graph.

2. The second type of plot, described in Figures 5, 6, and 7, takes the form of scatter plots. These plots visualize the mutual information of both the encoder and decoder in relation to the corresponding reconstruction error.

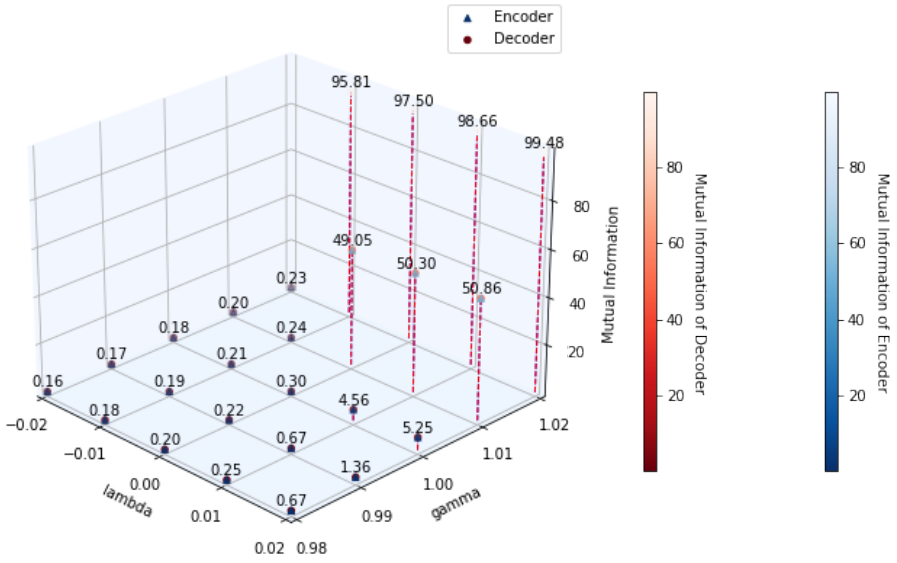

Figure 4: Plot of mutual information w.r.t. (γ, λ) pair.

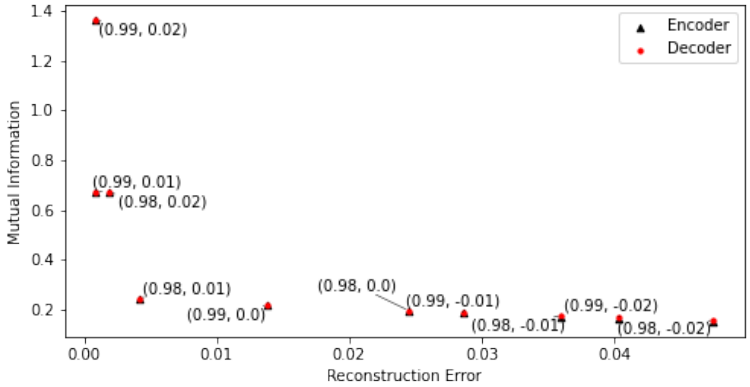

Figure 5: Plot of mutual information w.r.t. reconstruction error for each (γ, λ) pair, given $\gamma < 1$.

3. Lastly, the third type of plot, presented in Figure 8, is presented as a boxplot. This format offers a comparative representation of the mutual information values associated with both the encoder and decoder across three distinct scenarios of γ: $\gamma < 1$, $\gamma = 1$, and $\gamma > 1$.

9.2 Analysis of Numerical Solutions

In Figure 4, a positive correlation becomes evident between mutual information and the parameter γ. Specifically, for a fixed λ, increasing γ values correspond to higher mutual information, while lower γ values lead to reduced mutual information. Moreover, it is apparent that when $\gamma > 1$, the mutual information could potentially increase without bound for any $\lambda \geq 0$. This observation is confirmed by the boxplot presented in Figure 8 for cases where $\gamma > 1$.

However, the issue of mutual information explosion can also arise with γ values not exceeding 1. For instance, as demonstrated in Figure 4 for a fixed $\gamma = 1$, if we increase λ from 0 to a positive value such as $\lambda = 0.01$, the mutual information rapidly escalates from 0.3 to 4.56. This highlights the sensitivity of mutual information to small variations in hyperparameter values. A similar early blow-up phenomenon is also evident in the boxplot of Figure 8 for $\gamma = 1$. Notably, the range of mutual information values exhibits greater dispersion and a significantly higher standard deviation compared to the case of $\gamma = 1$ with arbitrary

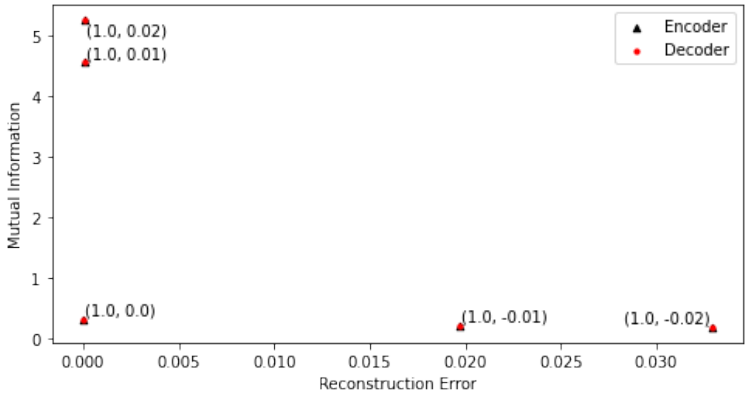

Figure 6: Plot of mutual information w.r.t. reconstruction error for each (γ, λ) pair, given $\gamma = 1$.

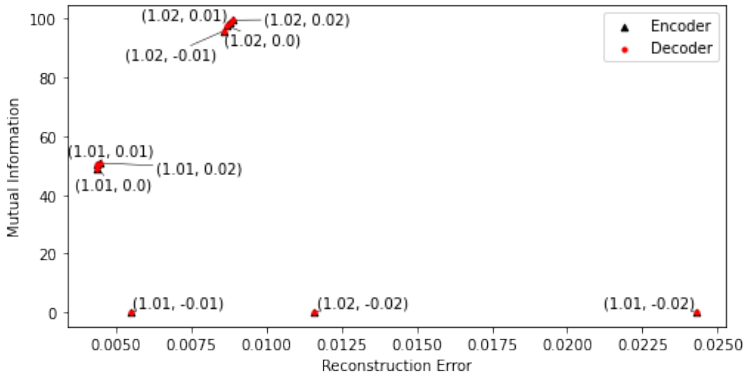

Figure 7: Plot of mutual information w.r.t. reconstruction error for each (γ, λ) pair, given $\gamma > 1$.

$\mathbf{\Sigma_Z}$ and diagonalized $\mathbf{\Sigma_Z}$, as seen in Figure 3. This divergence is attributed to the intricate interplay introduced by the hyperparameter λ.

Nevertheless, the incorporation of λ offers enhanced control over mutual information values, effectively addressing and stabilizing the previously observed blow-up issue. For instance, scenarios without λ (i.e., $\lambda = 0$) and $\gamma > 1$ were susceptible to mutual information escalation within the γ-VAE framework, regardless of the decoder noise covariance structure. However, the incorporation of λ provides a mechanism to counteract this blow-up phenomenon. As exemplified in Figure 4, selecting a negative λ value (e.g., -0.01) alongside a positive γ value (e.g., 1.01) maintains mutual information at a relatively moderate level of approximately 0.24. This instance underscores the potential of a negative λ in mitigating the blow-up tendency often associated with $\gamma > 1$.

The scatter plots illustrating the relationship between mutual information and reconstruction error for various (γ, λ) pairs, as depicted in Figures 5, 6, and 7, underscore our capacity to influence reconstruction accuracy through the incorporation of the additional hyperparameter λ. Notably, all reconstruction error values showcased in these plots consistently fall well below the predefined threshold. Additionally, these plots accentuate the intricate relationship between mutual information and reconstruction error. Mutual information functions as a metric that quantifies the shared information between variables, with the mutual information of the encoder serving as an indicator of the link between the latent space and the original data. The typical trend between mutual information and reconstruction error is inverse: as mutual information increases, the latent space becomes more informative, which, in turn, augments reconstruction precision by leveraging the insights contained within the latent space. Conversely, lower mutual information indicates an inadequately informed latent space, potentially leading to suboptimal data reconstruction due to the limited insights provided by the latent space. However, it is important to note that the linearity of this correlation is influenced by various factors, including data complexity, model architecture, and training quality. While

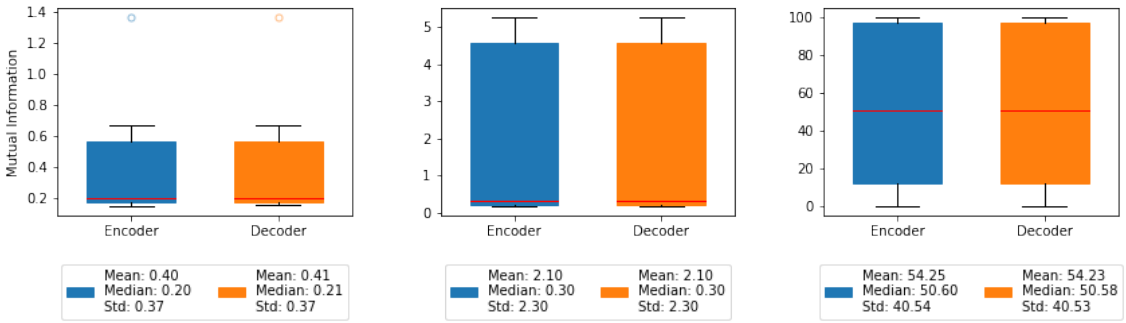

Figure 8: Plot of mutual information, given $\gamma < 1$ (left), $\gamma = 1$ (middle), $\gamma > 1$ (right).

this correlation is generally observed in well-trained models, exceptions may arise due to inadequate training or convergence issues.

Numerical evidence derived from our scatter plots substantiates the existence of a linear correlation between mutual information and reconstruction error. Specifically, in Figure 5, when γ values are fixed, the inverse relationship emerges between mutual information and reconstruction error. Heightened mutual information aligns with improved reconstruction accuracy, whereas increased reconstruction error corresponds to reduced mutual information. However, it is imperative to note that this relationship holds valid only for algorithmically converged optimal solutions, exempt from mutual information escalation. Consider the scenario of $\gamma > 1$ illustrated in Figure 7: with a fixed $\gamma = 1.02$ and non-negative λ values, the algorithm fails to achieve convergence, yielding an invalid optimal solution. In this scenario, the mutual information at $\lambda = 0.01$ is lower than that at $\lambda = 0.02$ ($98.6923 < 99.5084$). According to the anticipated linear correlation between mutual information and reconstruction error, one would expect a higher reconstruction error at $\lambda = 0.01$ compared to $\lambda = 0.02$. However, empirical findings defy this expectation, with $\lambda = 0.01$ yielding a lower value than $\lambda = 0.02$ ($0.0088 < 0.0089$).

10 Conclusion

The objective of this study was to enhance the accuracy of reconstruction in the linear Gaussian β-VAE model. This was achieved by introducing three variations of the β-VAE model: γ-VAE with an arbitrary positive definite $\mathbf{\Sigma_Z}$, γ-VAE with a diagonalized positive definite $\mathbf{\Sigma_Z}$, and $\gamma\lambda$-VAE with a diagonalized positive definite $\mathbf{\Sigma_Z}$.

- Initially, we derived closed-form solutions for all proposed variations using two distinct methods: the gradient-based approach and the alternating iteration algorithm. Through analytical analysis, we demonstrated the consistency between the gradient-based and iterative solutions for both the γ-VAE and $\gamma\lambda$-VAE loss functions, highlighting the robustness of our findings.

- Subsequently, we conducted numerical experiments to compare the first two γ-VAE problems. Our findings revealed that mutual information within the interval $0 < \gamma < 1$ was considerably smaller compared to values for $\gamma > 1$, regardless of decoder noise covariance structure. Notably, we observed the mutual information's high sensitivity to even slight alterations in the parameter γ. This sensitivity manifested as an abrupt blow-up phenomenon in mutual information immediately after γ exceeded 1. The linear model we considered achieved the best approximation of second-order statistics when $\gamma = 1$, which can be attributed to its inherent simplicity. This observation emphasized the importance of considering model complexity and architecture when interpreting the impact of the parameter γ.

  We analytically demonstrated that $\gamma < 1$ led to a unique optimal solution for the γ-VAE with arbitrary $\mathbf{\Sigma_Z}$, while numerical experiments revealed compromised uniqueness in the case of diagonalized $\mathbf{\Sigma_Z}$. Moreover, for $\gamma = 1$, multiple optimal solutions emerged, with the minimum γ-VAE loss function aligning with the input data entropy $H(\boldsymbol{Y})$. Analytical and numerical analyses highlighted convergence challenges associated with singular matrices for $\gamma > 1$ in both γ-VAE variations.

We also illustrated that adopting diagonalized positive definite $\mathbf{\Sigma_Z}$ significantly influenced numerical outcomes and reconstruction accuracy. Analytically, arbitrary $\mathbf{\Sigma_Z}$ resulted in trivial solutions for $0 < \gamma < 1$, achieving complete data recovery but zero mutual information and reconstruction error. Empirical observations indicated that using diagonalized $\mathbf{\Sigma_Z}$ considerably reduced reconstruction error and yielded more consistent outcomes, particularly for $\gamma > 1$.

- Within the $\gamma\lambda$-VAE framework, it was observed that, for a fixed non-negative λ, mutual information within the interval $0 < \gamma < 1$ was notably lower compared to values for $\gamma > 1$. Additionally, we emphasized the sensitivity of mutual information to slight fluctuations in hyperparameter values. However, the introduction of λ allowed for improved control over mutual information values, effectively addressing the earlier blow-up issue. Our numerical experiments highlighted the enhanced manipulation of reconstruction accuracy through the λ hyperparameter, consistently maintaining reconstruction error values below predefined thresholds. Finally, we emphasized the relationship between mutual information and reconstruction error, confirming that its linearity was influenced by data complexity, model architecture, and training quality. While this relationship generally held for well-trained models, exceptions could arise from suboptimal training or convergence difficulties.

In conclusion, through the introduction of three variants of the β-VAE model and comprehensive numerical analyses, we demonstrated the potential for improved reconstruction accuracy by incorporating additional hyperparameters and modifying the underlying framework.

11 Disentanglement

11.1 Existing Disentanglement Metric Score

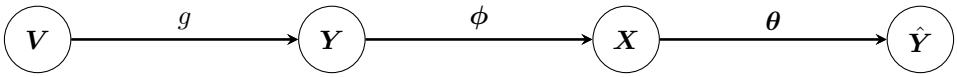

Figure 9: Markov chain diagram of the β-VAE model with an additional generative transition $\boldsymbol{Y} = g(\boldsymbol{V})$.

In Figure 9, each realization of the input data $\boldsymbol{Y} \in \mathbb{R}^n$ is generated from a function of the random variable $\boldsymbol{V} \in \mathbb{R}^s$, representing generative parameters. Here, $s \leq n$. The latent variable $\boldsymbol{X}$ is inferred from the input data $\boldsymbol{Y}$ without prior knowledge of $\boldsymbol{V}$. The β-VAE framework used in this paper is a data compression method closely related to information theory. In cases where little information is available about the nature of the observed data $\boldsymbol{Y}$, the extracted latent parameters $(x_1, x_2, \ldots, x_m)$ can be used to describe the underlying generative parameters $(v_1, v_2, \ldots, v_s)$ that generate the data.

Disentanglement [12] refers to a property of the latent space where variations in a single latent variable correspond to variations in a single generative parameter, resulting in interpretable and transferable representations. The objective is to obtain a disentangled representation of the data, where each latent parameter in $\boldsymbol{X}$ corresponds exclusively to a distinct generative parameter in $\boldsymbol{V}$, capturing the underlying factors of variation in the data. To evaluate disentanglement, we use a disentanglement metric score denoted as S_D. This score measures the correlation between each latent parameter and a single generative parameter, even in scenarios with limited knowledge of these parameters. The disentanglement metric score S_D is defined as:

$$S_D = \frac{1}{m} \sum_i \frac{\max_j |\mathrm{cov}(x_i, v_j)|}{\sum_j |\mathrm{cov}(x_i, v_j)|} \tag{9}$$

Here, x_i represents the i-th component of the latent vector for all $i \in \{1, \cdots, m\}$, and v_j represents the j-th component of the generative parameter vector for all $j \in \{1, \cdots, s\}$. The disentanglement metric score S_D is calculated by taking the ratio of the maximum determinant of the covariance between a latent parameter x_i and any generative parameter v_j to the sum of the determinants of covariances between x_i and all generative parameters v_j. This normalization captures the extent to which the latent parameter is correlated with individual generative parameters. The range of the disentanglement score S_D is $[1/s, 1]$.

The upper bound of 1 indeed signifies perfect disentanglement, where each latent parameter x_i corresponds exclusively to a distinct generative parameter v_j. However, it's essential to emphasize that achieving this upper bound is often very challenging and may not be attainable in practice due to the complexities of real-world data and generative processes. The lower bound of $1/s$ represents a situation where there is no disentanglement. However, it's important to clarify that this lower bound doesn't imply "weak" or "non-existent" correlations between latent and generative parameters. Instead, it means that the correlations exist, but they are distributed in such a way that no single latent parameter is effectively capturing a specific underlying factor of variation in the data. In other words, there might be correlations, but they are not structured in a disentangled manner.

The disentanglement metric score S_D has several limitations:

- The score is designed based on the assumption that each latent parameter should be correlated with only a single generative parameter. However, in real-world scenarios, the relationship between latent parameters and generative parameters may exhibit non-linear behaviors, which can lead to incomplete evaluations.
- The paper [12] specifically considers the scenario where the dimensions of the latent space and generative parameters are equal ($m = s$). However, even in this case, the disentanglement metric score (9) may not fully capture the complexity of the model, potentially limiting its ability to handle diverse data distributions.

To overcome these limitations, our objective is to generalize this metric score and investigate its performance under various scenarios. By exploring a more comprehensive range of scenarios, we can gain a deeper understanding of disentanglement in the model and develop a more robust evaluation framework that can effectively assess the quality of disentangled representations across different data distributions and complex relationships between latent and generative parameters.

11.2 Linear Gaussian Data

Achieving the ideal notion of complete disentanglement, where each latent dimension precisely corresponds to a single generative parameter variation, may not always be fully attainable or practical. Nonetheless, it remains a valuable objective to encourage some level of disentanglement or independence among latent dimensions while considering the presence of correlations and overlaps in the generative parameter space. To address this, we adopt the assumption that the set of latent parameters $(x_1, x_2, \ldots, x_m)$ is statistically *independent*, implying that variations in one latent dimension do not influence the others. On the other hand, we recognize that generative parameters $(v_1, v_2, \ldots, v_s)$ may exhibit *correlations*, meaning that variations in one generative parameter may be related to variations in others.

To investigate disentanglement within the context of our linear problem, we consider a linear Gaussian setting represented by the generative model $(\{v_i\}_{i=1}^s, \boldsymbol{Y})$, where the input data $\boldsymbol{Y}$ is defined as:

$$\boldsymbol{Y} = \sum_{i=1}^{s} v_i \Gamma_i + \tilde{\boldsymbol{Z}}, \tag{10}$$

In this model, $\Gamma_i \in \mathbb{R}^n$ represents a collection of independent eigenvectors obtained from a set of s linearly independent vectors. Each of these vectors corresponds to a standard basis vector within $\mathbb{R}^n$. The noise, denoted as $\tilde{\mathbf{Z}}$, is drawn from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 < 1$. The generative variable $\mathbf{V}$ follows a zero-mean multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{V}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{V}}$ may be a non-full rank matrix indicating correlations among the generative parameters $(v_1, v_2, \ldots, v_s)$. Since $\boldsymbol{Y} = \boldsymbol{\Gamma V} + \tilde{\boldsymbol{Z}}$, the covariance matrix for the data $\boldsymbol{Y}$ is expressed as:

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{Y}} &= \boldsymbol{\Gamma} \boldsymbol{\Sigma}_{\mathbf{V}} \boldsymbol{\Gamma}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{Z}}} \\ &= \boldsymbol{\Gamma} \boldsymbol{\Sigma}_{\mathbf{V}} \boldsymbol{\Gamma}^{\mathsf{T}} + \sigma^2 \mathbf{I}_n, \end{aligned} \tag{11}$$

Here, $\boldsymbol{\Gamma}$ is a $n \times s$ matrix created by stacking the eigenvectors Γ_i, while $\boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}}$ signifies the covariance matrix of the noise.

Utilizing the updated $\boldsymbol{\Sigma_Y}$, the encoding process defined in equation (2) within the linear Gaussian framework can be reformulated as follows:

$$\begin{aligned}\boldsymbol{X} &= \mathbf{B}\boldsymbol{Y} + \boldsymbol{W} \\ &= \mathbf{B}(\boldsymbol{\Gamma V} + \tilde{\mathbf{Z}}) + \boldsymbol{W} \\ &= \mathbf{B}\boldsymbol{\Gamma V} + (\mathbf{B}\tilde{\mathbf{Z}} + \boldsymbol{W}),\end{aligned} \tag{12}$$

where $\mathbf{B} \in \mathbb{R}^{m\times n}$ denotes the encoding matrix and $\boldsymbol{W} \in \mathbb{R}^m$ represents the the encoder noise that follows a Gaussian distribution of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_W})$, where $\boldsymbol{\Sigma_W}$ is a positive definite matrix. Moreover, $\boldsymbol{W}$ is assumed to be independent of the input data $\boldsymbol{Y}$. The decoding process remains consistent, as detailed in equation (3):

$$\hat{\boldsymbol{Y}} = \mathbf{A}\hat{\boldsymbol{X}} + \boldsymbol{Z},$$

where $\mathbf{A} \in \mathbb{R}^{n\times m}$ denotes the decoding matrix, $\hat{\boldsymbol{X}} \in \mathbb{R}^m$ represents a sample drawn from the latent variable distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, and $\boldsymbol{Z} \in \mathbb{R}^n$ stands for the decoder noise following a distribution of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_Z})$, where $\boldsymbol{\Sigma_Z}$ is a positive definite matrix. Additionally, $\boldsymbol{Z}$ is assumed to be independent of $\hat{\boldsymbol{X}}$.

The mean and covariance matrices of both the encoder and decoder are still computed using the framework outlined in system (4). Similarly, the mutual information of the encoder and decoder remains consistent, as described in system (5). The primary distinction lies in the formula of $\boldsymbol{\Sigma_Y}$, which is now governed by the equation (11).

$$\begin{aligned}(\boldsymbol{\mu_X}, \boldsymbol{\Sigma_X}) &= (\mathbf{0}, \mathbf{B}\boldsymbol{\Sigma_Y}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}) \\ (\boldsymbol{\mu}_{\hat{\boldsymbol{Y}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}}) &= (\mathbf{0}, \mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma_Z}) \\ I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X}) &= \frac{1}{2}\log\frac{|\mathbf{B}\boldsymbol{\Sigma_Y}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}|}{|\boldsymbol{\Sigma_W}|} \\ I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}}; \hat{\boldsymbol{Y}}) &= \frac{1}{2}\log\frac{|\mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma_Z}|}{|\boldsymbol{\Sigma_Z}|}.\end{aligned}$$

Using the new assumption within the linear Gaussian setup, incorporating the additional generative model for data $\boldsymbol{Y}$, enables the computation of mutual information between the generative variable $\boldsymbol{V}$ and the latent variable $\boldsymbol{X}$ through the following expression:

$$\begin{aligned}I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X}) &= h(\boldsymbol{X}) - h(\boldsymbol{X}|\boldsymbol{V}) \\ &= h(\boldsymbol{X}) - h(\mathbf{B}\tilde{\mathbf{Z}} + \boldsymbol{W}) \\ &= h(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_X})) - h\left(\mathcal{N}(\mathbf{0}, \mathbf{B}\boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W})\right) \\ &= h(\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_X})) - h\left(\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{B}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W})\right) \\ &= \frac{1}{2}\log((2\pi e)^m|\boldsymbol{\Sigma_X}|) - \frac{1}{2}\log((2\pi e)^m|\sigma^2\mathbf{B}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}|) \\ &= \frac{1}{2}\log\frac{|\mathbf{B}\boldsymbol{\Sigma_Y}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}|}{|\sigma^2\mathbf{B}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}|} \\ &= \frac{1}{2}\log\frac{|\mathbf{B}(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{V}}\boldsymbol{\Gamma}^\mathsf{T} + \sigma^2\mathbf{I}_n)\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}|}{|\sigma^2\mathbf{B}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma_W}|}.\end{aligned} \tag{13}$$

11.2.1 Problem Formulation

We consider a specific scenario where the input data dimension is $n = 4$, the generative parameter dimension is $s = 4$, and the latent dimension is $m = 2$. The generative variable $\mathbf{V}$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_V})$, where the covariance matrix $\boldsymbol{\Sigma_V}$ is singular, indicating correlations among the generative parameters. The generative parameters can be grouped into two pairs of correlated variables: (v_1, v_2) and (v_3, v_4). Specifically, there exists a scaling factor α such that v_2 is related to v_1 as $v_2 = \alpha v_1$. Similarly, there exists a scaling factor β such that v_4 is related to v_3 as $v_4 = \beta v_3$. We assume that the two independent latent parameters x_1 and x_2 can potentially capture the variations in the generative parameters (v_1, v_2) and (v_3, v_4), respectively. This relationship is depicted in the Venn diagram shown in Figure 10. However, to

validate this assumption and assess the effectiveness of disentanglement, further analysis using appropriate measures is required. This will provide insights into how well the latent dimensions describe the underlying variations in the generative parameters and whether disentanglement is achieved.

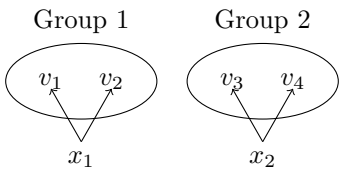

Figure 10: Venn diagram of correlated generative parameters using $s = 4$ and $m = 2$.

Suppose we start by sampling v_1 and v_3 from Gaussian distribution $\mathcal{N}(0, \sigma^2_{v_{13}})$. Subsequently, we introduce two scaling factors, denoted as (α, β), which we employ to compute v_2 and v_4 according to the stipulations $v_2 = \alpha v_1$ and $v_4 = \beta v_3$. Consequently, we acquire the set of 4 generative parameters as

$$[v_1, v_2, v_3, v_4] = [v_1, \alpha v_1, v_3, \beta v_3].$$

Following this, v_2 and v_4 are governed by Gaussian distributions, specifically $\mathcal{N}(0, \alpha^2\sigma^2_{v_{13}})$ and $\mathcal{N}(0, \beta^2\sigma^2_{v_{13}})$, respectively. In relation to the covariance matrix of the generative variable $\boldsymbol{V}$, represented as $\boldsymbol{\Sigma_V}$, this results in:

$$\begin{aligned}
\boldsymbol{\Sigma_V} &= \mathbb{E}[(\boldsymbol{V} - \mathbb{E}[\boldsymbol{V}])(\boldsymbol{V} - \mathbb{E}[\boldsymbol{V}])^\mathsf{T}] \\
&= \begin{bmatrix} \sigma^2_{v_1} & \sigma_{v_1,v_2} & \sigma_{v_1,v_3} & \sigma_{v_1,v_4} \\ \sigma_{v_2,v_1} & \sigma^2_{v_2} & \sigma_{v_2,v_3} & \sigma_{v_2,v_4} \\ \sigma_{v_3,v_1} & \sigma_{v_3,v_2} & \sigma^2_{v_3} & \sigma_{v_3,v_4} \\ \sigma_{v_4,v_1} & \sigma_{v_4,v_2} & \sigma_{v_4,v_3} & \sigma^2_{v_4} \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2_{v_{13}} & \mathbb{E}[v_1 v_2^\mathsf{T}] & 0 & 0 \\ \mathbb{E}[v_2 v_1^\mathsf{T}] & \alpha^2\sigma^2_{v_{13}} & 0 & 0 \\ 0 & 0 & \sigma^2_{v_{13}} & \mathbb{E}[v_3 v_4^\mathsf{T}] \\ 0 & 0 & \mathbb{E}[v_4 v_3^\mathsf{T}] & \beta^2\sigma^2_{v_{13}} \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2_{v_{13}} & \alpha\sigma^2_{v_{13}} & 0 & 0 \\ \alpha\sigma^2_{v_{13}} & \alpha^2\sigma^2_{v_{13}} & 0 & 0 \\ 0 & 0 & \sigma^2_{v_{13}} & \beta\sigma^2_{v_{13}} \\ 0 & 0 & \beta\sigma^2_{v_{13}} & \beta^2\sigma^2_{v_{13}} \end{bmatrix}.
\end{aligned} \tag{14}$$

We proceed by constructing the matrix $\boldsymbol{\Gamma}$ formed by stacking 4 independent eigenvectors, each corresponding to a standard basis vector within $\mathbb{R}^4$. Next, we choose a small value for σ^2. This value is then used to define the covariance matrix for the noise $\tilde{\boldsymbol{Z}}$ as $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Z}}} = \sigma^2\mathbf{I}_4$. By employing equation (11), we derive the covariance matrix for $\boldsymbol{\Sigma_Y}$ as follows:

$$\boldsymbol{\Sigma_Y} = \boldsymbol{\Gamma}\boldsymbol{\Sigma_V}\boldsymbol{\Gamma}^\mathsf{T} + \sigma^2\mathbf{I}_4.$$

To evaluate the disentanglement for this scenario, which demonstrates the capability of latent parameters x_1 and x_2 to effectively represent two groups of generative parameters (v_1, v_2) and (v_3, v_4), respectively, as illustrated in Figure 10, we employ the mutual information formula. In this context, with a latent dimension of $m = 2$, we initially divide the set of generative parameters $\mathcal{V}_4 = \{v_1, v_2, v_3, v_4\}$ into two distinct groups denoted as v_{s_1} and v_{s_2}. Evaluating the proficiency of each latent parameter in capturing the associated generative parameter group necessitates the computation of mutual information between each x_i and its corresponding v_{s_i} for $i = 1, 2$. These computed individual mutual information values are then aggregated. Since the generative parameter groups could exhibit correlations, quantifying the mutual information linking the two groups becomes essential. This correlation is quantified using $I(v_{s_1}; v_{s_2})$. Consequently, our objective lies in assessing whether the latent parameters can effectively encapsulate the underlying generative parameters. To achieve this, our focus shifts towards maximizing the subsequent expression:

$$\max I(x_1; v_{s_1}) + I(x_2; v_{s_2}) - I(v_{s_1}; v_{s_2}). \tag{15}$$

Here,

$$\begin{aligned} &v_{s_1} \subset \mathcal{V}_4, v_{s_2} \subset \mathcal{V}_4 \\ &v_{s_1} \cup v_{s_2} = \mathcal{V}_4 \\ &v_{s_1} \cap v_{s_2} = \varnothing. \end{aligned} \tag{16}$$

We are about to delve into the process of partitioning the generative parameters set $\mathcal{V}_4$. Given that $\mathcal{V}_4$ comprises 4 elements and we intend to distribute these elements into two distinct groups, v_{s_1} and v_{s_2}, while adhering to the conditions outlined in the system (16), we must address three distinct cases:

1. **One element in v_{s_1} and three in v_{s_2}**:

   Given that the set $\mathcal{V}_4$ consists of 4 elements, there are 4 distinct ways to partition $\mathcal{V}_4$, precisely corresponding to the 4 different ways of selecting elements for v_{s_1}.

2. **Two elements in v_{s_1} and two in v_{s_2}**:

   Our primary focus resides in the selection of elements for v_{s_1}, as the composition of v_{s_2} inherently emerges from the remaining components within set $\mathcal{V}_4$. Considering the inclusion of 2 elements within v_{s_1}, and taking into account the presence of 4 elements in set $\mathcal{V}_4$, our aim is to select 2 elements from this particular set. This goal is accomplished through the process of calculating combinations without repetition. Hence, there are $\binom{4}{2} = 6$ distinct ways to partition the set $\mathcal{V}_4$ into two groups, each containing 2 elements.

3. **Three elements in v_{s_1} and one in v_{s_2}**:

   Similar to the second scenario, we have a total of $\binom{4}{3} = 4$ possible ways to choose elements for v_{s_1}.

Hence, the total number of ways to partition a set of 4 generative parameters into 2 distinct groups, satisfying the conditions outlined in equation 16, is 14.

Let $\mathcal{I}_2 = I(x_1; v_{s_1}) + I(x_2; v_{s_2}) - I(v_{s_1}; v_{s_2})$ be the mutual information-based metric that we aim to maximize in our scenario. Since the mutual information between two Gaussian variables is given by

$$I(\boldsymbol{X}; \boldsymbol{Y}) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_{\boldsymbol{X}}||\boldsymbol{\Sigma}_{\boldsymbol{Y}}|}{|\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}|},$$

the metric $\mathcal{I}_2$ can be equivalently expressed as follows:

$$\begin{aligned} \mathcal{I}_2 &= I(x_1; v_{s_1}) + I(x_2; v_{s_2}) - I(v_{s_1}; v_{s_2}) \\ &= \frac{1}{2} \left(\log \frac{|\sigma_{x_1}^2||\boldsymbol{\Sigma}_{v_{s_1}}|}{|\boldsymbol{\Sigma}_{x_1,v_{s_1}}|} + \log \frac{|\sigma_{x_2}^2||\boldsymbol{\Sigma}_{v_{s_2}}|}{|\boldsymbol{\Sigma}_{x_2,v_{s_2}}|} - \log \frac{|\boldsymbol{\Sigma}_{v_{s_1}}||\boldsymbol{\Sigma}_{v_{s_2}}|}{|\boldsymbol{\Sigma}_{v_{s_1},v_{s_2}}|} \right) \\ &= \frac{1}{2} \left(\log \frac{|\boldsymbol{\Sigma}_{v_{s_1}}|}{|\boldsymbol{\Sigma}_{x_1,v_{s_1}}|} + \log \frac{|\boldsymbol{\Sigma}_{v_{s_2}}|}{|\boldsymbol{\Sigma}_{x_2,v_{s_2}}|} - \log \frac{|\boldsymbol{\Sigma}_{v_{s_1}}||\boldsymbol{\Sigma}_{v_{s_2}}|}{|\boldsymbol{\Sigma}_{v_{s_1},v_{s_2}}|} \right). \end{aligned} \tag{17}$$

We now determine the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{V}$, which can be represented as follows:

$$\begin{aligned} \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{V} \end{bmatrix} &\sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{V}}, \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{V}}) \\ &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{V}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{X}} & \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{V}} \\ \boldsymbol{\Sigma}_{\boldsymbol{V},\boldsymbol{X}} & \boldsymbol{\Sigma}_{\boldsymbol{V}} \end{bmatrix} \right) \\ &= \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{X}} & \mathbb{E}[\boldsymbol{X}\boldsymbol{V}^\mathsf{T}] \\ [\mathbb{E}[\boldsymbol{X}\boldsymbol{V}^\mathsf{T}]]^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{V}} \end{bmatrix} \right) \\ &= \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \mathbf{I}_2 & \mathbb{E}[[\mathbf{B}\boldsymbol{\Gamma}\boldsymbol{V} + (\mathbf{B}\tilde{\mathbf{Z}} + \boldsymbol{W})]\boldsymbol{V}^\mathsf{T}] \\ [\mathbb{E}[\boldsymbol{X}\boldsymbol{V}^\mathsf{T}]]^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{V}} \end{bmatrix} \right) \\ &= \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \mathbf{I}_2 & \mathbf{B}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{V}} \\ [\mathbb{E}[\boldsymbol{X}\boldsymbol{V}^\mathsf{T}]]^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{V}} \end{bmatrix} \right) \\ &= \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \mathbf{I}_2 & \mathbf{B}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{V}} \\ \boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{\Gamma}^\mathsf{T}\mathbf{B}^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{V}} \end{bmatrix} \right), \end{aligned} \tag{18}$$

where $\mathbf{\Sigma_V}$ is described in equation (14). So, the joint distribution of x_i and v_{s_i} can be computed as

$$\begin{aligned}\begin{bmatrix} x_i \\ v_{s_i} \end{bmatrix} &\sim \mathcal{N}(\boldsymbol{\mu}_{x_i,v_{s_i}}, \mathbf{\Sigma}_{x_i,v_{s_i}}) \\ &= \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \mathbf{\Sigma}_{x_i,v_{s_i}} \\ \mathbf{\Sigma}_{v_{s_i},x_i} & \mathbf{\Sigma}_{v_{s_i}} \end{bmatrix}\right), \end{aligned} \tag{19}$$

where the covariance matrix $\mathbf{\Sigma}_{x_i,v_{s_i}}$ forms a component of the larger covariance matrix $\mathbf{\Sigma_{X,V}}$ that encompasses all elements associated with x_i and v_{s_i}.

Notice that the determinant $|\mathbf{\Sigma}_{v_{s_1},v_{s_2}}|$ in equation (17) is always zero when $\mathbf{\Sigma_V}$ is singular. Consequently, the mutual information between two sets of generative parameters, denoted as $I(v_{s_1}; v_{s_2})$, becomes undefined in this scenario. To circumvent the issue posed by a singular matrix, we aim to revise the assumption stated in the problem formulation. Specifically, we seek to ensure that $\mathbf{\Sigma_V}$ is nonsingular, thereby enabling computable and meaningful results. We still begin by sampling v_1 and v_3 from the Gaussian distribution, denoted as $\mathcal{N}(0, \sigma^2_{v_{13}})$. Then, we introduce two scaling factors, denoted as (α, β), to generate v_2 and v_4. In addition to these scaling factors, we incorporate small Gaussian noise variables, represented by z_2 and z_4, each following a Gaussian distribution with variances $\sigma^2_{z_2}$ and $\sigma^2_{z_4}$, respectively. Consequently, the generation of v_2 and v_4 is now governed by the following equations: $v_2 = \alpha v_1 + z_2$ and $v_4 = \beta v_3 + z_4$. This yields the set of four generative parameters as

$$[v_1, v_2, v_3, v_4] = [v_1, \alpha v_1 + z_2, v_3, \beta v_3 + z_4].$$

Following this modification, v_2 and v_4 are subject to normal distributions with variances $\alpha^2\sigma^2_{v_1} + \sigma^2_{z_2}$ and $\beta^2\sigma^2_{v_3} + \sigma^2_{z_4}$, i.e., $\mathcal{N}(0, \alpha^2\sigma^2_{v_{13}} + \sigma^2_{z_2})$ and $\mathcal{N}(0, \beta^2\sigma^2_{v_{13}} + \sigma^2_{z_4})$, respectively. Consequently, the covariance matrix of the generative variable $\mathbf{V}$, as defined in equation (14), is now given by:

$$\mathbf{\Sigma_V} = \begin{bmatrix} \sigma^2_{v_{13}} & \alpha\sigma^2_{v_{13}} & 0 & 0 \\ \alpha\sigma^2_{v_{13}} & \alpha^2\sigma^2_{v_{13}} + \sigma^2_{z_2} & 0 & 0 \\ 0 & 0 & \sigma^2_{v_{13}} & \beta\sigma^2_{v_{13}} \\ 0 & 0 & \beta\sigma^2_{v_{13}} & \beta^2\sigma^2_{v_{13}} + \sigma^2_{z_4} \end{bmatrix}. \tag{20}$$

Therefore, instead of employing $\mathbf{\Sigma_V}$ in equation (14) to calculate the joint distribution of $(\mathbf{X}, \mathbf{V})$, we utilize the revised matrix as defined in equation (20).

11.2.2 Numerical Analysis

In this section, we use two arrays of hyperparameters: $\gamma = [0.98, 1.02]$ and $\lambda = [-0.02, 0.02]$, each incremented by 0.01. This setup aligns with the description in Section 9. For every pair of (γ, λ) within these arrays, we seek to find the optimal solution $(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W})$ for the $\gamma\lambda$-VAE loss function (8). Our optimization process adheres to two constraints: the requirement for a diagonalized positive definite $\mathbf{\Sigma_Z}$ and a reconstruction error tolerance of 5%. We employ **Lemmas 5** and **6** as integral components of our optimization approach. Since each of these arrays contains 5 elements, we explore a total of 25 unique combinations of (γ, λ). To ascertain the uniqueness of the optimal solution, we repeat the algorithm 5 times for each specific combination of (γ, λ). It is crucial to note that there may exist multiple solutions for any given (γ, λ) combination. Consequently, we select and retain only the solution that minimizes the reconstruction error among these possibilities for each respective combination.

In contrast to our previous numerical investigation, we deviate from the practice of initiating with a random value for the covariance matrix of the input data $\mathbf{Y}$. Instead, we incorporate the $\gamma\lambda$-VAE model with an additional generative transition $\mathbf{Y} = g(\mathbf{V})$. Therefore, the data $\mathbf{Y}$ is generated using the generative variable $\mathbf{V}$, such that $\mathbf{Y} = \mathbf{\Gamma V} + \tilde{\mathbf{Z}}$. Hence, we initiate the process by selecting parameter values, as delineated in Table 7, in order to construct $\mathbf{\Sigma_V}$. Subsequently, we leverage this covariance matrix to compute $\mathbf{\Sigma_Y}$, as outlined in Table 8.

We will now proceed to present the 14 partitions, each defining the division of the set of 4 generative parameters into 2 distinct groups:

1. $v_{s_1} = \{v_1\}$ and $v_{s_2} = \{v_2, v_3, v_4\}$
2. $v_{s_1} = \{v_2\}$ and $v_{s_2} = \{v_1, v_3, v_4\}$

α	β	$\sigma_{v_{13}}$	σ_{z_2}	σ_{z_4}	σ	$\mathbf{\Gamma}$
2	4	0.05	0.02	0.03	0.04	$[\boldsymbol{e}_1, \boldsymbol{e}_4, \boldsymbol{e}_2, \boldsymbol{e}_3]$

Table 7: Parameter values used in numerical analysis in Section 11.2.2.

$\mathbf{\Sigma_V}$	$\mathbf{\Sigma_Y}$
$\begin{bmatrix} 0.0025 & 0.005 & 0 & 0 \\ 0.005 & 0.0104 & 0 & 0 \\ 0 & 0 & 0.0025 & 0.01 \\ 0 & 0 & 0.01 & 0.0409 \end{bmatrix}$	$\begin{bmatrix} 0.0041 & 0 & 0 & 0.005 \\ 0 & 0.0041 & 0.01 & 0 \\ 0 & 0.01 & 0.0425 & 0 \\ 0.005 & 0 & 0 & 0.012 \end{bmatrix}$

Table 8: $\mathbf{\Sigma_V}$ and $\mathbf{\Sigma_Y}$ used in numerical analysis in Section 11.2.2.

3. $v_{s_1} = \{v_3\}$ and $v_{s_2} = \{v_1, v_2, v_4\}$
4. $v_{s_1} = \{v_4\}$ and $v_{s_2} = \{v_1, v_2, v_3\}$
5. $v_{s_1} = \{v_1, v_2\}$ and $v_{s_2} = \{v_3, v_4\}$
6. $v_{s_1} = \{v_1, v_3\}$ and $v_{s_2} = \{v_2, v_4\}$
7. $v_{s_1} = \{v_1, v_4\}$ and $v_{s_2} = \{v_2, v_3\}$
8. $v_{s_1} = \{v_2, v_3\}$ and $v_{s_2} = \{v_1, v_4\}$
9. $v_{s_1} = \{v_2, v_4\}$ and $v_{s_2} = \{v_1, v_3\}$
10. $v_{s_1} = \{v_3, v_4\}$ and $v_{s_2} = \{v_1, v_2\}$
11. $v_{s_1} = \{v_1, v_2, v_3\}$ and $v_{s_2} = \{v_4\}$
12. $v_{s_1} = \{v_1, v_2, v_4\}$ and $v_{s_2} = \{v_3\}$
13. $v_{s_1} = \{v_1, v_3, v_4\}$ and $v_{s_2} = \{v_2\}$
14. $v_{s_1} = \{v_2, v_3, v_4\}$ and $v_{s_2} = \{v_1\}$.

In order to assess alternative metrics for quantifying disentanglement and to conduct a comparative analysis with our mutual information-based metric $\mathcal{I}_2$, we introduce the correlation-based metric S_D. This metric, initially presented in [12] and elaborated upon in Section 11.1, is now applied within the framework of our linear Gaussian setting, as outlined in Section 11.2.1. The computation of the metric S_D, defined by equation (9), is described as follows:

$$S_D = \frac{1}{2} \sum_{i=1}^{2} \frac{\max_j |\mathrm{cov}(x_i, v_j)|}{\sum_{j=1}^{4} |\mathrm{cov}(x_i, v_j)|}. \tag{21}$$

Here x_i denotes the i-th component of the latent vector for all $i = 1, 2$, while v_j represents the j-th component of the generative parameter vector for all $j \in \{1, \ldots, 4\}$.

For every pair of (γ, λ), we select the optimal solution with the minimal reconstruction error. From the optimal solution, we present the following information in Table 9: the reconstruction error, denoted as $\mathcal{L}_{\mathrm{rec}}$; the mutual information of the encoder, represented as $I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$; the mutual information between the generative variable $\mathbf{V}$ and the latent variable $\mathbf{X}$, computed using equation (13); the metric S_D value obtained through equation (21); and the maximum value of $\mathcal{I}_2$ alongside the corresponding partition.

According to the data presented in Table 9, we have the following findings:

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	S_D	Max $\mathcal{I}_2$	Partition
$(0.98, -0.02)$	0.0015	1.8422	1.2945	0.6496	1.2945	5
$(0.98, -0.01)$	0.0015	1.8424	1.2962	0.7186	0.5052	5
$(0.98, 0)$	0.0015	1.8425	1.298	0.6538	1.246	10
$(0.98, 0.01)$	0.0015	1.8427	1.2997	0.7109	0.677	10
$(0.98, 0.02)$	0.0015	1.8429	1.3015	0.6794	1.0221	5
$(0.99, -0.02)$	0.0007	1.863	1.3015	0.6503	1.3015	5
$(0.99, -0.01)$	0.0007	1.863	1.3049	0.6522	1.2647	5
$(0.99, 0)$	0.0007	1.8632	1.3082	0.6516	1.3082	10
$(0.99, 0.01)$	0.0007	1.8634	1.3117	0.6521	1.3107	10
$(0.99, 0.02)$	0.0007	1.8636	1.3151	0.6555	1.2306	5
$(1, -0.02)$	0	2.0854	1.1247	0.6813	1.1185	10
$(1, -0.01)$	0	1.9612	1.1856	0.6717	1.1814	10
$(1, 0)$	0	1.8834	1.3182	0.7066	0.7435	5
$(1, 0.01)$	0	1.9805	1.5316	0.6669	1.3914	10
$(1, 0.02)$	0.0001	2.1512	1.7115	0.7048	0.5605	10
$(1.01, -0.02)$	0.0031	47.3344	0.941	0.6923	0.9381	10
$(1.01, -0.01)$	0.0031	46.7199	0.941	0.6918	0.9434	5
$(1.01, 0)$	0.0031	39.5042	1.4779	0.7311	1.3091	5
$(1.01, 0.01)$	0.0049	46.762	2.6472	0.6838	1.62	5
$(1.01, 0.02)$	0.0049	47.3582	2.6472	0.6639	1.2142	5
$(1.02, -0.02)$	0.0061	95.8937	0.941	0.6934	0.9283	5
$(1.02, -0.01)$	0.0061	95.208	0.941	0.691	0.954	10
$(1.02, 0)$	0.0095	89.5331	2.1102	0.7857	2.236	5
$(1.02, 0.01)$	0.0096	95.3382	2.6472	0.7116	1.7949	5
$(1.02, 0.02)$	0.0096	95.9531	2.6472	0.6976	0.8888	5

Table 9: Reconstruction error, mutual information, and disentanglement metric values for each (γ, λ) pair given 4 generative parameters.

- For all considered pairs (γ, λ), the value of $\mathcal{I}_2$ reaches its maximum at either Partition 5 or 10. In Partition 5, the 4 generative parameters are partitioned into 2 groups: $v_{s_1} = \{v_1, v_2\}$ and $v_{s_2} = \{v_3, v_4\}$. Partition 10 closely resembles Partition 5, except that the elements from the first and second groups are interchanged, resulting in $v_{s_1} = \{v_3, v_4\}$ and $v_{s_2} = \{v_1, v_2\}$. It is important to note that both partitions share a common characteristic: each group consists of two elements, where one is a linear transformation of the other. Specifically, $v_2 = \alpha v_1 + z_2$ and $v_4 = \beta v_3 + z_4$. Therefore, the metric $\mathcal{I}_2$ serves as an effective disentanglement metric in this scenario, revealing which latent parameters can accurately capture each group of dependent generative parameters.

- For various combinations of (γ, λ), the disentanglement score S_D consistently falls within the range of $[0.25, 1]$ and remains above 0.5. It is well understood that the closer the value of S_D is to 1, the better the disentanglement. However, these intermediate values merely suggest that while there is some degree of disentanglement, it may not be perfect.

  In this scenario, $\mathcal{I}_2$ proves to be more useful compared to S_D because it allows us to effectively determine which group of generative parameters can be captured using either of the two latent parameters. Nonetheless, it is essential to acknowledge that the metric S_D is designed under the presumption that each latent parameter should be correlated with only one generative parameter. In our specific scenario, we observe that one latent parameter can effectively capture more than one generative parameter, rendering S_D inaccurate for our model.

11.3 Numerical Investigation of Scenario Expansion

In this section, we will modify the problem formulation described in Section 11.2.1 to consider two new scenarios:

1. In the first scenario, we change the dimension of the generative variable $\boldsymbol{V}$ to $s = 3$, while keeping the dimension of the input data $\boldsymbol{Y}$ as $n = 4$, and the dimension of the latent variable $\boldsymbol{X}$ as $m = 2$. We will then consider three cases:

   (a) **Independence of Generative Parameters**: We initiate by independently sampling each of the three generative parameters, v_1, v_2, and v_3, from their respective Gaussian distributions, which are $\mathcal{N}(0, \sigma_{v_1}^2)$, $\mathcal{N}(0, \sigma_{v_2}^2)$, and $\mathcal{N}(0, \sigma_{v_3}^2)$. Consequently, the covariance matrix of the generative variable $\mathbf{V}$ becomes

   $$\boldsymbol{\Sigma_V} = \begin{bmatrix} \sigma_{v_1}^2 & 0 & 0 \\ 0 & \sigma_{v_2}^2 & 0 \\ 0 & 0 & \sigma_{v_3}^2 \end{bmatrix}. \tag{22}$$

   (b) **Linear Dependence of v_1 and v_2, with Independence of v_3**: In this case, we begin by independently sampling v_1 and v_3 from Gaussian distributions, specifically $\mathcal{N}(0, \sigma_{v_1}^2)$ and $\mathcal{N}(0, \sigma_{v_3}^2)$, respectively. We then introduce a scaling factor, denoted as α, to calculate v_2 as per the equation $v_2 = \alpha v_1 + z_2$, where z_2 follows a Gaussian distribution, specifically $z_2 \sim \mathcal{N}(0, \sigma_{z_2}^2)$. So, the covariance matrix of the generative variable $\mathbf{V}$ is given by

   $$\boldsymbol{\Sigma_V} = \begin{bmatrix} \sigma_{v_1}^2 & \alpha\sigma_{v_1}^2 & 0 \\ \alpha\sigma_{v_1}^2 & \alpha^2\sigma_{v_1}^2 + \sigma_{z_2}^2 & 0 \\ 0 & 0 & \sigma_{v_3}^2 \end{bmatrix}. \tag{23}$$

   (c) **Linear Dependence of v_2 and v_3 on v_1**: In this scenario, we initially sample v_1 from a Gaussian distribution: $v_1 \sim \mathcal{N}(0, \sigma_{v_1}^2)$. Subsequently, we introduce two scaling factors, denoted as α and β, to calculate v_2 and v_3 respectively, according to the equations $v_2 = \alpha v_1 + z_2$ and $v_3 = \beta v_1 + z_3$, where z_2 and z_3 follow Gaussian distributions, specifically $z_2 \sim \mathcal{N}(0, \sigma_{z_2}^2)$ and $z_3 \sim \mathcal{N}(0, \sigma_{z_3}^2)$. As a result, the covariance matrix for the generative variable $\mathbf{V}$ transforms into

   $$\boldsymbol{\Sigma_V} = \begin{bmatrix} \sigma_{v_1}^2 & \alpha\sigma_{v_1}^2 & \beta\sigma_{v_1}^2 \\ \alpha\sigma_{v_1}^2 & \alpha^2\sigma_{v_1}^2 + \sigma_{z_2}^2 & \alpha\beta\sigma_{v_1}^2 \\ \beta\sigma_{v_1}^2 & \alpha\beta\sigma_{v_1}^2 & \beta^2\sigma_{v_1}^2 + \sigma_{z_3}^2 \end{bmatrix}. \tag{24}$$

   Considering that the data $\mathbf{Y}$ is generated according to the equation $\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{V} + \tilde{\mathbf{Z}}$, in all three of these cases, we maintain a constant value for the 4×3 matrix $\boldsymbol{\Gamma}$, which is formed by stacking three independent eigenvectors, each corresponding to a standard basis vector within $\mathbb{R}^4$. Subsequently, we select a small value for σ^2. This chosen value is then utilized to define the covariance matrix for the noise $\tilde{\boldsymbol{Z}}$ as $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{Z}}} = \sigma^2\mathbf{I}_4$. So, we calculate the covariance matrix for $\boldsymbol{\Sigma_Y}$ as follows:

   $$\boldsymbol{\Sigma_Y} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{V}}\boldsymbol{\Gamma}^{\mathsf{T}} + \sigma^2\mathbf{I}_4.$$

2. In the second scenario, we increase the dimension of the latent variable $\boldsymbol{X}$ to $m = 3$, while keeping the dimension of $\boldsymbol{V}$ at $s = 3$, and the dimension of $\boldsymbol{Y}$ at $n = 4$. We also explore three distinct cases as delineated in the first scenario.

The objective is to assess the effectiveness of our proposed information-based metric $\mathcal{I}_2$ and the correlation-based metric S_D as disentanglement metrics under these modified scenarios.

11.3.1 $(s, n, m) = (3, 4, 2)$

Similar to Section 11.2.2, we continue to utilize two arrays of hyperparameters: $\gamma = [0.98, 1.02]$ and $\lambda = [-0.02, 0.02]$, with increments of 0.01 for each. For each pair of (γ, λ) contained within these arrays, our objective is to determine the optimal solution $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma_Z}, \boldsymbol{\Sigma_W})$ for the $\gamma\lambda$-VAE loss function defined in

equation (8). Our optimization process still adheres to two constraints: the requirement for a diagonalized positive definite $\mathbf{\Sigma_Z}$ and a reconstruction error tolerance of 5%. To ascertain the uniqueness of the optimal solution, we iterate the algorithm 5 times for each specific combination of (γ, λ). In cases where multiple solutions exist for a given (γ, λ) combination, we select and retain only the solution that minimizes the reconstruction error among these possibilities for that particular combination. Given that the dimension of the latent variable remains unchanged at $m = 2$, the formula for computing the information-based metric $\mathcal{I}_2$ remains the same as equation (17).

However, in contrast to the problem described in Section 11.2.1, where we established the existence of 14 distinct ways to partition the set $\mathcal{V}_4$ consisting of 4 generative parameters into 2 distinct groups, satisfying the conditions outlined in system (16), the new scenario involves only 3 generative parameters. Consequently, there are only 6 distinct ways to partition the set $\mathcal{V}_3 = \{v_1, v_2, v_3\}$ into 2 distinct groups while still meeting the conditions outlined in system (16). Here, $\mathcal{V}_4$ has been replaced by $\mathcal{V}_3$. For each pair (γ, λ), this leads to the computation of 6 different values of $\mathcal{I}_2$, each corresponding to a specific partition. We then present only the highest value of $\mathcal{I}_2$ along with its corresponding partition, as our objective is to maximize this metric. we now list the 6 partitions, each of which divides the set of 3 generative parameters into two distinct groups, denoted as v_{s_1} and v_{s_2}:

1. $v_{s_1} = \{v_1\}$ and $v_{s_2} = \{v_2, v_3\}$
2. $v_{s_1} = \{v_2\}$ and $v_{s_2} = \{v_1, v_3\}$
3. $v_{s_1} = \{v_3\}$ and $v_{s_2} = \{v_1, v_2\}$
4. $v_{s_1} = \{v_1, v_2\}$ and $v_{s_2} = \{v_3\}$
5. $v_{s_1} = \{v_1, v_3\}$ and $v_{s_2} = \{v_2\}$
6. $v_{s_1} = \{v_2, v_3\}$ and $v_{s_2} = \{v_1\}$

Notice that Partition 1 is a permutation symmetry of Partition 6, as are Partitions 2 and 5, and Partitions 3 and 4.

Next, we select the specific parameter values to be employed consistently across all cases within the first scenario, as outlined in Table 10. Using this table, we can calculate the corresponding $\mathbf{\Sigma_V}$ and $\mathbf{\Sigma_Y}$ for each

α	β	σ_{v_1}	σ_{v_2}	σ_{v_3}	σ_{z_2}	σ_{z_3}	σ	$\mathbf{\Gamma}$
2	4	0.01	0.02	0.03	0.02	0.03	0.04	$[\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3]$

Table 10: Parameter values used across all cases within Scenario 1.

case in Scenario 1, as outlined in Table 11.

For each pair of (γ, λ), we choose the optimal solution that yields the minimum reconstruction error. From this optimal solution, we provide the following information in Tables 12, 13, and 14, corresponding to the three cases within Scenario 1: reconstruction error, encoder mutual information, mutual information between generative variable $\boldsymbol{V}$ and latent variable $\boldsymbol{X}$, disentanglement metric S_D value, and the maximum value of $\mathcal{I}_2$ along with its corresponding partition.

Based on the information presented in these three tables, we draw the following conclusions:

- Table 12 presents numerical results for Case 1 within Scenario 1, assuming independence among all three generative parameters denoted as v_1, v_2, and v_3. However, due to the constraint of having only two available latent parameters, specifically x_1 and x_2, it is impossible to validly partition these three independent generative parameters into two distinct groups that satisfy the conditions outlined in (16).

  We observe that the optimal partition varies across different combinations of (γ, λ). Even when a particular (γ, λ) is fixed, conducting numerical investigations multiple times leads to different optimal partition numbers. For example, given $(\gamma, \lambda) = (1.01, -0.02)$, the optimal partitions listed in Table 12 are 2 and 6. However, for other numerical simulations, we may obtain Partitions 3, 5, and 6. Additionally, it's important to note that these optimal partitions may not be permutation symmetries of each other. For instance, Partitions 2 and 6 are not permutation symmetries of others, and the same

Cases	$\boldsymbol{\Sigma_V}$	$\boldsymbol{\Sigma_Y}$
1	$\begin{bmatrix} 0.0001 & 0 & 0 \\ 0 & 0.0004 & 0 \\ 0 & 0 & 0.0009 \end{bmatrix}$	$\begin{bmatrix} 0.0017 & 0 & 0 & 0 \\ 0 & 0.002 & 0 & 0 \\ 0 & 0 & 0.0025 & 0 \\ 0 & 0 & 0 & 0.0016 \end{bmatrix}$
2	$\begin{bmatrix} 0.0001 & 0.0002 & 0 \\ 0.0002 & 0.0008 & 0 \\ 0 & 0 & 0.0009 \end{bmatrix}$	$\begin{bmatrix} 0.0017 & 0.0002 & 0 & 0 \\ 0.0002 & 0.0024 & 0 & 0 \\ 0 & 0 & 0.0025 & 0 \\ 0 & 0 & 0 & 0.0016 \end{bmatrix}$
3	$\begin{bmatrix} 0.0001 & 0.0002 & 0.0004 \\ 0.0002 & 0.0008 & 0.0008 \\ 0.0004 & 0.0008 & 0.0025 \end{bmatrix}$	$\begin{bmatrix} 0.0017 & 0.0002 & 0.0004 & 0 \\ 0.0002 & 0.0024 & 0.0008 & 0 \\ 0.0004 & 0.0008 & 0.0041 & 0 \\ 0 & 0 & 0 & 0.0016 \end{bmatrix}$

Table 11: $\boldsymbol{\Sigma_V}$ and $\boldsymbol{\Sigma_Y}$ for each case within Scenario 1.

applies to Partitions 3, 5, and 6. Hence, it is not sufficient to conclude that $\mathcal{I}_2$ consistently reaches its maximum value at a fixed partition for a specific combination of (γ, λ).

In cases where the dimension of the latent variable is less than the dimension of the generative variable, and all generative parameters are independent, establishing the most effective method of mapping latent parameters to generative parameters presents a significant challenge. This is due to the absence of a consistent optimal partition across different combinations of (γ, λ), making it difficult to unequivocally ascertain which latent parameter should be used to capture each specific generative parameter.

- Table 13 presents numerical results for Case 2 within Scenario 1, where generative parameters v_1 and v_2 are assumed to be linearly dependent, while they remain independent from v_3. Among the six listed partitions, it is observed that either Partition 3 or Partition 4 accurately represents the relationship among the generative parameters, and these partitions are permutation symmetries of each other. In Partition 3, group v_{s_1} contains the independent generative parameter v_3, while group v_{s_2} comprises the two generative parameters v_1 and v_2. In Partition 4, the elements in each group are switched, but the configuration remains consistent. The numerical results presented in Table 13 demonstrate that the disentanglement metric $\mathcal{I}_2$ attains its maximum value at either Partition 3 or Partition 4. This underscores the effectiveness of $\mathcal{I}_2$ as a metric for evaluating disentanglement in cases where there is precisely one partition that accurately describes the relationship among generative parameters, with permutations of this partition not counted as new solutions.

- Table 14 presents numerical results for Case 3 within Scenario 1, where we assume that generative parameters v_2 and v_3 are linearly dependent on v_1. In this scenario, it is theoretically possible to capture all three generative parameters using a single latent parameter, rendering the use of two latent parameters redundant. However, for the sake of the current analysis, we presume a necessity to partition these three generative parameters into two distinct groups, with neither group being empty. The valid partition that accurately describes the relationship among these generative parameters in this case adheres to the following rule: one group contains the two generative parameters that are linearly dependent, while the other group contains the remaining one. However, all 6 partitions satisfy this rule. Thus, the question arises: how can we determine the optimal partition among these partitions?

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	S_D	Max $\mathcal{I}_2$	Partition
$(0.98, -0.02)$	0	0	0	0.6429	0	All 6 partitions
$(0.98, -0.01)$	0	0	0	0.6429	0	All 6 partitions
$(0.98, 0)$	0	0	0	0.6429	0	All 6 partitions
$(0.98, 0.01)$	0	0	0	0.6429	0	All 6 partitions
$(0.98, 0.02)$	0	0	0	0.6429	0	All 6 partitions
$(0.99, -0.02)$	0	0	0	0.6429	0	All 6 partitions
$(0.99, -0.01)$	0	0	0	0.6429	0	All 6 partitions
$(0.99, 0)$	0	0	0	0.6429	0	All 6 partitions
$(0.99, 0.01)$	0	0	0	0.6429	0	All 6 partitions
$(0.99, 0.02)$	0	0	0	0.6429	0	All 6 partitions
$(1, -0.02)$	0	0.0004	0.0001	0.6429	0.0001	1, 3, 5
$(1, -0.01)$	0	0.0023	0.0004	0.6429	0.0004	1, 3, 5
$(1, 0)$	0	0.0008	0.0001	0.6429	0.0001	3, 6
$(1, 0.01)$	0	0.0011	0	0.6429	0	All 6 partitions
$(1, 0.02)$	0	0.0031	0.0011	0.6426	0.0008	1, 4
$(1.01, -0.02)$	0.004	44.872	0.1419	0.6638	0.1423	2, 6
$(1.01, -0.01)$	0.004	45.8062	0.1419	0.6634	0.1013	2, 6
$(1.01, 0)$	0.004	45.6605	0.1419	0.6638	0.1409	2, 6
$(1.01, 0.01)$	0.004	45.5546	0.1419	0.6638	0.1411	2, 6
$(1.01, 0.02)$	0.0034	46.0782	0.0303	0.6442	0.0305	2, 3, 6
$(1.02, -0.02)$	0.0067	95.0585	0.0303	0.6442	0.0187	1, 4, 5
$(1.02, -0.01)$	0.0067	95.0066	0.0303	0.6442	0.0278	1, 4, 5
$(1.02, 0)$	0.0067	94.7948	0.0303	0.6442	0.0187	2, 3, 6
$(1.02, 0.01)$	0.0067	94.4385	0.0303	0.6442	0.0221	1, 4, 5
$(1.02, 0.02)$	0.0067	94.6457	0.0303	0.6442	0.0294	1, 4, 5

Table 12: Reconstruction error, mutual information, and disentanglement metric values for each (γ, λ) pair in Case 1 of Scenario 1.

Indeed, we observe consistency in the column indicating the optimal partition in Table 14. According to the numerical results, the $\mathcal{I}_2$ metric reaches its maximum value at either Partition 2 or Partition 5, and these partitions are permutation symmetries of each other. In Partition 2, group v_{s_1} comprises v_2, and group v_{s_2} includes v_1 and v_3. Conversely, in Partition 5, the grouping is reversed, with group v_{s_1} containing v_1 and v_3, and group v_{s_2} containing v_2.

Based on the optimal partition, it suggests that generative parameter v_2 should be captured by one latent parameter independently, while the remaining generative parameters, namely v_1 and v_3, can be captured using the other latent parameter. The rationale behind grouping v_1 and v_3 together instead of pairing v_1 and v_2 is rooted in the correlation coefficients employed to derive v_3 (denoted as β) and v_2 (denoted as α) from v_1. Given that $\beta > \alpha$, it follows that the covariance between v_1 and v_3 exceeds the covariance between v_1 and v_2, expressed as $\beta\sigma_{v_1}^2 > \alpha\sigma_{v_1}^2$, signifying a stronger linear dependence between v_1 and v_3.

Despite the absence of an ideal partition where all generative parameters can be captured using only one latent parameter, the $\mathcal{I}_2$ metric remains a robust tool for effectively evaluating disentanglement. It can identify the optimal partition from a set of multiple partitions, provided that not all of them are permutation symmetries of each other.

11.3.2 $(s, n, m) = (3, 4, 3)$

Given that both the latent dimension and the dimension of the generative variable are 3, we proceed to partition the set of these 3 generative parameters, denoted as $\mathcal{V}_3 = \{v_1, v_2, v_3\}$, into three distinct generative

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	S_D	Max $\mathcal{I}_2$	Partition
$(0.98, -0.02)$	0.0152	0.0789	0.02	0.5046	0.0199	3
$(0.98, -0.01)$	0.0152	0.0789	0.02	0.5046	0.0196	3
$(0.98, 0)$	0.0152	0.0789	0.02	0.5046	0.0192	3
$(0.98, 0.01)$	0.0151	0.079	0.02	0.5046	0.019	3
$(0.98, 0.02)$	0.0151	0.079	0.02	0.5046	0.0122	3
$(0.99, -0.02)$	0.0075	0.0892	0.0224	0.5051	0.019	3
$(0.99, -0.01)$	0.0075	0.0892	0.0224	0.5051	0.0128	3
$(0.99, 0)$	0.0074	0.0892	0.0224	0.5052	0.0195	3
$(0.99, 0.01)$	0.0074	0.0893	0.0225	0.5052	0.0209	3
$(0.99, 0.02)$	0.0074	0.0893	0.0225	0.5052	0.0192	3
$(1, -0.02)$	0.0001	0.0997	0.0238	0.5054	0.0182	4
$(1, -0.01)$	0	0.0995	0.0242	0.5055	0.0232	4
$(1, 0)$	0	0.0993	0.0249	0.5057	0.0246	3
$(1, 0.01)$	0	0.1	0.0261	0.506	0.0229	3
$(1, 0.02)$	0.0001	0.1006	0.0269	0.5062	0.0241	3
$(1.01, -0.02)$	0.0034	46.5913	0.0303	0.5042	0.0304	3
$(1.01, -0.01)$	0.0048	45.5206	0.2027	0.5441	0.1482	3
$(1.01, 0)$	0.0034	45.2469	0.0303	0.5041	0.0206	3
$(1.01, 0.01)$	0.0048	46.1141	0.2027	0.5466	0.1982	3
$(1.01, 0.02)$	0.0048	45.2942	0.2027	0.5437	0.1376	3
$(1.02, -0.02)$	0.0095	95.3216	0.2027	0.5472	0.2049	3
$(1.02, -0.01)$	0.0068	94.3106	0.0303	0.5042	0.0306	3
$(1.02, 0)$	0.0099	95.1758	0.2535	0.4871	0.1553	3
$(1.02, 0.01)$	0.0068	95.2	0.0303	0.5042	0.0294	4
$(1.02, 0.02)$	0.0095	95.0535	0.2027	0.5464	0.1919	3

Table 13: Reconstruction error, mutual information, and disentanglement metric values for each (γ, λ) pair in Case 2 of Scenario 1.

parameter groups, labeled as v_{s_1}, v_{s_2}, and v_{s_3}. Consequently, we modify the information-based metric from $\mathcal{I}_2$ to $\mathcal{I}_3$ as follows:

$$\mathcal{I}_3 = I(x_1; v_{s_1}) + I(x_2; v_{s_2}) + I(x_3; v_{s_3}) - I(v_{s_1}; v_{s_2}) - I(v_{s_1}; v_{s_3}) - I(v_{s_2}; v_{s_3}). \tag{25}$$

Here,

$$\begin{aligned} &v_{s_1} \subset \mathcal{V}_3, v_{s_2} \subset \mathcal{V}_3, v_{s_3} \subset \mathcal{V}_3 \\ &v_{s_1} \cup v_{s_2} \cup v_{s_3} = \mathcal{V}_3 \\ &v_{s_1} \cap v_{s_2} \cap v_{s_3} = \varnothing. \end{aligned} \tag{26}$$

So, our objective is to maximize $\mathcal{I}_3$, which can be expanded as below:

$$\begin{aligned} \mathcal{I}_3 &= \sum_{m=1}^{3} I(x_m; v_{s_m}) - I(v_{s_1}; v_{s_2}) - I(v_{s_1}; v_{s_3}) - I(v_{s_2}; v_{s_3}) \\ &= \frac{1}{2}\left(\sum_{m=1}^{3} \log \frac{|\boldsymbol{\Sigma}_{v_{s_m}}|}{|\boldsymbol{\Sigma}_{x_m, v_{s_m}}|} - \log \frac{|\boldsymbol{\Sigma}_{v_{s_1}}||\boldsymbol{\Sigma}_{v_{s_2}}|}{|\boldsymbol{\Sigma}_{v_{s_1}, v_{s_2}}|} - \log \frac{|\boldsymbol{\Sigma}_{v_{s_1}}||\boldsymbol{\Sigma}_{v_{s_3}}|}{|\boldsymbol{\Sigma}_{v_{s_1}, v_{s_3}}|} - \log \frac{|\boldsymbol{\Sigma}_{v_{s_2}}||\boldsymbol{\Sigma}_{v_{s_3}}|}{|\boldsymbol{\Sigma}_{v_{s_2}, v_{s_3}}|} \right). \end{aligned} \tag{27}$$

Notice that there are 6 ways to partition the 3 generative parameters into 3 distinct groups that satisfy the system of conditions (26). Furthermore, all these 6 partitions exhibit permutation symmetry among each other.

1. $v_{s_1} = \{v_1\}$, $v_{s_2} = \{v_2\}$, and $v_{s_3} = \{v_3\}$

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	S_D	Max $\mathcal{I}_2$	Partition
$(0.98, -0.02)$	0.0152	0.2726	0.1485	0.7275	-0.2191	2
$(0.98, -0.01)$	0.0151	0.2727	0.1486	0.7274	-0.2156	2
$(0.98, 0)$	0.0151	0.2728	0.1486	0.7271	-0.2114	2
$(0.98, 0.01)$	0.0151	0.2729	0.1487	0.728	-0.2378	5
$(0.98, 0.02)$	0.015	0.273	0.1487	0.727	-0.2096	2
$(0.99, -0.02)$	0.0088	0.2895	0.1589	0.7257	-0.2074	2
$(0.99, -0.01)$	0.0087	0.2896	0.159	0.7259	-0.2102	2
$(0.99, 0)$	0.0087	0.2898	0.159	0.7254	-0.2013	2
$(0.99, 0.01)$	0.0087	0.2899	0.1591	0.7257	-0.2079	2
$(0.99, 0.02)$	0.0086	0.29	0.1592	0.7262	-0.2188	2
$(1, -0.02)$	0.0001	0.321	0.1778	0.7216	-0.1806	2
$(1, -0.01)$	0.0001	0.3209	0.1774	0.7221	-0.1862	2
$(1, 0)$	0	0.3216	0.1781	0.7222	-0.1903	2
$(1, 0.01)$	0.0001	0.3219	0.1783	0.7231	-0.2133	2
$(1, 0.02)$	0.0001	0.3222	0.1785	0.7219	-0.187	2
$(1.01, -0.02)$	0.0049	46.5503	0.6396	0.7126	0.0719	5
$(1.01, -0.01)$	0.0048	46.5336	0.4892	0.6845	-0.0923	5
$(1.01, 0)$	0.0049	46.3831	0.6396	0.6964	0.2176	5
$(1.01, 0.01)$	0.0049	46.5854	0.6396	0.713	0.1123	2
$(1.01, 0.02)$	0.0049	47.0485	0.6396	0.7001	0.0916	2
$(1.02, -0.02)$	0.0095	95.475	0.4892	0.6825	-0.0558	2
$(1.02, -0.01)$	0.0095	94.7711	0.4892	0.6706	0.0538	2
$(1.02, 0)$	0.0095	95.5315	0.4892	0.6613	0.1127	2
$(1.02, 0.01)$	0.0098	96.0466	0.6396	0.7094	0.1586	2
$(1.02, 0.02)$	0.0095	94.8074	0.4892	0.6809	-0.0348	2

Table 14: Reconstruction error, mutual information, and disentanglement metric values for each (γ, λ) pair in Case 3 of Scenario 1.

2. $v_{s_1} = \{v_1\}$, $v_{s_2} = \{v_3\}$, and $v_{s_3} = \{v_2\}$
3. $v_{s_1} = \{v_2\}$, $v_{s_2} = \{v_1\}$, and $v_{s_3} = \{v_3\}$
4. $v_{s_1} = \{v_2\}$, $v_{s_2} = \{v_3\}$, and $v_{s_3} = \{v_1\}$
5. $v_{s_1} = \{v_3\}$, $v_{s_2} = \{v_1\}$, and $v_{s_3} = \{v_2\}$
6. $v_{s_1} = \{v_3\}$, $v_{s_2} = \{v_2\}$, and $v_{s_3} = \{v_1\}$

With the increase in the dimension of the latent variable to 3, the metric S_D becomes

$$S_D = \frac{1}{3} \sum_{i=1}^{3} \frac{\max_j |\text{cov}(x_i, v_j)|}{\sum_{j=1}^{3} |\text{cov}(x_i, v_j)|}. \tag{28}$$

Here x_i denotes the i-th component of the latent vector, while v_j represents the j-th component of the generative parameter vector.

Following the methodology outlined in Section 11.3.1, we examine three distinct cases as specified in Section 11.3. To ensure uniformity across all cases within the second scenario, we adopt consistent parameter values as presented in Table 10. Utilizing the information in this table, we calculate the covariance matrices $\boldsymbol{\Sigma_V}$ and $\boldsymbol{\Sigma_Y}$ for each case within Scenario 2. These computed covariance matrices align with those observed in Scenario 1, as documented in Table 11. For each pair of (γ, λ), we identify the optimal solution that

minimizes the reconstruction error. Subsequently, we present the following data in Tables 15, 16, and 17, corresponding to the three cases within Scenario 2: reconstruction error, encoder mutual information, mutual information between the generative variable $\boldsymbol{V}$ and the latent variable $\boldsymbol{X}$, the disentanglement metric S_D computed using equation (28), and the maximum value of $\mathcal{I}_3$ along with its associated partition determined by equation (27).

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V};\boldsymbol{X})$	S_D	Max $\mathcal{I}_3$	Partition
$(0.98, -0.02)$	0	0	0	0.3214	0	All 6 partitions
$(0.98, -0.01)$	0	0	0	0.3214	0	All 6 partitions
$(0.98, 0)$	0	0	0	0.3214	0	All 6 partitions
$(0.98, 0.01)$	0	0	0	0.3214	0	All 6 partitions
$(0.98, 0.02)$	0	0	0	0.3214	0	All 6 partitions
$(0.99, -0.02)$	0	0	0	0.3214	0	All 6 partitions
$(0.99, -0.01)$	0	0	0	0.3214	0	All 6 partitions
$(0.99, 0)$	0	0	0	0.3214	0	All 6 partitions
$(0.99, 0.01)$	0	0	0	0.3214	0	All 6 partitions
$(0.99, 0.02)$	0	0	0	0.3214	0	All 6 partitions
$(1, -0.02)$	0	0.0008	0.0002	0.3214	0.0002	5
$(1, -0.01)$	0	0.0014	0.0004	0.3214	0.0002	1, 3, 5, 6
$(1, 0)$	0	0.0048	0.0016	0.3213	0.0011	1, 3
$(1, 0.01)$	0	0.0031	0.0007	0.3214	0.0007	5, 6
$(1, 0.02)$	0	0.0087	0.0018	0.3213	0.0017	6
$(1.01, -0.02)$	0.004	68.4638	0.1419	0.3282	0.0702	3
$(1.01, -0.01)$	0.004	69.2841	0.1419	0.3282	0.0763	1
$(1.01, 0)$	0.005	68.8959	0.3347	0.3129	0.1557	2
$(1.01, 0.01)$	0.005	69.2016	0.3347	0.3119	0.2582	6
$(1.01, 0.02)$	0.004	68.3431	0.1419	0.3282	0.0706	2
$(1.02, -0.02)$	0.0079	143.5911	0.1419	0.3284	0.1251	1
$(1.02, -0.01)$	0.0079	142.3279	0.1419	0.3284	0.1266	6
$(1.02, 0)$	0.0079	142.5249	0.1419	0.3283	0.0775	4
$(1.02, 0.01)$	0.0099	143.319	0.2535	0.306	0.1468	3
$(1.02, 0.02)$	0.0079	142.7511	0.1419	0.3283	0.1035	5

Table 15: Reconstruction error, mutual information, and disentanglement metric values for each (γ, λ) pair in Case 1 of Scenario 2.

Based on these three tables, we arrive at the following insights:

- Table 15 presents numerical results for Case 1 within Scenario 2, assuming independence among all three generative parameters denoted as v_1, v_2, and v_3. In this case, all six listed partitions accurately describe the relationship among the generative parameters. Similar to Case 3 within Scenario 1, our goal is to determine the optimal partition among these partitions. However, since all listed partitions are permutation symmetries of each other, there is a unique optimal partition in this case. Moreover, if we want to precisely determine which latent parameter should be used to capture a specific generative parameter, it remains uncertain because the partition number varies each time we run a new numerical simulation for any combination of (γ, λ).

  Now, turning our attention to the metric S_D, we observe that for any combination of (γ, λ), all values of S_D consistently remain below $1/3$, which is the lower bound of S_D as defined. This consistent pattern suggests that using S_D in this case provides inaccurate results. According to [12], this metric is expected to perform well when the assumption is that each latent parameter should be correlated with only one generative parameter. In our scenario, all three generative parameters are independent, and the same holds for three independent latent parameters. Therefore, theoretically, each generative parameter can be effectively learned using exactly one latent parameter. However, numerical experiments have

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	S_D	Max $\mathcal{I}_3$	Partition
$(0.98, -0.02)$	0.0152	0.0789	0.02	0.2515	-0.3311	1
$(0.98, -0.01)$	0.0152	0.0789	0.02	0.2515	-0.3293	3
$(0.98, 0)$	0.0152	0.0789	0.02	0.2515	-0.3293	2
$(0.98, 0.01)$	0.0151	0.079	0.02	0.2515	-0.3303	4
$(0.98, 0.02)$	0.0151	0.079	0.02	0.2515	-0.33	3
$(0.99, -0.02)$	0.0075	0.0892	0.0224	0.2517	-0.3257	2
$(0.99, -0.01)$	0.0075	0.0892	0.0224	0.2517	-0.329	4
$(0.99, 0)$	0.0074	0.0892	0.0224	0.2517	-0.326	2
$(0.99, 0.01)$	0.0074	0.0893	0.0225	0.2517	-0.3295	6
$(0.99, 0.02)$	0.0074	0.0893	0.0225	0.2517	-0.3293	5
$(1, -0.02)$	0.0001	0.1041	0.0236	0.2518	-0.327	5
$(1, -0.01)$	0	0.0996	0.0242	0.2518	-0.3269	6
$(1, 0)$	0	0.0996	0.0241	0.2518	-0.3269	6
$(1, 0.01)$	0	0.1011	0.0258	0.252	-0.3241	2
$(1, 0.02)$	0.0001	0.1017	0.0273	0.2521	-0.3236	2
$(1.01, -0.02)$	0.0049	67.8445	0.2281	0.2651	-0.1644	2
$(1.01, -0.01)$	0.005	69.3147	0.4513	0.2528	-0.0722	5
$(1.01, 0)$	0.0049	69.4386	0.2281	0.265	-0.205	4
$(1.01, 0.01)$	0.0049	68.3424	0.2281	0.2653	-0.1583	2
$(1.01, 0.02)$	0.0049	68.8344	0.2281	0.2665	-0.1306	4
$(1.02, -0.02)$	0.0099	141.6833	0.4513	0.2596	-0.0297	5
$(1.02, -0.01)$	0.0099	142.2384	0.4513	0.2488	-0.1343	1
$(1.02, 0)$	0.0097	142.9661	0.2281	0.2652	-0.1548	5
$(1.02, 0.01)$	0.0099	143.6558	0.4259	0.2613	-0.0104	3
$(1.02, 0.02)$	0.0097	142.1888	0.2281	0.2654	-0.1842	6

Table 16: Reconstruction error, mutual information, and disentanglement metric values for each (γ, λ) pair in Case 2 of Scenario 2.

demonstrated that S_D fails to properly quantify disentanglement in our setup. In this context, it should ideally yield very high values, reaching 1.

- Table 16 presents numerical results for Case 2 within Scenario 2, assuming a linear dependency between generative parameters v_1 and v_2, while they remain independent from v_3. Similarly, Table 17 provides numerical results for Case 3 within Scenario 2, where we assume that generative parameters v_2 and v_3 are linearly dependent on v_1. In both cases, it is impossible to validly partition these three independent generative parameters into three distinct groups that satisfy the assumption. Although the six listed partitions in both cases fail to accurately describe the relationship among generative parameters, unlike Case 1 of Scenario 1, we have a consistent optimal partition across various combinations of (γ, λ) for both tables because all listed partitions are permutation symmetries of each other. However, if we fix the combination (γ, λ), the partition number will vary each time we run a new numerical simulation. Due to all listed partitions being permutation symmetries of each other and the varying value of the optimal partition for any given (γ, λ), we cannot determine the optimal way to map latent parameters to generative parameters using metric $\mathcal{I}_3$.

11.3.3 Summary of Numerical Results

In our examples, we have not considered the scenario where the dimension of the latent variable exceeds that of the generative variable. This is because it would lead to the presence of at least one redundant latent parameter that lacks any meaningful information about the generative variable. Therefore, we restrict our analysis to two cases: one where the dimension of the latent variable is less than or equal to the dimension

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	S_D	Max $\mathcal{I}_3$	Partition
$(0.98, -0.02)$	0.0152	0.2726	0.1485	0.3652	-0.9318	5
$(0.98, -0.01)$	0.0151	0.2727	0.1486	0.3653	-0.9408	2
$(0.98, 0)$	0.0151	0.2728	0.1486	0.3652	-0.9355	6
$(0.98, 0.01)$	0.0151	0.2729	0.1487	0.3652	-0.9364	1
$(0.98, 0.02)$	0.015	0.273	0.1487	0.3652	-0.9343	3
$(0.99, -0.02)$	0.0088	0.2895	0.1589	0.3646	-0.9194	5
$(0.99, -0.01)$	0.0087	0.2896	0.159	0.3647	-0.923	5
$(0.99, 0)$	0.0087	0.2898	0.159	0.3647	-0.9271	6
$(0.99, 0.01)$	0.0087	0.2899	0.1591	0.3647	-0.9275	3
$(0.99, 0.02)$	0.0086	0.29	0.1592	0.3643	-0.9083	1
$(1, -0.02)$	0.0001	0.3188	0.174	0.3641	-0.9229	6
$(1, -0.01)$	0.0001	0.3213	0.1778	0.3637	-0.9165	6
$(1, 0)$	0	0.3209	0.1771	0.3635	-0.9023	6
$(1, 0.01)$	0.0001	0.3219	0.1782	0.3635	-0.9055	5
$(1, 0.02)$	0.0001	0.3221	0.1783	0.3637	-0.9139	6
$(1.01, -0.02)$	0.0048	68.9475	0.4892	0.3493	-0.7088	1
$(1.01, -0.01)$	0.0049	69.7457	0.6396	0.3564	-0.636	2
$(1.01, 0)$	0.0048	70.2334	0.4892	0.3505	-0.7352	2
$(1.01, 0.01)$	0.0048	69.187	0.4892	0.3477	-0.6776	4
$(1.01, 0.02)$	0.0048	69.0806	0.4892	0.3513	-0.7359	3
$(1.02, -0.02)$	0.0098	145.0052	0.6396	0.3586	-0.6215	3
$(1.02, -0.01)$	0.0095	141.2475	0.4892	0.3503	-0.7176	6
$(1.02, 0)$	0.0095	143.3375	0.4892	0.3513	-0.7696	4
$(1.02, 0.01)$	0.0098	143.824	0.6396	0.3479	-0.4527	6
$(1.02, 0.02)$	0.0095	144.0293	0.4892	0.3503	-0.7456	1

Table 17: Reconstruction error, mutual information, and disentanglement metric values for each (γ, λ) pair in Case 3 of Scenario 2.

of the generative variable, as described in Sections 11.3.1 and 11.3.2.

Using the metrics $\mathcal{I}_2$ and $\mathcal{I}_3$ as examples, we can infer the following pros and cons that also apply to the disentanglement metric $\mathcal{I}_m$, where m represents the dimension of the latent variable.

- **Pros**: This metric proves to be highly effective when there exists only one partition that accurately describes the relationship among generative parameters, as demonstrated in Case 2 of Scenario 1. Permutation symmetries of this partition are not considered as separate solutions. Moreover, even in cases where multiple correct partitions exist, and if not all these partitions are permutation symmetries of others, as demonstrated in Case 3 of Scenario 1, then metric $\mathcal{I}_m$ can identify the exact optimal partition from a set of multiple correct partitions.

- **Cons**: In cases where $\mathcal{I}_m$ can be effectively applied to disentangle, its first limitation pertains to potential computational complexity, especially when dealing with high-dimensional generative and latent variables. This increased complexity arises from the possible proliferation of partitions and the growing intricacy of the formula used to compute $\mathcal{I}_m$.

  The next limitation arises in situations where $\mathcal{I}_m$ struggles or cannot effectively assess disentanglement. Firstly, $\mathcal{I}_m$ fails to evaluate disentanglement for cases where the dimension of the latent variable is less than the dimension of the generative variable, and all generative parameters are independent, as observed in Case 1 of Scenario 1. In these scenarios, the optimal partition varies across different combinations of (γ, λ). Even when a specific (γ, λ) is fixed, conducting numerical investigations multiple times results in varying optimal partition numbers. Consequently, it is inappropriate to conclude that $\mathcal{I}_m$ consistently reaches its maximum value at a fixed partition for a specific combination of (γ, λ).

Secondly, using $\mathcal{I}_m$ to evaluate disentanglement is ineffective when the dimension of the latent and generative variables are the same, and there exists dependency among generative parameters. In these cases, we cannot determine the optimal way to map latent parameters to generative parameters, as all listed partitions are permutation symmetries of each other, as demonstrated in Cases 2 and 3 in Scenario 2.

11.4 Extension of Mutual Information-Based Metric $\mathcal{I}_3$

Upon examining Tables 15, 16, and 17 obtained in Section 11.3.2, it is evident that the potential of partitioning has not been fully utilized in generating meaningful numerical results. The manner in which we partition the three generative parameters into three distinct groups, adhering to the conditions outlined in (26), consistently leads to a set of partitions, all of which exhibit permutation symmetries with respect to each other. Consequently, this results in a unique optimal partition where the metric $\mathcal{I}_3$ attains its maximum value. To address this issue, it is necessary to consider an extension of the $\mathcal{I}_3$ metric in which each group of generative parameters v_{s_m}, for $m \in \{1, 2, 3\}$, can be empty. This extension aims to eliminate permutation symmetries among all partitions and generate more meaningful numerical results. In essence, this enables scenarios where not all available latent parameters are required to capture information from the generative parameters.

We continue to examine the scenario in which both the generative and latent variables have dimensions equal to 3, as previously discussed in Section 11.3.2. However, we will explore the partitioning of generative parameters while considering the possibility that not all three latent parameters are necessary for capturing information from these three generative parameters. This exploration leads to the consideration of three distinct cases:

- **Case 1: None of the groups are empty**

  In this case, we have a total of 6 ways to partition the 3 generative parameters into 3 distinct groups, with each group containing exactly 1 generative parameter. This partitioning method has been previously discussed in Section 11.3.2.

- **Case 2: One group is empty**

  In this case, only 2 out of the 3 latent parameters, denoted as x_i and x_j, where $i, j \in \{1, 2, 3\}$ and $i \neq j$, are utilized to capture information from the 3 generative parameters. This leaves the third latent parameter unused for other purposes. Consequently, we can partition the set of 3 generative parameters, represented as $\mathcal{V}_3 = \{v_1, v_2, v_3\}$, into 3 distinct groups, with 2 of these groups being non-empty. Specifically, each non-empty group is associated with one of the utilized latent parameters and is labeled as v_{s_i} and v_{s_j}, respectively. On the other hand, the remaining group, denoted as v_{s_k}, corresponds to the unused latent parameter x_k, where $k \in \{1, 2, 3\}$ and $k \notin \{i, j\}$, and this group will be empty. For distinct values of i, j, and k chosen from the set $\{1, 2, 3\}$, the system of conditions (26) can be modified as follows:

$$\begin{aligned} &v_{s_k} = \varnothing \\ &v_{s_i} \subset \mathcal{V}_3, v_{s_j} \subset \mathcal{V}_3 \\ &v_{s_i} \cup v_{s_j} = \mathcal{V}_3 \\ &v_{s_i} \cap v_{s_j} = \varnothing. \end{aligned} \tag{29}$$

- **Case 3: Two groups are empty**

  In this case, only 1 out of the 3 latent parameters are used to capture information from the 3 generative parameters, leaving the two other latent parameters unused for other purposes. In this context, the set of 3 generative parameters can still be partitioned into 3 distinct groups. But there is an exactly 1 group is non-empty that contains all 3 generative parameters.

We now consider the complete list of potential partitions for Cases 2 and 3.

- **Case 2: One group is empty**

Suppose latent parameter x_1 is not utilized, resulting in an empty associated group of generative parameters, v_{s_1}. In such a scenario, there are 6 possible ways to partition the three generative parameters into three distinct groups that satisfy the system of conditions (29).

1. $v_{s_1} = \varnothing$, $v_{s_2} = \{v_1\}$, and $v_{s_3} = \{v_2, v_3\}$
2. $v_{s_1} = \varnothing$, $v_{s_2} = \{v_2\}$, and $v_{s_3} = \{v_1, v_3\}$
3. $v_{s_1} = \varnothing$, $v_{s_2} = \{v_3\}$, and $v_{s_3} = \{v_1, v_2\}$
4. $v_{s_1} = \varnothing$, $v_{s_2} = \{v_1, v_2\}$, and $v_{s_3} = \{v_3\}$
5. $v_{s_1} = \varnothing$, $v_{s_2} = \{v_1, v_3\}$, and $v_{s_3} = \{v_2\}$
6. $v_{s_1} = \varnothing$, $v_{s_2} = \{v_2, v_3\}$, and $v_{s_3} = \{v_1\}$

Given that 1 out of the 3 latent parameters will not be used, it is possible for both v_{s_2} and v_{s_3} to be empty as well. Consequently, there are a total of 18 distinct ways to partition the 3 generative parameters in this case.

- **Case 3: Two groups are empty**

  In this case, since 2 groups are empty, one group must contain all generative parameters, resulting in the following list of 3 partitions.

  1. $v_{s_1} = \{v_1, v_2, v_3\}$, $v_{s_2} = \varnothing$, and $v_{s_3} = \varnothing$
  2. $v_{s_1} = \varnothing$, $v_{s_2} = \{v_1, v_2, v_3\}$, and $v_{s_3} = \varnothing$
  3. $v_{s_1} = \varnothing$, $v_{s_2} = \varnothing$, and $v_{s_3} = \{v_1, v_2, v_3\}$

Combining the 6 partitions from Case 1 with the 21 partitions from Cases 2 and 3, there are a total of 27 partitions when partitioning the 3 generative parameters into 3 distinct groups, considering scenario where not all latent parameters must be used to capture information from the given generative parameters.

To assess disentanglement in this scenario, we introduce a modified information-based metric denoted as $\tilde{\mathcal{I}}_3$. For distinct values of i, j, and k chosen from the set $\{1, 2, 3\}$, this metric is defined as follows:

$$\tilde{\mathcal{I}}_3 = \begin{cases} \sum_{m=1}^{3} I(x_m; v_{s_m}) - \sum_{i<j} I(v_{s_i}; v_{s_j}) & \text{if } v_{s_i}, v_{s_j}, v_{s_k} \neq \varnothing \\ I(x_i; v_{s_i}) + I(x_j; v_{s_j}) - I(v_{s_i}; v_{s_j}) & \text{if } v_{s_i}, v_{s_j} \neq \varnothing \text{ and } v_{s_k} = \varnothing \\ I(x_i; v_{s_i}) & \text{if } v_{s_i} \neq \varnothing \text{ and } v_{s_j} = v_{s_k} = \varnothing. \end{cases} \tag{30}$$

11.4.1 Numerical Results

To investigate disentanglement using the metric $\tilde{\mathcal{I}}_3$ under three scenarios, as described in Section 11.3, which include the independence of 3 generative parameters, linear dependence of v_1 and v_2 with independence of v_3, and linear dependence of v_2 and v_3 on v_1, we need to establish parameter values that will remain consistent across all these scenarios. We will reuse the parameter values presented in Table 10 in Section 11.3.2, which will yield the same values for $\boldsymbol{\Sigma_V}$ and $\boldsymbol{\Sigma_Y}$, as detailed in Table 11.

Applying the methodology described in Section 11.2.2, we have derived numerical results for three scenarios characterizing various relationships among generative parameters. These results are displayed in Tables 18, 19, and 20. The tables include information regarding reconstruction error, encoder mutual information, mutual information between the generative variable $\boldsymbol{V}$ and the latent variable $\boldsymbol{X}$, as well as the maximum value of $\tilde{\mathcal{I}}_3$ alongside its corresponding partition determined by equation (30).

- **Scenario 1: Independence of Generative Parameters**

  In this scenario, the task is to partition 3 independent generative parameters into 3 distinct groups, such that each group contains precisely 1 generative parameter. Consequently, we have identified 6 partitions that accurately describe the relationships among the given generative parameters in this context, specifically denoted as Partitions 12 to 17. It is worth noting that since each group consists of only 1 element, all of these 6 partitions are permutation symmetries of each other.

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	Max $\tilde{\mathcal{I}}_3$	Partition
$(0.98, -0.02)$	0	0	0	0	All partitions
$(0.98, -0.01)$	0	0	0	0	All partitions
$(0.98, 0)$	0	0	0	0	All partitions
$(0.98, 0.01)$	0	0	0	0	All partitions
$(0.98, 0.02)$	0	0	0	0	All partitions
$(0.99, -0.02)$	0	0	0	0	All partitions
$(0.99, -0.01)$	0	0	0	0	All partitions
$(0.99, 0)$	0	0	0	0	All partitions
$(0.99, 0.01)$	0	0	0	0	All partitions
$(0.99, 0.02)$	0	0	0	0	All partitions
$(1, -0.02)$	1.074e-05	0.0007	0.0001	5.323e-05	11
$(1, -0.01)$	2.93e-06	0.0009	0.0001	7.08e-05	15
$(1, 0)$	1.497e-05	0.0031	0.0008	0.0007	14
$(1, 0.01)$	3.26e-06	0.0405	0.0140	0.0081	17
$(1, 0.02)$	5.23e-06	0.0057	0.0006	0.0003	2
$(1.01, -0.02)$	0.0039	68.6932	0.1418	0.0708	21
$(1.01, -0.01)$	0.0039	69.6490	0.1418	0.1030	3
$(1.01, 0)$	0.0039	69.4459	0.1418	0.1390	9
$(1.01, 0.01)$	0.0049	68.9825	0.2534	0.2080	16
$(1.01, 0.02)$	0.0049	69.3505	0.3347	0.2196	4
$(1.02, -0.02)$	0.0098	142.4277	0.3650	0.2242	17
$(1.02, -0.01)$	0.0098	143.6928	0.2534	0.2101	13
$(1.02, 0)$	0.0098	144.2456	0.3650	0.2819	12
$(1.02, 0.01)$	0.0098	142.2954	0.2534	0.1830	13
$(1.02, 0.02)$	0.0098	142.3931	0.3650	0.1672	7

Table 18: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 1.

For each pair of (γ, λ), we determine the maximal value of the metric $\tilde{\mathcal{I}}_3$ by examining all 27 possible ways to partition the 3 generative parameters into 3 distinct groups. The partition corresponding to the maximum $\tilde{\mathcal{I}}_3$ value is considered the optimal partition for that specific pair of (γ, λ). If this optimal partition falls within the range of Partitions 12 to 17, we will highlight them in red within Tables 18. This highlighting indicates that the metric $\tilde{\mathcal{I}}_3$ has effectively disentangled the relationship among the given generative parameters for the given pair of (γ, λ).

It's important to observe that the metric $\tilde{\mathcal{I}}_3$ has not been able to disentangle the relationship among these 3 given generative parameters for all combinations of (γ, λ). Specifically, in the scenario where all generative parameters are independent, we can only achieve a successful disentanglement in 32% of cases.

- **Scenario 2: Linear Dependence of v_1 and v_2, with Independence of v_3**

  Given that 2 generative parameters, v_1 and v_2, exhibit linear dependence, they should be grouped together, while the independent generative parameter v_3 should be allocated to a separate group. Since only 2 latent parameters are necessary to capture information from these 3 generative parameters, there must be precisely 1 group that remains empty.

  In this scenario, there exist 6 partitions that accurately represent the relationships among the 3 given generative parameters. These partitions are Partitions 4, 5, 11, 20, 21, and 24. Each of these partitions consists of exactly 1 empty group, 1 group containing $\{v_1, v_2\}$, and the remaining group containing v_3.

  Furthermore, notice that Partition 4 is a permutation symmetry of Partition 5, Partition 11 is a permutation symmetry of Partition 21, and Partition 20 is a permutation symmetry of Partition 24.

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	Max $\tilde{\mathcal{I}}_3$	Partition
$(0.98, -0.02)$	0.0152	0.0788	0.0199	0.0157	21
$(0.98, -0.01)$	0.0151	0.0788	0.0199	0.0187	1
$(0.98, 0)$	0.0151	0.0789	0.0199	0.0097	21
$(0.98, 0.01)$	0.0151	0.0789	0.0200	0.0110	1
$(0.98, 0.02)$	0.0151	0.0790	0.0200	0.0133	21
$(0.99, -0.02)$	0.0074	0.0891	0.0224	0.0123	1
$(0.99, -0.01)$	0.0074	0.0892	0.0224	0.0146	21
$(0.99, 0)$	0.0074	0.0892	0.0224	0.0129	1
$(0.99, 0.01)$	0.0073	0.0892	0.0224	0.0130	1
$(0.99, 0.02)$	0.0073	0.0893	0.0224	0.0148	21
$(1, -0.02)$	5.384e-05	0.0996	0.0238	0.0178	24
$(1, -0.01)$	2.692e-05	0.3247	0.0943	0.0833	21
$(1, 0)$	2.25e-06	0.0999	0.0245	0.0164	4
$(1, 0.01)$	2.636e-05	0.1436	0.0267	0.0154	24
$(1, 0.02)$	5.293e-05	0.1007	0.0264	0.0231	4
$(1.01, -0.02)$	0.0048	68.3902	0.2281	0.1821	21
$(1.01, -0.01)$	0.0049	68.9910	0.2534	0.1287	11
$(1.01, 0)$	0.0049	69.3434	0.4258	0.3160	11
$(1.01, 0.01)$	0.0049	68.8717	0.4512	0.2658	4
$(1.01, 0.02)$	0.0049	69.9942	0.4258	0.2192	20
$(1.02, -0.02)$	0.0098	142.8238	0.4512	0.3116	4
$(1.02, -0.01)$	0.0098	140.5503	0.4512	0.1857	8
$(1.02, 0)$	0.0098	142.6699	0.4512	0.2874	21
$(1.02, 0.01)$	0.0098	142.9741	0.4512	0.2830	24
$(1.02, 0.02)$	0.0098	142.8761	0.4512	0.3799	24

Table 19: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 2.

Similarly to Scenario 1, the partition associated with the maximum value of $\tilde{\mathcal{I}}_3$ is considered the optimal partition for a specific pair of (γ, λ). If this optimal partition aligns with the partitions that accurately describe the relationship among the 3 generative parameters in Scenario 2, as listed above, we will highlight them in red within Tables 19. This highlighting serves to indicate that the metric $\tilde{\mathcal{I}}_3$ has effectively disentangled the relationship among the given generative parameters for the specified pair of (γ, λ).

We observe that the metric $\tilde{\mathcal{I}}_3$ has not been able to disentangle the relationship among these 3 given generative parameters for all combinations of (γ, λ). However, the probability of discovering the correct optimal partition in Scenario 2, where 2 out of 3 generative parameters are linearly dependent, is much higher compared to that of Scenario 1, where no dependency exists among all generative parameters. Specifically, in the scenario where v_1 and v_2 are linearly dependent on each other while independent from v_3, we achieve a successful disentanglement rate of 76%, significantly higher than the 32% success rate in Scenario 1.

- **Scenario 3: Linear Dependence of v_2 and v_3 on v_1**

  Since all 3 generative parameters are linearly dependent, they should be grouped together. In this case, only 1 latent parameter is needed to capture information from these generative parameters, leaving 2 latent parameters unused. Consequently, 2 groups are empty. Therefore, we have 3 partitions that accurately describe the relationship among the 3 given generative parameters in this scenario: Partitions 1, 8, and 27. Notice that all these 3 partitions are permutation symmetries of each other.

  Similarly to the two previous scenarios, we will highlight the optimal partition that accurately describes

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	Max $\tilde{\mathcal{I}}_3$	Partition
$(0.98, -0.02)$	0.0151	0.2726	0.1485	0.0999	1
$(0.98, -0.01)$	0.0151	0.2727	0.1485	0.1415	27
$(0.98, 0)$	0.0151	0.2728	0.1486	0.1022	1
$(0.98, 0.01)$	0.0150	0.2728	0.1486	0.1064	1
$(0.98, 0.02)$	0.0150	0.2729	0.1487	0.0900	27
$(0.99, -0.02)$	0.0087	0.2895	0.1588	0.1110	1
$(0.99, -0.01)$	0.0087	0.2896	0.1589	0.1209	1
$(0.99, 0)$	0.0086	0.2897	0.1590	0.1393	1
$(0.99, 0.01)$	0.0086	0.2898	0.1591	0.1426	27
$(0.99, 0.02)$	0.0086	0.2900	0.1591	0.0926	1
$(1, -0.02)$	0.0001	0.3192	0.1747	0.1377	1
$(1, -0.01)$	5.948e-05	0.3204	0.1767	0.1567	27
$(1, 0)$	1e-06	0.3208	0.1771	0.0887	27
$(1, 0.01)$	6.342e-05	0.3261	0.1782	0.1437	1
$(1, 0.02)$	0.0001	0.3236	0.1788	0.1406	1
$(1.01, -0.02)$	0.0049	70.0361	0.6395	0.3969	27
$(1.01, -0.01)$	0.0049	70.3412	0.6395	0.4916	1
$(1.01, 0)$	0.0049	70.5561	0.6395	0.2711	1
$(1.01, 0.01)$	0.0049	68.8737	0.6395	0.2494	1
$(1.01, 0.02)$	0.0049	69.2710	0.6395	0.2580	8
$(1.02, -0.02)$	0.0097	142.4882	0.6395	0.2595	1
$(1.02, -0.01)$	0.0097	142.9027	0.6395	0.3408	1
$(1.02, 0)$	0.0094	143.0366	0.4891	0.3174	1
$(1.02, 0.01)$	0.0094	143.0789	0.4891	0.4263	8
$(1.02, 0.02)$	0.0097	141.7545	0.6395	0.4738	1

Table 20: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 3.

the relationship among the 3 generative parameters in this scenario in red within Tables 20. This highlighting serves to indicate that the metric $\tilde{\mathcal{I}}_3$ has effectively disentangled the relationship among the given generative parameters for the specified pair of (γ, λ).

It's noteworthy that the metric $\tilde{\mathcal{I}}_3$ has been able to disentangle the relationship among these 3 given generative parameters for all combinations of (γ, λ). This demonstrates that $\tilde{\mathcal{I}}_3$ is highly effective for scenarios where all generative parameters are fully correlated.

11.5 Comparison of Disentanglement Metrics

11.5.1 Information-based Metrics

In this section, our goal is to establish a robust framework for evaluating the effectiveness of our proposed disentanglement metric $\tilde{\mathcal{I}}_3$ in comparison to other contemporary metrics. Our metric is grounded in mutual information theory, providing a foundational basis for comparison with metrics sharing this theoretical foundation. One such metric for comparison is the **Mutual Information Gap (MIG)**, introduced by Chen et al. [11]. Both our proposed metric $\tilde{\mathcal{I}}_3$ and MIG fall under the category of information-based metrics, utilizing principles of information theory to calculate a disentanglement score by estimating the mutual information between generative variables $\boldsymbol{V}$ and latent variables $\boldsymbol{X}$.

The MIG score is built upon the property of *compactness*, asserting that the subset of the latent dimensions affected by a generative factor should be as small as possible [13]. Ideally, only one latent dimension completely describes a generative factor. This property stems from the rationale that if one latent variable effectively models a ground truth generative factor, there is no necessity for other latent variables to redun-

dantly convey information about the same generative factor. Based on the assumption of discrete generative factors and independent latent variables, the computation of the MIG score involves the following steps:

1. Calculate the mutual information between each factor v_j and every latent dimension $\boldsymbol{X}$.
2. Classify latent dimensions based on the extent of information they encode about a specific generative factor, recognizing that each generative factor may have high mutual information with multiple latent dimensions.
3. Determine the difference between the top two highest values of mutual information for each generative factor; this difference defines the *gap*.
4. Normalize the gap by dividing it by the entropy of the corresponding generative factor. The MIG score is then determined by the formula:

$$\mathrm{MIG}(\boldsymbol{V}, \boldsymbol{X}) = \frac{1}{s} \sum_{j=1}^{s} \frac{1}{H(v_j)} \left[I(x_{i^{(j)}}; v_j) - I(x_{i'^{(j)}}; v_j) \right], \tag{31}$$

   where $i^{(j)} = \arg\max_i I(x_i; v_j)$ and $i'^{(j)} = \arg\max_{i \neq i^{(j)}} I(x_i; v_j)$ represent the indices associated with the maximum and second-maximum values of $I(x_i; v_j)$, respectively; and s denotes the number of known ground truth generative factors.

The MIG score is defined as the average, normalized difference (*gap*) between the top two latent variables with the highest mutual information. In an ideal disentangled representation, a single latent dimension should significantly encode the information of a specific generative factor, and the second most significant dimension should possess a small mutual information value, leading to a high gap. Consequently, a higher MIG score serves as an indicator of better disentanglement.

11.5.2 Predictor-based Metrics

Another disentanglement metric that can be used to compare with our proposed metric $\tilde{\mathcal{I}}_3$ is **Separated Attribute Predictability (SAP)**, introduced by Kumar et al. [10]. SAP falls within the category of predictor-based metrics, utilizing regressors or classifiers to predict generative factors from latent dimensions. The predictor's analysis is then employed to assess the usefulness of each latent dimension in predicting the generative factors [13].

For continuous generative factors, the R^2 score of linear regression is computed to predict generative factor values from each dimension of the learned representation. In the case of discrete generative factors, a linear classifier is trained. The SAP score is then calculated as the average difference in prediction errors between the two most predictive latent dimensions for each generative factor. This concept is analogous to the gap idea in the MIG metric. However, SAP holds an advantage over MIG as it is applicable to both continuous and discrete generative factors.

To apply this metric under the assumption of continuous generative variables, the following steps are undertaken:

1. Construct a score matrix $\mathbf{S}$ with dimensions $m \times s$, where m is the number of latent variables and s is the number of generative factors. Each ij-th entry of matrix $\mathbf{S}$, denoted by $S_{i,j}$, quantifies how well the j-th generative factor is predicted using only the i-th latent variable. The R^2 score of the regression, computed as follows, determines $S_{i,j}$:

$$S_{i,j} = \left[\frac{\mathrm{cov}(x_i, v_j)}{\sqrt{\mathrm{var}(x_i)}\sqrt{\mathrm{var}(v_j)}} \right]^2. \tag{32}$$

   This R^2 score ranges from 0 to 1, where a score of 1 indicates that a linear function of the i-th latent variable explains all the variability in the j-th generative factor. For inactive latent dimensions, $S_{i,j}$ is set to 0.

2. Once all values of the score matrix $\mathbf{S}$ are computed, calculate the difference between the top two entries with the highest scores for each column of the matrix, corresponding to the top two most predictive latent dimensions.

3. The SAP score is computed as the mean of the s differences and can be expressed as follows:

$$\mathrm{SAP}(\boldsymbol{V}, \boldsymbol{X}) = \frac{1}{s} \sum_{j=1}^{s} \left(S_{i^{(j)},j} - S_{i'^{(j)},j} \right). \tag{33}$$

   Here, $i^{(j)} = \arg\max_i S_{i,j}$ and $i'^{(j)} = \arg\max_{i \neq i^{(j)}} S_{i,j}$ represent the indices associated with the highest and second-highest scores for generative factor v_j, respectively.

Achieving a high SAP score involves evaluating the difference between the two highest entries in each score matrix column. This criterion dictates the need for precisely one high-scoring entry per column to achieve a high SAP score. A high SAP score suggests that each generative factor is primarily captured by only one latent variable, emphasizing a more distinct and isolated representation of generative factors. Therefore, a higher SAP score typically corresponds to a more effective disentanglement of generative factors from the latent variables.

11.6 Comparison of $\tilde{\mathcal{I}}_3$ and SAP in a Linear Gaussian Framework

11.6.1 $(s, n, m) = (3, 4, 3)$

Note that the MIG metric relies on the assumption of discrete ground truth generative factors. Consequently, the MIG metric is not applicable when addressing scenarios with continuous generative factors. If applied in such cases, the entropy of each generative factor may yield negative values, resulting in a negative MIG score. This departure from the defined range $[0, 1]$, as outlined in equation (31), highlights the metric's limitation in handling continuous generative factors. Conversely, both our proposed metric $\tilde{\mathcal{I}}_3$ and the SAP metric can effectively handle scenarios involving continuous generative factors. Given that our numerical simulations thus far involve a linear Gaussian setting using continuous variables, we will specifically evaluate the effectiveness of the SAP metric in studying disentanglement compared to $\tilde{\mathcal{I}}_3$ under the Gaussian linear setting. Notably, the MIG metric will be assessed in comparison with the $\tilde{\mathcal{I}}_3$ metric using discrete variables, utilizing a real dataset for scenarios involving non-Gaussian distributions.

We revisit the three scenarios detailed in Section 11.3: independence of generative factors, partial correlation among generative factors, and complete correlation among generative factors. Our aim is to assess the effectiveness of the SAP metric in comparison with $\tilde{\mathcal{I}}_3$ under each scenario, discerning the advantages and limitations of each metric. To achieve this objective, we conduct numerical simulations under the assumption of (s, n, m), corresponding to the dimensions of generative variables (3), input data (4), and latent variables (3), respectively. We continue to examine 25 pairs of (γ, λ) from two arrays: $\gamma = \{0.98, 0.99, 1, 1.01, 1.02\}$ and $\lambda = \{-0.02, -0.01, 0, 0.01, 0.02\}$. For each pair of (γ, λ) across the three scenarios, we investigate whether each metric successfully captures disentanglement. In instances where the metric successfully identifies disentanglement, we denote it with a check mark. In cases where the metric does not, we leave the result tables blank.

1. **Scenario 1: Independence of Generative Factors**

   In this scenario, where three given generative factors are independent, and there are three latent variables in total, each latent variable should fully capture information related to one generative factor.

   Table 21 provides comprehensive information, including the reconstruction error, encoder mutual information, mutual information between the generative variable $\boldsymbol{V}$ and the latent variable $\boldsymbol{X}$, the $\tilde{\mathcal{I}}_3$ score obtained using the metric $\tilde{\mathcal{I}}_3$, the success of disentanglement as determined by the metric $\tilde{\mathcal{I}}_3$, the SAP score obtained using the metric SAP, and the success of disentanglement determined by the SAP metric.

   - For each pair of (γ, λ), disentanglement success for $\tilde{\mathcal{I}}_3$ is attained when it identifies the partition yielding the maximum value of $\tilde{\mathcal{I}}_3$ among all 27 partitions. In this specific partition, all three

generative factors must be divided into three distinct groups. Each group, associated with exactly one latent variable, contains exactly one factor. This implies that each latent variable is used to capture information related to precisely one generative factor.

- Determining the success of disentanglement for the SAP metric involves several steps. Initially, we utilize the score matrix $\mathbf{S}$ with dimensions 3×3, where each ij-th entry $S_{i,j}$ is calculated using equation (32). As each column j of the matrix $\mathbf{S}$ corresponds to the j-th generative factor and contains three rows of values, we extract the row index i of the entry with the highest value. This index represents the index of latent variable that should be utilized to capture information about the specific factor under consideration. In Scenario 1, to ensure a fair comparison of disentanglement with the $\tilde{\mathcal{I}}_3$ metric, we follow the same criterion: each factor must be captured by a distinct latent variable. This implies that we indicate the SAP metric has successfully achieved disentanglement when the latent variable index associated with the entry having the highest value in each column of the generative factor is unique.

By Table 21, the metric $\tilde{\mathcal{I}}_3$ shows ineffective for studying disentanglement when all generative factors are independent. Specifically, in this scenario, the successful disentanglement rate is only 32%. In contrast, the SAP metric appears more effective in studying disentanglement, demonstrating a success rate of 44%.

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	$\tilde{\mathcal{I}}_3$ Score	$\tilde{\mathcal{I}}_3$ Success	SAP Score	SAP Success
$(0.98, -0.02)$	0	0	0	0		0	
$(0.98, -0.01)$	0	0	0	0		0	
$(0.98, 0)$	0	0	0	0		0	
$(0.98, 0.01)$	0	0	0	0		0	
$(0.98, 0.02)$	0	0	0	0		0	
$(0.99, -0.02)$	0	0	0	0		0	
$(0.99, -0.01)$	0	0	0	0		0	
$(0.99, 0)$	0	0	0	0		0	✓
$(0.99, 0.01)$	0	0	0	0		0	
$(0.99, 0.02)$	0	0	0	0		0	
$(1, -0.02)$	2.045e-05	0.0004	0.0001	0.0001		0	
$(1, -0.01)$	3.443e-05	0.0003	0	0		0	
$(1, 0)$	1.557e-05	0.0023	0.0003	0.0002	✓	0.0001	✓
$(1, 0.01)$	1.2e-06	0.0269	0.0012	0.0005		0.0001	
$(1, 0.02)$	2.795e-05	0.0337	0.0067	0.0056	✓	0.003	✓
$(1.01, -0.02)$	0.0050	69.5785	0.365	0.2628	✓	0.1105	✓
$(1.01, -0.01)$	0.0050	68.5482	0.365	0.2161	✓	0.0652	✓
$(1.01, 0)$	0.0050	69.1519	0.365	0.3091	✓	0.1608	✓
$(1.01, 0.01)$	0.0050	69.2936	0.365	0.3066	✓	0.1557	✓
$(1.01, 0.02)$	0.004	68.5367	0.1419	0.0614		0.0066	✓
$(1.02, -0.02)$	0.0099	143.457	0.2535	0.1408		0.0312	
$(1.02, -0.01)$	0.0099	144.2767	0.2535	0.208	✓	0.0971	
$(1.02, 0)$	0.0099	143.0818	0.2535	0.2137		0.1072	✓
$(1.02, 0.01)$	0.0099	142.418	0.2535	0.1997		0.0896	✓
$(1.02, 0.02)$	0.0099	143.1381	0.365	0.3353	✓	0.1811	✓

Table 21: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 1.

By observing Figure 11, it is noticeable that $\tilde{\mathcal{I}}_3$ and SAP scores across various (γ, λ) pairs exhibit a consistent pattern. Specifically, as $\tilde{\mathcal{I}}_3$ increases, the SAP score also increases, and conversely, when $\tilde{\mathcal{I}}_3$ decreases, the SAP score tends to decrease as well.

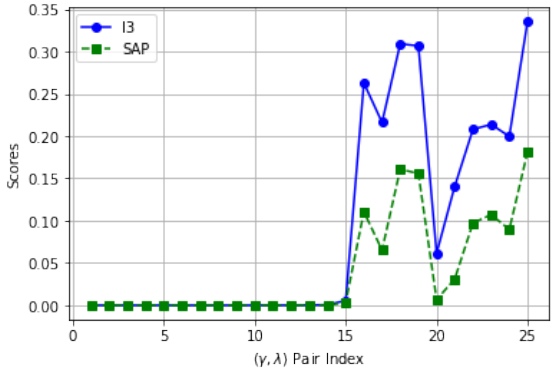

Figure 11: $\mathcal{I}_3$ vs. SAP scores for each (γ, λ) pair in Scenario 1.

2. **Scenario 2: Linear Dependence of v_1 and v_2, with Independence of v_3**

   Given that two generative factors, v_1 and v_2, exhibit linear dependence, they should be captured using the same latent variable, while the independent generative factor v_3 should be studied by another latent variable. Thus, one latent variable remains unused.

   - Similar to Scenario 1, disentanglement success for $\tilde{\mathcal{I}}_3$ for each pair of (γ, λ) in Scenario 2 is achieved by identifying the partition that yields the maximum value of $\tilde{\mathcal{I}}_3$ among all 27 partitions. However, in this specific partition for the second scenario, the three generative factors are distributed across two distinct groups associated with two distinct latent variables, and the third group is empty, indicating that one latent variable remains unused. Specifically, the group with two correlated generative factors must be v_1 and v_2, while the second group contains the remaining independent generative factor, which must be v_3.
   - We employ similar steps as in Scenario 1 to assess the success of disentanglement for the SAP metric in Scenario 2. To ensure a fair comparison when evaluating disentanglement with the $\tilde{\mathcal{I}}_3$ metric, we deem the SAP metric successful in achieving disentanglement under the same criterion: two correlated generative factors must be captured using a single latent variable, while the independent factor is represented by the second latent variable. This implies that the latent variable index of the entry with the highest value in the first two columns of matrix $\mathbf{S}$ must be the same, while the latent variable index of the entry with the highest value in the third column must be different. Each column of matrix $\mathbf{S}$ represents a single generative factor, so having the same latent variable indices in the first two columns implies that the two generative factors v_1 and v_2 should be associated with the same latent variable, while the last generative factor v_3 can be learned using the second latent variable.

   Examining Table 22, the metric $\tilde{\mathcal{I}}_3$ is quite effective for studying disentanglement when partial correlation among generative factors exists. Specifically, in the second scenario, the successful disentanglement rate is 68%, which is much higher than 32% in the first scenario when no correlation exists. Similar to the first scenario, the SAP metric also appears more effective in studying disentanglement in this case, demonstrating a success rate of 92%.

3. **Scenario 3: Linear Dependence of v_2 and v_3 on v_1**

   Since all three generative factors are fully correlated, only one latent variable is needed to capture information from these generative factors, leaving two latent variables unused.

   - Similar to the first two scenarios, disentanglement success for $\tilde{\mathcal{I}}_3$ in Scenario 3 is achieved for each pair of (γ, λ) when the simulation returns the partition that yields the maximum value of $\tilde{\mathcal{I}}_3$ among all 27 partitions. However, in this specific partition for the third scenario, all three generative factors are grouped together in a single group.

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	$\tilde{\mathcal{I}}_3$ Score	$\tilde{\mathcal{I}}_3$ Success	SAP Score	SAP Success
$(0.98, -0.02)$	0.0152	0.0789	0.02	0.0152		0.0194	✓
$(0.98, -0.01)$	0.0152	0.0789	0.02	0.0175		0.0261	✓
$(0.98, 0)$	0.0152	0.0789	0.02	0.0186		0.0296	✓
$(0.98, 0.01)$	0.0151	0.079	0.02	0.0199	✓	0.034	✓
$(0.98, 0.02)$	0.0151	0.079	0.02	0.0132		0.0151	✓
$(0.99, -0.02)$	0.0075	0.0892	0.0224	0.0104		0.0048	✓
$(0.99, -0.01)$	0.0075	0.0892	0.0224	0.0193	✓	0.0289	✓
$(0.99, 0)$	0.0074	0.0892	0.0224	0.0132		0.0088	✓
$(0.99, 0.01)$	0.0074	0.0893	0.0225	0.015		0.0133	✓
$(0.99, 0.02)$	0.0074	0.0893	0.0225	0.02		0.0321	✓
$(1, -0.02)$	5.371e-05	0.1004	0.0239	0.0175	✓	0.0225	✓
$(1, -0.01)$	2.69e-05	0.0999	0.025	0.0138	✓	0.0054	✓
$(1, 0)$	4.13e-06	0.1004	0.0248	0.0149	✓	0.0091	✓
$(1, 0.01)$	3.561e-05	0.0996	0.0256	0.0205	✓	0.0288	✓
$(1, 0.02)$	4.966e-05	0.1668	0.0516	0.027	✓	0.0083	✓
$(1.01, -0.02)$	0.0050	68.974	0.4259	0.2281	✓	0.1062	✓
$(1.01, -0.01)$	0.0050	68.7437	0.4513	0.3149	✓	0.1959	✓
$(1.01, 0)$	0.0050	69.9153	0.4513	0.3136	✓	0.2882	✓
$(1.01, 0.01)$	0.0050	68.2196	0.4513	0.311	✓	0.2466	✓
$(1.01, 0.02)$	0.0050	69.6443	0.4513	0.4314	✓	0.3877	✓
$(1.02, -0.02)$	0.0099	145.3234	0.2535	0.1784	✓	0.0792	✓
$(1.02, -0.01)$	0.0099	142.6586	0.4513	0.3884	✓	0.2809	✓
$(1.02, 0)$	0.0099	143.0361	0.4259	0.2128	✓	0.0702	✓
$(1.02, 0.01)$	0.0097	143.469	0.2281	0.0946	✓	0.058	
$(1.02, 0.02)$	0.0097	141.9153	0.2281	0.1722	✓	0.205	

Table 22: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 2.

- We also employ similar steps as in Scenario 1 to assess the success of disentanglement for the SAP metric in Scenario 3. To ensure a fair comparison when evaluating disentanglement with the $\tilde{\mathcal{I}}_3$ metric, we consider the SAP metric successful in achieving disentanglement under the same criterion: all three generative factors must be captured using a single latent variable. This implies that the latent variable index of the entry with the highest value in all three columns of the matrix $\mathbf{S}$ must be the same.

Examining Table 23, the metric $\tilde{\mathcal{I}}_3$ is highly effective for studying disentanglement when full correlation among generative factors exists. Specifically, in the third scenario, the successful disentanglement rate using the metric $\tilde{\mathcal{I}}_3$ is 100%, which is much higher than 32% in the first scenario when no correlation exists and 68% in the second scenario when partial correlation exists. Unlike the first two scenarios, in the third scenario, the SAP metric is less effective in studying disentanglement compared to metric $\tilde{\mathcal{I}}_3$, although the success rate is still high at 88%.

By observing Figure 12, it is noticeable that $\tilde{\mathcal{I}}_3$ and SAP scores across multiple (γ, λ) pairs once again follow a similar pattern.

11.6.2 $(s, n, m) = (3, 4, 2)$

In this section, we assess the effectiveness of the SAP metric in comparison with the metric $\tilde{\mathcal{I}}_2$ across three distinct scenarios outlined in Section 11.3. These scenarios include the independence of generative factors, partial correlation among generative factors, and complete correlation among generative factors.

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	$\tilde{\mathcal{I}}_3$ Score	$\tilde{\mathcal{I}}_3$ Success	SAP Score	SAP Success
$(0.98, -0.02)$	0.0152	0.2726	0.1485	0.1391	✓	0.3499	✓
$(0.98, -0.01)$	0.0152	0.2727	0.1486	0.0823	✓	0.0741	✓
$(0.98, 0)$	0.0151	0.2728	0.1486	0.0684	✓	0.0082	✓
$(0.98, 0.01)$	0.0151	0.2729	0.1487	0.137	✓	0.3435	✓
$(0.98, 0.02)$	0.0150	0.273	0.1487	0.1369	✓	0.3423	✓
$(0.99, -0.02)$	0.0088	0.2895	0.1589	0.1108	✓	0.1936	✓
$(0.99, -0.01)$	0.0087	0.2896	0.159	0.1007	✓	0.1961	✓
$(0.99, 0)$	0.0087	0.2898	0.159	0.0818	✓	0.1045	✓
$(0.99, 0.01)$	0.0087	0.2899	0.1591	0.0748	✓	0.0153	✓
$(0.99, 0.02)$	0.0086	0.29	0.1592	0.0946	✓	0.1754	✓
$(1, -0.02)$	0.0001	0.3209	0.1777	0.1278	✓	0.2405	✓
$(1, -0.01)$	5.912e-05	0.3204	0.1766	0.1244	✓	0.2285	✓
$(1, 0)$	1e-06	0.3235	0.1781	0.0792	✓	0.026	✓
$(1, 0.01)$	6.349e-05	0.3228	0.1783	0.1717	✓	0.4239	✓
$(1, 0.02)$	0.0001	0.3216	0.1777	0.0859	✓	0.1081	✓
$(1.01, -0.02)$	0.0048	69.8466	0.4892	0.226	✓	0.2267	✓
$(1.01, -0.01)$	0.0050	69.6363	0.6396	0.2887	✓	0.3364	✓
$(1.01, 0)$	0.0049	68.5649	0.6396	0.4225	✓	0.7438	✓
$(1.01, 0.01)$	0.0050	69.799	0.6563	0.4014	✓	0.5976	
$(1.01, 0.02)$	0.0049	70.4844	0.6396	0.32	✓	0.446	
$(1.02, -0.02)$	0.0095	143.8766	0.4892	0.3825	✓	0.6029	✓
$(1.02, -0.01)$	0.0098	143.5167	0.6396	0.5095	✓	0.8608	✓
$(1.02, 0)$	0.0095	143.4126	0.4892	0.1871	✓	0.0318	✓
$(1.02, 0.01)$	0.0098	143.3949	0.6396	0.2691	✓	0.4328	
$(1.02, 0.02)$	0.0095	143.9086	0.4892	0.4335	✓	0.7498	✓

Table 23: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 3.

The formulation of the metric $\tilde{\mathcal{I}}_3$ is guided by the intuition of partitioning the given generative factors into three distinct groups. While some of these groups may be empty, the union of all groups must be equal to the total number of given generative factors. Conversely, the metric $\tilde{\mathcal{I}}_2$ is designed to address scenarios where the dimension of the latent variable $\boldsymbol{X}$ is 2. This implies the need to partition all generative factors into two distinct groups, allowing for the possibility that one of these groups may be empty. Consequently, two distinct cases arise:

- **Case 1: None of the groups are empty**

  In this case, there are a total of 6 ways to partition the set of 3 generative factors, v_1, v_2, v_3, into 2 distinct groups, v_{s_1} and v_{s_2}. Each group of generative factors is associated with a fixed latent variable, denoted as x_1 for v_{s_1} and x_2 for v_{s_2}, respectively. The associations are determined as follows.

  1. $v_{s_1} = \{v_1\}$ and $v_{s_2} = \{v_2, v_3\}$
  2. $v_{s_1} = \{v_2\}$ and $v_{s_2} = \{v_1, v_3\}$
  3. $v_{s_1} = \{v_3\}$ and $v_{s_2} = \{v_1, v_2\}$
  4. $v_{s_1} = \{v_1, v_2\}$ and $v_{s_2} = \{v_3\}$
  5. $v_{s_1} = \{v_1, v_3\}$ and $v_{s_2} = \{v_2\}$
  6. $v_{s_1} = \{v_2, v_3\}$ and $v_{s_2} = \{v_1\}$

- **Case 2: One group is empty**

  In this case, all three given generative factors must be included in one group, either v_{s_1} or v_{s_2}.

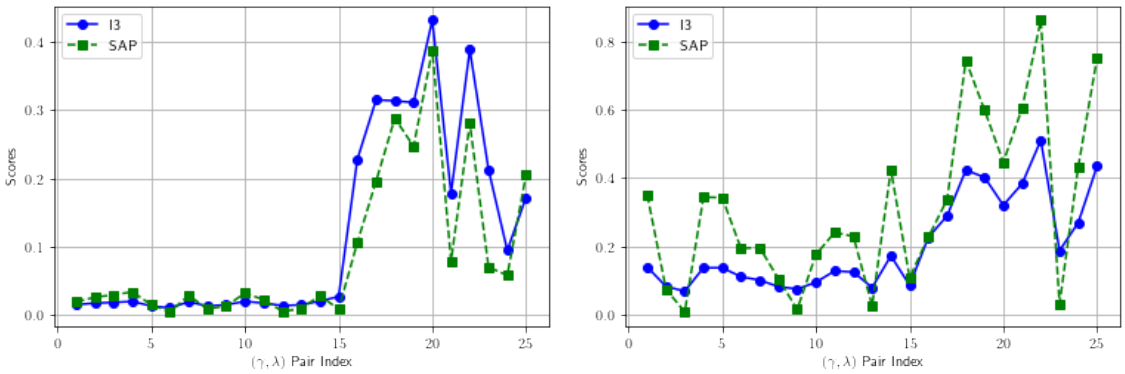

Figure 12: $\mathcal{I}_3$ vs. SAP scores for each (γ, λ) pair in Scenario 2 (left) and Scenario 3 (right).

1. $v_{s_1} = \{v_1, v_2, v_3\}$ and $v_{s_2} = \varnothing$
2. $v_{s_1} = \varnothing$ and $v_{s_2} = \{v_1, v_2, v_3\}$

From these two cases, a total of 8 possible ways exist to partition the set of 3 generative factors into 2 distinct groups, considering the case of one group may be empty.

To evaluate disentanglement in this context, we employ the $\tilde{\mathcal{I}}_2$ score, which can be computed as follows.

$$\tilde{\mathcal{I}}_2 \text{ score} = \begin{cases} I(x_1; v_{s_1}) + I(x_2; v_{s_2}) - I(v_{s_1}; v_{s_2}) & \text{if } v_{s_1}, v_{s_2} \neq \varnothing \\ I(x_i; v_{s_i}) & \text{if } v_{s_i} \neq \varnothing \text{ and } v_{s_j} = \varnothing \text{ for } i \neq j. \end{cases} \tag{34}$$

We conduct numerical simulations under the given assumption $(s, n, m) = (3, 4, 2)$. The parameter values for all three scenarios are consistent and described in Table 10. The covariance matrices $\mathbf{\Sigma_V}$ and $\mathbf{\Sigma_Y}$ used for these scenarios are detailed in Table 11. Our investigation involves the analysis of 25 pairs of parameters (γ, λ), selected from two arrays: $\gamma = \{0.98, 0.99, 1, 1.01, 1.02\}$ and $\lambda = \{-0.02, -0.01, 0, 0.01, 0.02\}$. Across the three distinct scenarios, we assess the effectiveness of metrics SAP and $\tilde{\mathcal{I}}_2$ in capturing disentanglement. The evaluation criteria for each metric are defined by the following rules:

- **Scenario 1: Independence of Generative Factors**
  - **Metric $\tilde{\mathcal{I}}_2$**: For each pair of (γ, λ), we calculate the $\tilde{\mathcal{I}}_2$ score using the system (34) for every possible partition. The partition with the maximum $\tilde{\mathcal{I}}_2$ score is then identified. If this partition accurately characterizes the relationship among the given generative factors, we deem metric $\tilde{\mathcal{I}}_2$ successful in studying disentanglement for that specific (γ, λ) pair. In Scenario 1, where the three generative factors are independent, each factor should ideally correspond to a distinct latent variable. However, with only two latent variables available, it is impossible to form three distinct groups, each containing exactly one generative factor. Consequently, all listed partitions are considered equivalent and are deemed correct ways to partition the given generative factors.
  - **Metric SAP**: Evaluating the efficacy of disentanglement for the SAP metric involves a series of steps. Initially, we construct a correlation score matrix $\mathbf{S}$ with dimensions 2×3, where each entry value is computed using equation (32). Each of the three columns, from left to right, in this matrix corresponds to a generative factor v_1, v_2, v_3, respectively, and each of the two rows, from top to bottom, corresponds to a latent variable x_1, x_2, respectively.

    For each column j, we identify the row index i of the entry with the highest value. This implies that the latent variable x_i should be employed to capture generative factor v_j due to its strongest correlation with v_j among all available latent variables. Subsequently, we use a ✓ to indicate the entry with the highest value for each column. For instance, as shown in Table 24, this signifies that latent variable x_1 should be utilized to capture two generative factors, v_1 and v_3, while the remaining latent variable x_2 should be employed to capture generative factor v_2.

S	v_1	v_2	v_3
x_1	✓		✓
x_2		✓	

Table 24: Example score matrix **S** for Scenario 1.

In Scenario 1, we affirm that the SAP metric successfully studies disentanglement if each generative factor is captured by a distinct latent variable. This assertion implies that the latent variable index associated with the entry having the highest value in each column must be unique. However, given the limitation of having only 2 latent variables, it is necessary to employ at least one latent variable to study 2 or more generative factors. Therefore, in this scenario, we can conclude that the SAP metric successfully studies disentanglement if either one or both latent variables are used to capture information from generative factors. Consequently, the SAP metric can always be used to study disentanglement in this scenario.

- **Scenario 2: Linear Dependence of v_1 and v_2, with Independence of v_3**
  - **Metric $\tilde{\mathcal{I}}_2$**: In this scenario, metric $\tilde{\mathcal{I}}_2$ is considered successful in studying disentanglement for each pair of (γ, λ) if the maximum $\tilde{\mathcal{I}}_2$ score occurs at a partition where group v_{s_i} contains 2 correlated generative factors v_1 and v_2, while the group v_{s_j}, for $i \neq j$, contains the independent generative factor v_3.
  - **Metric SAP**: In this scenario, metric SAP is considered successful in studying disentanglement if both generative factors v_1 and v_2 exhibit the strongest correlation with the same latent variable x_i, while v_3 has the strongest correlation with x_j, for $i \neq j$. This implies that the row index of the entry with the highest value in the first two columns of matrix **S** must be the same, while the row index of the entry with the highest value in the third column must be different, as exemplified in Table 25.

S	v_1	v_2	v_3
x_1			✓
x_2	✓	✓	

Table 25: Example score matrix **S** for Scenario 2.

- **Scenario 3: Linear Dependence of v_2 and v_3 on v_1**
  - **Metric $\tilde{\mathcal{I}}_2$**: In this scenario, metric $\tilde{\mathcal{I}}_2$ is considered successful in studying disentanglement for each pair of (γ, λ) if the maximum $\tilde{\mathcal{I}}_2$ score occurs at a partition where group v_{s_i} contains all 3 given generative factors due to complete correlation, while the remaining group is empty.
  - **Metric SAP**: In this scenario, metric SAP is considered successful in studying disentanglement if all generative factors exhibit the strongest correlation with the same latent variable. This implies that the row index of the entry with the highest value for all 3 columns of matrix **S** must be unique, as exemplified in Table 26.

S	v_1	v_2	v_3
x_1			
x_2	✓	✓	✓

Table 26: Example score matrix **S** for Scenario 3.

We now present the numerical results for each scenario in the case where the dimension of the latent variable is less than the dimension of generative factors. We aim to investigate whether the conclusions drawn from numerical simulations remain consistent when the dimensions of these variables are not equal.

1. **Scenario 1: Independence of Generative Factors**

   We will omit the results for Scenario 1 since this scenario, characterized by the independence of generative factors, is of lesser concern. In this scenario, the specific approach to utilizing latent variables for capturing information from generative variables does not contribute meaningfully to the comparison of the effectiveness of disentanglement between the two metrics.

2. **Scenario 2: Linear Dependence of v_1 and v_2, with Independence of v_3**

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	$\tilde{\mathcal{I}}_2$ Score	$\tilde{\mathcal{I}}_2$ Success	SAP Score	SAP Success
$(0.98, -0.02)$	0.0152	0.0789	0.02	0.02		0.0342	✓
$(0.98, -0.01)$	0.0152	0.0789	0.02	0.0134		0.0119	✓
$(0.98, 0)$	0.0152	0.0789	0.02	0.0169		0.0237	✓
$(0.98, 0.01)$	0.0151	0.079	0.02	0.0103		0.0013	✓
$(0.98, 0.02)$	0.0151	0.079	0.02	0.0113		0.0047	✓
$(0.99, -0.02)$	0.0075	0.0892	0.0224	0.0223		0.0379	✓
$(0.99, -0.01)$	0.0075	0.0892	0.0224	0.0223		0.038	✓
$(0.99, 0)$	0.0074	0.0892	0.0224	0.0194		0.0282	✓
$(0.99, 0.01)$	0.0074	0.0893	0.0225	0.0221		0.037	✓
$(0.99, 0.02)$	0.0074	0.0893	0.0225	0.0221		0.0372	✓
$(1, -0.02)$	5.376e-05	0.0996	0.0238	0.0204	✓	0.0291	✓
$(1, -0.01)$	2.819e-05	1.007	0.0229	0.0229	✓	0.039	✓
$(1, 0)$	1e-06	0.0995	0.0249	0.0246	✓	0.0416	✓
$(1, 0.01)$	2.673e-05	0.0996	0.0257	0.0217	✓	0.0306	✓
$(1, 0.02)$	5.243e-05	0.1012	0.027	0.064	✓	0.0437	✓
$(1.01, -0.02)$	0.0048	46.282	0.2027	0.1941		0.2571	✓
$(1.01, -0.01)$	0.0049	45.4903	0.2281	0.2137		0.2867	✓
$(1.01, 0)$	0.0050	45.6119	0.2535	0.2268	✓	0.1307	✓
$(1.01, 0.01)$	0.0034	45.0386	0.0303	0.0276		0.0322	✓
$(1.01, 0.02)$	0.0049	46.1263	0.2281	0.1996		0.244	✓
$(1.02, -0.02)$	0.0097	95.1093	0.2281	0.1914		0.2189	✓
$(1.02, -0.01)$	0.0097	93.4183	0.2281	0.1956		0.2307	✓
$(1.02, 0)$	0.0097	95.1079	0.2281	0.1955		0.2305	✓
$(1.02, 0.01)$	0.0095	95.4747	0.2027	0.2045		0.279	✓
$(1.02, 0.02)$	0.0097	95.8823	0.2281	0.1441		0.1335	✓

Table 27: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 2.

   In contrast to the scenario where the dimension of the latent variable equals the dimension of the generative variable, as discussed in Section 11.6.1, the effectiveness of the metric $\tilde{\mathcal{I}}_2$ is significantly reduced in Scenario 2 when the dimension of the latent variable is less than the dimension of the generative variable. Specifically, the successful disentanglement rate using metric $\tilde{\mathcal{I}}_2$ in Scenario 2 is only 24%, as indicated in Table 27, which is lower than the rate achieved with metric $\tilde{\mathcal{I}}_3$ at 68%, as shown in Table 22. In contrast, the metric SAP demonstrates superior performance in disentanglement, achieving a success rate of 100%, outperforming metric $\tilde{\mathcal{I}}_2$ in the presence of partial correlation among generative factors.

3. **Scenario 3: Linear Dependence of v_2 and v_3 on v_1**

   In Scenario 3, where full correlation exists among generative factors, similar to the case where the dimension of the latent variable is equal to the dimension of the generative variable, the metric $\tilde{\mathcal{I}}_2$ shows highly effective for studying disentanglement even when the dimension of the latent variable is less than the dimension of the generative variable, achieving a success rate of 100%, as depicted

(γ, λ)	$\mathcal{L}_{\text{rec}}$	$I_{\boldsymbol{\phi}}(\boldsymbol{Y}; \boldsymbol{X})$	$I_{\boldsymbol{\phi}}(\boldsymbol{V}; \boldsymbol{X})$	$\tilde{\mathcal{I}}_2$ Score	$\tilde{\mathcal{I}}_2$ Success	SAP Score	SAP Success
$(0.98, -0.02)$	0.0152	0.2726	0.1485	0.1383	✓	0.3417	✓
$(0.98, -0.01)$	0.0151	0.2727	0.1486	0.1252	✓	0.2809	✓
$(0.98, 0)$	0.0151	0.2728	0.1486	0.1098	✓	0.2069	✓
$(0.98, 0.01)$	0.0151	0.2729	0.1487	0.1417	✓	0.3567	✓
$(0.98, 0.02)$	0.0150	0.273	0.1487	0.1442	✓	0.3681	✓
$(0.99, -0.02)$	0.0088	0.2895	0.1589	0.1538	✓	0.3875	✓
$(0.99, -0.01)$	0.0087	0.2896	0.159	0.092	✓	0.0962	✓
$(0.99, 0)$	0.0087	0.2898	0.159	0.1557	✓	0.3955	✓
$(0.99, 0.01)$	0.0087	0.2899	0.1591	0.101	✓	0.1402	✓
$(0.99, 0.02)$	0.0086	0.29	0.1592	0.0877	✓	0.0737	✓
$(1, -0.02)$	0.0001	0.3209	0.1777	0.1571	✓	0.3593	✓
$(1, -0.01)$	6.337e-05	0.3213	0.1779	0.1576	✓	0.361	✓
$(1, 0)$	1e-06	0.3214	0.1778	0.1593	✓	0.3678	✓
$(1, 0.01)$	6.356e-05	0.3219	0.1783	0.1515	✓	0.3329	✓
$(1, 0.02)$	0.0001	0.3222	0.1785	0.1402	✓	0.2821	✓
$(1.01, -0.02)$	0.0049	46.8817	0.6396	0.3097	✓	0.4431	
$(1.01, -0.01)$	0.0049	46.8959	0.6396	0.489	✓	0.7012	✓
$(1.01, 0)$	0.0049	46.5917	0.6396	0.5122	✓	0.8421	✓
$(1.01, 0.01)$	0.0049	46.6654	0.6396	0.4399	✓	0.6631	
$(1.01, 0.02)$	0.0049	46.4838	0.6396	0.3231	✓	0.4596	
$(1.02, -0.02)$	0.0098	95.2912	0.6396	0.5152	✓	0.8261	✓
$(1.02, -0.01)$	0.0095	95.7142	0.4892	0.4549	✓	0.8021	✓
$(1.02, 0)$	0.0098	95.187	0.6396	0.5065	✓	0.866	✓
$(1.02, 0.01)$	0.0095	95.6982	0.4892	0.4858	✓	0.8462	✓
$(1.02, 0.02)$	0.0098	94.8639	0.6396	0.4178	✓	0.6355	

Table 28: Reconstruction error, mutual information, and disentanglement metric value for each (γ, λ) pair in Scenario 3.

in Table 28. While the SAP metric exhibits slightly lower effectiveness in studying disentanglement compared to metric $\tilde{\mathcal{I}}_2$, the success rate remains very high at 88%.

By examining Figure 13, which illustrates the variations of $\tilde{\mathcal{I}}_2$ and SAP scores across 25 pairs of (γ, λ), a notable distinction emerges. In contrast to the scenario where the dimension of the latent variable is equal to the dimension of the generative variable, it is evident that, when the dimension of the latent variable is less than the dimension of the generative variable, the $\tilde{\mathcal{I}}_2$ score no longer adheres to the pattern exhibited by the SAP score for all considered pairs of (γ, λ). For instance, in Scenario 2 at pair index 18, where $(\gamma, \lambda) = (1.01, 0)$, the SAP score decreases while the $\tilde{\mathcal{I}}_2$ score increases.

11.6.3 Summary of Numerical Results

Observing Figure 14, which depicts the distribution of $\mathcal{I}_3$ and SAP scores across three scenarios representing no correlation, partial correlation, and full correlation among generative factors, it is evident that both the mean and median of both metric scores increase. Specifically, the mean and median of both metric scores are highest in Scenario 3 and lowest in Scenario 1. The same results hold when we replace $\mathcal{I}_3$ with $\mathcal{I}_2$, as shown in Figure 15. Since a higher SAP score indicates better disentanglement, this trend suggests that both metrics, $\mathcal{I}_m$ and SAP, are most effective when there is full correlation among generative factors, less efficient when partial correlation exists, and inefficient when no correlation exists.

Metric SAP exhibits advantages over metric $\mathcal{I}_m$ in scenarios with non-full correlation among generative factors, demonstrating a higher disentanglement success rate. This superiority is particularly evident when the dimension of the latent variable is less than the dimension of the generative variable. Additionally, metric SAP computes the score faster, as it does not need to check through multiple partitions like metric $\mathcal{I}_m$ does.

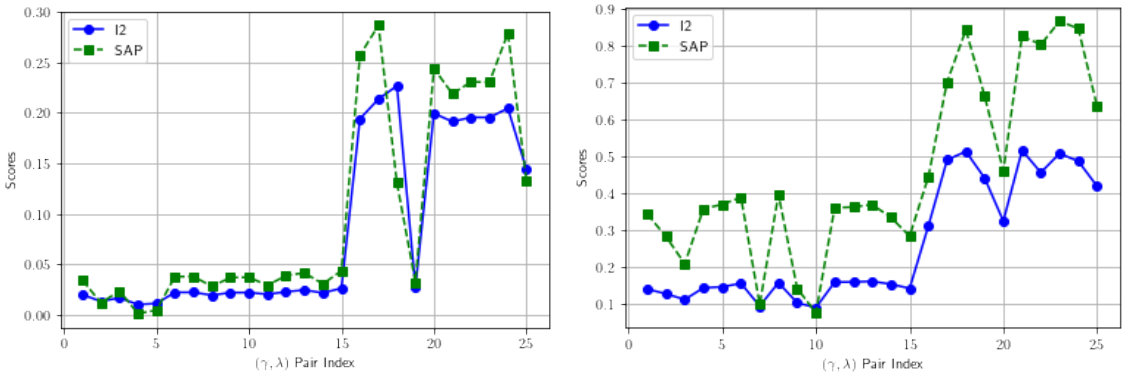

Figure 13: $\mathcal{I}_2$ vs. SAP scores for each (γ, λ) pair in Scenario 2 (left) and Scenario 3 (right).

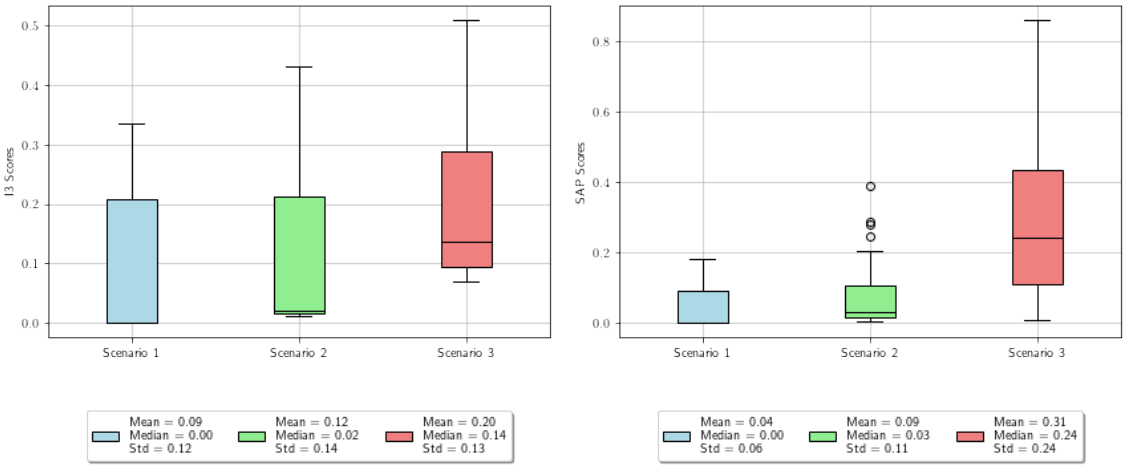

Figure 14: Distribution of $\mathcal{I}_3$ (left) and SAP (right) scores across scenarios given $(s, n, m) = (3, 4, 3)$.

Metric SAP is also highly efficient in cases where full correlation among generative factors exists. It achieves a high disentanglement success rate, and its fast computation is a significant advantage compared to the $\mathcal{I}_m$ metric. However, if accuracy is prioritized, using metric $\tilde{\mathcal{I}}_m$ will yield a 100% disentanglement success rate. Therefore, there is a trade-off between computational efficiency and accuracy between these two metrics in scenarios with full correlation among generative factors. For other cases, the SAP metric clearly excels in both computational efficiency and accuracy.

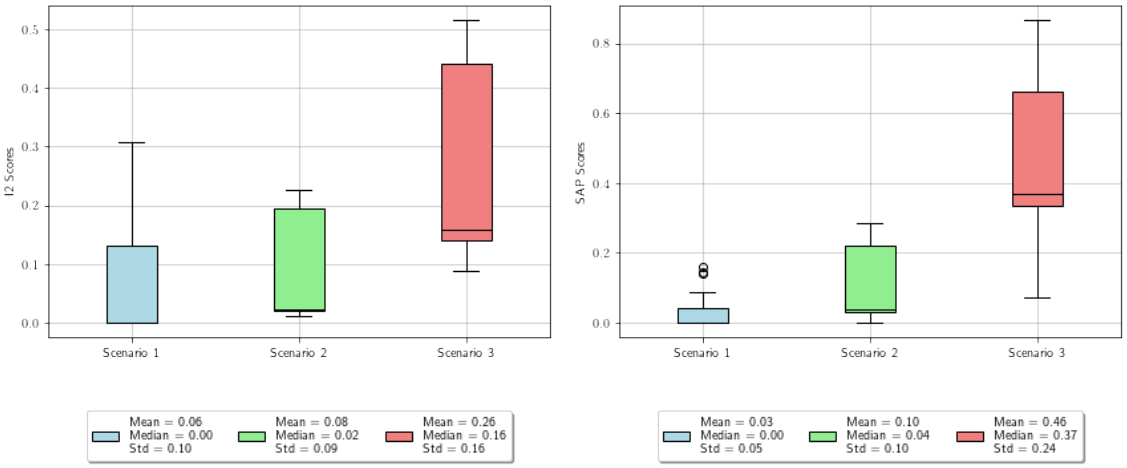

Figure 15: Distribution of $\mathcal{I}_2$ (left) and SAP (right) scores across scenarios given $(s, n, m) = (3, 4, 2)$.

References

[1] S. Arimoto. "An algorithm for computing the capacity of arbitrary discrete memoryless channels". In: *IEEE Transactions on Information Theory* 18.1 (1972), pp. 14–20. DOI: 10.1109/TIT.1972.1054753.

[2] R. Blahut. "Computation of channel capacity and rate-distortion functions". In: *IEEE Transactions on Information Theory* 18.4 (1972), pp. 460–473. DOI: 10.1109/TIT.1972.1054855.

[3] H. V. Henderson and S. R. Searle. "On Deriving the Inverse of a Sum of Matrices". In: *SIAM Review* 23.1 (1981), pp. 53–60. DOI: 10.1137/1023004. eprint: https://doi.org/10.1137/1023004. URL: https://doi.org/10.1137/1023004.

[4] Tzon-Tzer Lu and Sheng-Hua Shiou. "Inverses of 2 × 2 block matrices". In: *Computers Mathematics with Applications* 43.1 (2002), pp. 119–129. ISSN: 0898-1221. DOI: https://doi.org/10.1016/S0898-1221(01)00278-4. URL: https://www.sciencedirect.com/science/article/pii/S0898122101002784.

[5] "Rate Distortion Theory". In: *Elements of Information Theory*. John Wiley Sons, Ltd, 2005. Chap. 10, pp. 301–346. ISBN: 9780471748823. DOI: https://doi.org/10.1002/047174882X.ch10. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/047174882X.ch10. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/047174882X.ch10.

[6] Kaare Brandt Petersen and Michael Syskind Pedersen. "The Matrix Cookbook". In: 2006. URL: https://api.semanticscholar.org/CorpusID:1221763.

[7] Philip D. Powell. *Calculating Determinants of Block Matrices*. 2011. arXiv: 1112.4379 [math.RA].

[8] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy2fzU9gl.

[9] Alexander A. Alemi et al. *Fixing a Broken ELBO*. 2018. arXiv: 1711.00464 [cs.LG].

[10] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. *Variational Inference of Disentangled Latent Concepts from Unlabeled Observations*. 2018. arXiv: 1711.00848 [cs.LG].

[11] Ricky T. Q. Chen et al. *Isolating Sources of Disentanglement in Variational Autoencoders*. 2019. arXiv: 1802.04942 [cs.LG].

[12] Christian Jacobsen and Karthik Duraisamy. *Disentangling Generative Factors of Physical Fields Using Variational Autoencoders*. 2021. arXiv: 2109.07399 [physics.comp-ph].

[13] Marc-André Carbonneau et al. *Measuring Disentanglement: A Review of Metrics*. 2022. arXiv: 2012.09276 [cs.LG].

[14] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

Appendix

A Proofs of Supporting Lemmas

Lemma 7. *The mean* $\boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y}}$ *and covariance* $\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}$ *of the joint distribution* $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ *are*

$$\boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y}} = \mathbf{0}_{(m+n)\times 1} \quad \textit{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix}.$$

Proof. As $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, the mean and $(m+n)\times(m+n)$-covariance matrix of the joint distribution $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ are of the form

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y}} &= \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{bmatrix} = \mathbf{0}_{(m+n)\times 1} \\ \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{X}} & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix}. \end{aligned}$$

□

Lemma 8. *The mean* $\boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}$ *and covariance* $\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}$ *of the joint distribution* $q_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{y})$ *are*

$$\boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}} = \mathbf{0}_{(m+n)\times 1} \quad \textit{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}} = \begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix}.$$

Proof. Using the encoder model (2) and system (4), the mean and $(m+n)\times(m+n)$-covariance of the joint distribution $q_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{y})$ are of the form

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}} &= \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{bmatrix} = \mathbf{0}_{(m+n)\times 1} \\ \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}} &= \mathbb{E}\left[\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} [\boldsymbol{X}\boldsymbol{Y}^\mathsf{T}] \right] \\ &= \begin{bmatrix} \mathbb{E}[\boldsymbol{X}^2] & \mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^\mathsf{T}] \\ \mathbb{E}[\boldsymbol{Y}\boldsymbol{X}] & \mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\mathsf{T}] \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{X}} & \mathbb{E}[(\mathbf{B}\boldsymbol{Y}+\boldsymbol{W})\boldsymbol{Y}^\mathsf{T}] \\ [\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^\mathsf{T}]]^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\ [\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^\mathsf{T}]]^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix}. \end{aligned}$$

□

Lemma 9. *The mean* $\boldsymbol{\mu}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}$ *and covariance* $\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}$ *of the joint distribution* $p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ *are*

$$\boldsymbol{\mu}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}} = \mathbf{0}_{(m+n)\times 1} \quad \textit{and} \quad \boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{I}_m & \mathbf{A}^\mathsf{T} \\ \mathbf{A} & \mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}} \end{bmatrix}.$$

Proof. Using the decoder model (3) and system (4), the mean and $(m+n)\times(m+n)$-covariance of the joint

distribution $p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ are respectively described by

$$\begin{aligned}
\boldsymbol{\mu}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}} &= \begin{bmatrix}\boldsymbol{\mu}_{\hat{\boldsymbol{X}}}\\ \boldsymbol{\mu}_{\hat{\boldsymbol{Y}}}\end{bmatrix} = \mathbf{0}_{(m+n)\times 1}\\
\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}} &= \mathbb{E}\left[\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}[\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}^{\mathsf{T}}]\right]\\
&= \begin{bmatrix}\mathbb{E}[\hat{\boldsymbol{X}}^2] & \mathbb{E}[\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}^{\mathsf{T}}]\\ \mathbb{E}[\hat{\boldsymbol{Y}}\hat{\boldsymbol{X}}] & \mathbb{E}[\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}^{\mathsf{T}}]\end{bmatrix}\\
&= \begin{bmatrix}\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}}} & [\mathbb{E}[\hat{\boldsymbol{Y}}\hat{\boldsymbol{X}}]]^{\mathsf{T}}\\ \mathbb{E}[(\mathbf{A}\hat{\boldsymbol{X}}+\boldsymbol{Z})\hat{\boldsymbol{X}}] & \boldsymbol{\Sigma}_{\hat{\boldsymbol{Y}}}\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{I}_m & [\mathbb{E}[\hat{\boldsymbol{Y}}\hat{\boldsymbol{X}}]]^{\mathsf{T}}\\ \mathbf{A}\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}}} & \mathbf{A}\mathbf{A}^{\mathsf{T}}+\boldsymbol{\Sigma}_{\boldsymbol{Z}}\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{I}_m & \mathbf{A}^{\mathsf{T}}\\ \mathbf{A} & \mathbf{A}\mathbf{A}^{\mathsf{T}}+\boldsymbol{\Sigma}_{\boldsymbol{Z}}\end{bmatrix}.
\end{aligned}$$

□

Lemma 10. *The determinants of the covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}$, $\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}$, and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}$ are*

$$|\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}| = |\boldsymbol{\Sigma}_{\boldsymbol{Y}}|, \quad |\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}| = |\boldsymbol{\Sigma}_{\boldsymbol{W}}||\boldsymbol{\Sigma}_{\boldsymbol{Y}}|, \quad \textit{and} \quad |\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}| = |\boldsymbol{\Sigma}_{\boldsymbol{Z}}|.$$

Proof. Consider an arbitrary block matrix $\mathbf{M}$ that is composed of four submatrices $\mathbf{A},\mathbf{B},\mathbf{C}$, and $\mathbf{D}$ of dimensions $n\times n, n\times m, m\times n$, and $m\times m$, respectively. By [7], if $\mathbf{A}$ and $\mathbf{D}$ are invertible, then the determinant of $\mathbf{M}$ is given by

$$|\mathbf{M}| = \left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\right| = |\mathbf{A}||\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{D}||\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}|.$$

Using the determinant formula for block matrices, we get

$$\begin{aligned}
|\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}| &= \left|\begin{bmatrix}\mathbf{I}_m & \mathbf{0}_{m\times n}\\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}}\end{bmatrix}\right| = |\boldsymbol{\Sigma}_{\boldsymbol{Y}}||\mathbf{I}_m| = |\boldsymbol{\Sigma}_{\boldsymbol{Y}}|\\
|\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}| &= \left|\begin{bmatrix}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}}+\boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} & \boldsymbol{\Sigma}_{\boldsymbol{Y}}\end{bmatrix}\right|\\
&= |\boldsymbol{\Sigma}_{\boldsymbol{Y}}||(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}}+\boldsymbol{\Sigma}_{\boldsymbol{W}})-(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}})\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}})|\\
&= |\boldsymbol{\Sigma}_{\boldsymbol{W}}||\boldsymbol{\Sigma}_{\boldsymbol{Y}}|\\
|\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}| &= \left|\begin{bmatrix}\mathbf{I}_m & \mathbf{A}^{\mathsf{T}}\\ \mathbf{A} & \mathbf{A}\mathbf{A}^{\mathsf{T}}+\boldsymbol{\Sigma}_{\boldsymbol{Z}}\end{bmatrix}\right| = |\mathbf{I}_m||\mathbf{A}\mathbf{A}^{\mathsf{T}}+\boldsymbol{\Sigma}_{\boldsymbol{Z}}-\mathbf{A}\mathbf{A}^{\mathsf{T}}| = |\boldsymbol{\Sigma}_{\boldsymbol{Z}}|.
\end{aligned}$$

□

Lemma 11. *The inverse matrices of the covariance $\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}$ are*

$$\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}^{-1} = \begin{bmatrix}\mathbf{I}_m & \mathbf{0}_{m\times n}\\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\end{bmatrix} \quad \textit{and} \quad \boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}^{-1} = \begin{bmatrix}\mathbf{I}_m+\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}.$$

Proof. Using the same assumption for the block matrix $\mathbf{M}$ as in the proof of **Lemma 10**, by [4], the inverse of $\mathbf{M}$ is given by

$$\mathbf{M}^{-1} = \begin{bmatrix}\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\\ -(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\end{bmatrix}.$$

Using the inverse formula for block matrix gives

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}^{-1} &= \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} \end{bmatrix} \\
\boldsymbol{\Sigma}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}^{-1} &= \begin{bmatrix} \mathbf{I}_m & \mathbf{A}^\mathsf{T} \\ \mathbf{A} & \mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \mathbf{I}_m + \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1} \\ -(\mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1}\mathbf{A} & (\mathbf{A}\mathbf{A}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}} - \mathbf{A}\mathbf{A}^\mathsf{T})^{-1} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_m + \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} \\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} \end{bmatrix}.
\end{aligned}$$

□

B Proofs of Main Lemmas and Propositions

B.1 Proofs in Section 3

B.1.1 Proposition 1

Proof. Notice that $p_{\boldsymbol{Y}}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})$ and $p_{\boldsymbol{X}}(\boldsymbol{x})p_{\boldsymbol{Y}}(\boldsymbol{y})$ respectively define joint Gaussian distributions $q_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{y})$ and $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$. Since the KL divergence formula between two multivariate Gaussian distributions $p_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $p_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is given by

$$D_{KL}[p_1\|p_2] = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} - n + \mathrm{Tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right], \tag{35}$$

by **Lemmas 7**, **8**, **10**, and **11**, the *regularization term* in the γ-VAE loss function (7) can be computed as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}[D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]] \\
&= \int p_{\boldsymbol{Y}}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\log\frac{q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x})}dxdy \\
&= \int p_{\boldsymbol{Y}}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\log\frac{q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})p_{\boldsymbol{Y}}(\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x})p_{\boldsymbol{Y}}(\boldsymbol{y})}dxdy \\
&= D_{KL}[p_{\boldsymbol{Y}}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})p_{\boldsymbol{Y}}(\boldsymbol{y})] \\
&= D_{KL}[q_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{y})\|p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})] \\
&= D_{KL}[\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}, \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}})\|\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X},\boldsymbol{Y}}, \boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}})] \\
&= \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_{\boldsymbol{Y}}|}{|\boldsymbol{\Sigma}_{\boldsymbol{W}}||\boldsymbol{\Sigma}_{\boldsymbol{Y}}|} - (m+n) + \mathrm{Tr}\left(\begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix}\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} \end{bmatrix}\right)\right] \\
&= \frac{1}{2}\left[\log\frac{1}{|\boldsymbol{\Sigma}_{\boldsymbol{W}}|} - m - n + \mathrm{Tr}\begin{bmatrix} \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} & \mathbf{I}_n \end{bmatrix}\right] \\
&= \frac{1}{2}\left[-\log|\boldsymbol{\Sigma}_{\boldsymbol{W}}| - m - n + n + \mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}})\right] \\
&= \frac{1}{2}\left[\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}}) - \log|\boldsymbol{\Sigma}_{\boldsymbol{W}}| - m\right].
\end{aligned}$$

□

B.1.2 Proposition 2

Before proceeding to establish **Proposition 2**, it is necessary to prove the two following **Lemmas 12** and **13**.

Lemma 12. *The expected log-likelihood of the joint distribution $p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ is given by*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})]\\ &= -\frac{1}{2}\Bigg[(m+n)\log(2\pi)+\log|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|+\mathrm{Tr}\Bigg[(\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A})(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}+\boldsymbol{\Sigma}_{\boldsymbol{W}})\\ &\quad+\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}-\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}-\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\Bigg]\Bigg].\end{aligned}$$

Proof. Since the general multivariate Gaussian distribution of the probability density function for a random variable $\boldsymbol{X}\in\mathbb{R}^n$ is expressed as

$$p_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right], \tag{36}$$

by **Lemmas 9**, **10**, and **11**, the log-likelihood

$$\begin{aligned}&\log p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})\\ &=\log\left[\mathcal{N}\left(\mathbf{0}_{(m+n)\times 1},\begin{bmatrix}\mathbf{I}_m & \mathbf{A}^\mathsf{T}\\ \mathbf{A} & \mathbf{A}\mathbf{A}^\mathsf{T}+\boldsymbol{\Sigma}_{\boldsymbol{Z}}\end{bmatrix}\right)\right]\\ &=\log\left[\frac{1}{(2\pi)^{(m+n)/2}|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|^{1/2}}\exp\left(-\frac{1}{2}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\mathbf{I}_m & \mathbf{A}^\mathsf{T}\\ \mathbf{A} & \mathbf{A}\mathbf{A}^\mathsf{T}+\boldsymbol{\Sigma}_{\boldsymbol{Z}}\end{bmatrix}^{-1}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}\right)\right]\\ &=-\frac{m+n}{2}\log(2\pi)-\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|-\frac{1}{2}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}.\end{aligned} \tag{37}$$

Using **Lemma 8** gives

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}\left[\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}\right]\\ &=\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}\left[\mathrm{Tr}\left(\begin{bmatrix}\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}^\mathsf{T}\right)\right]\\ &=\mathrm{Tr}\left(\begin{bmatrix}\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}\left[\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}^\mathsf{T}\right]\right)\\ &=\mathrm{Tr}\left(\begin{bmatrix}\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}\right)\\ &=\mathrm{Tr}\left(\begin{bmatrix}\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}+\boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} & \boldsymbol{\Sigma}_{\boldsymbol{Y}}\end{bmatrix}\right)\\ &=\mathrm{Tr}[(\mathbf{I}_m+\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A})(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}+\boldsymbol{\Sigma}_{\boldsymbol{W}})+\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}-\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}-\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}].\end{aligned} \tag{38}$$

From equations (37) and (38), we can derive the expected log-likelihood $\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log[p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})]]$. □

Lemma 13. *The expected log-likelihood of the prior distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$ is given by*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\boldsymbol{X}}(\boldsymbol{x})]=-\frac{1}{2}\Big[m\log(2\pi)+\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}+\boldsymbol{\Sigma}_{\boldsymbol{W}})\Big].$$

Proof. By equations (4) and (36), the expected log-likelihood of $p_{\boldsymbol{X}}(\boldsymbol{x})$ can be computed as

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\boldsymbol{X}}(\boldsymbol{x})]$$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log \mathcal{N}(\mathbf{0}, \mathbf{I}_m)] \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}\left[\log\left[\frac{1}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)}\right)\right]\right] \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}\left[-\frac{m}{2}\log(2\pi) - \frac{1}{2}\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)}\right] \\
&= -\frac{m}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)}] \\
&= -\frac{m}{2}\log(2\pi) - \frac{1}{2}\operatorname{Tr}(\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}]) \\
&= -\frac{m}{2}\log(2\pi) - \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}) \\
&= -\frac{1}{2}\Big[m\log(2\pi) + \operatorname{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{W}})\Big].
\end{aligned}
$$

□

As $p_{\boldsymbol{X}}(\boldsymbol{x})p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})$ defines a joint distribution of $p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$, the *reconstruction term* in the γ-VAE loss function (7) can be written as

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})] \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}\left[\log\frac{p_{\boldsymbol{X}}(\boldsymbol{x})p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})}{p_{\boldsymbol{X}}(\boldsymbol{x})}\right] \\
&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})] - \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi}}[\log p_{\boldsymbol{X}}(\boldsymbol{x})]
\end{aligned}
$$

Using **Lemmas 12** and **13** give **Proposition 2**.

B.1.3 Lemma 1

Before deriving the optimal solution set $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$ for the γ-VAE loss function (7), we must initially determine its gradient, as outlined in **Lemma 14**.

Lemma 14. *The gradient of the objective function $\boldsymbol{\Gamma}_1$ is defined by*

$$
\begin{aligned}
\nabla\boldsymbol{\Gamma}_1(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) = \Big\langle &\frac{\partial\boldsymbol{\Gamma}_1}{\partial\mathbf{A}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}), \frac{\partial\boldsymbol{\Gamma}_1}{\partial\mathbf{B}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}), \\
&\frac{\partial\boldsymbol{\Gamma}_1}{\partial\boldsymbol{\Sigma}_{\boldsymbol{Z}}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}), \frac{\partial\boldsymbol{\Gamma}_1}{\partial\boldsymbol{\Sigma}_{\boldsymbol{W}}}(\mathbf{A}, \mathbf{W}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})\Big\rangle,
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial\boldsymbol{\Gamma}_1}{\partial\mathbf{A}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) &= \gamma[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{W}}) - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}}] \\
\frac{\partial\boldsymbol{\Gamma}_1}{\partial\mathbf{B}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) &= \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \gamma[\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}] \\
\frac{\partial\boldsymbol{\Gamma}_1}{\partial\boldsymbol{\Sigma}_{\boldsymbol{W}}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{W}}) &= \frac{1}{2}\Big[\mathbf{I}_m - \boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1} + \gamma\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\Big] \\
&\frac{\partial\boldsymbol{\Gamma}_1}{\partial\boldsymbol{\Sigma}_{\boldsymbol{Z}}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) \\
= \frac{\gamma}{2}\Big[&\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{W}})\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\Big].
\end{aligned}
$$

Proof. By [6], the partial derivatives of $\boldsymbol{\Gamma}_1$ with respect to $\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ can be calculated as follows:

$$
\frac{\partial\boldsymbol{\Gamma}_1}{\partial\mathbf{A}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})
$$

$$= -\frac{\gamma}{2}\left[\frac{\partial}{\partial \mathbf{A}}\left(\mathrm{Tr}[\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}})]\right)\right]$$

$$= -\frac{\gamma}{2}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} - 2\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}})\right]$$

$$= \gamma[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}}) - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}]$$

$$\frac{\partial \boldsymbol{\Gamma}_1}{\partial \mathbf{B}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$$

$$= \frac{\partial}{\partial \mathbf{B}}\left[\frac{1}{2}\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}) - \frac{\gamma}{2}\left(\mathrm{Tr}[\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}]\right)\right]$$

$$= \frac{1}{2}(2\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}) - \frac{\gamma}{2}\left[\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} + \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - 2\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right]$$

$$= \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \gamma[\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}]$$

$$\frac{\partial \boldsymbol{\Gamma}_1}{\partial \boldsymbol{\Sigma}_{\boldsymbol{Z}}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$$

$$= -\frac{\gamma}{2}\left[\frac{\partial}{\partial \boldsymbol{\Sigma}_{\boldsymbol{Z}}}\left(\mathrm{Tr}[\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}})] - \log|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|\right)\right]$$

$$= \frac{\gamma}{2}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \boldsymbol{\Sigma}_{\boldsymbol{W}})\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\right]$$

$$\frac{\partial \boldsymbol{\Gamma}_1}{\partial \boldsymbol{\Sigma}_{\boldsymbol{W}}}(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_{\boldsymbol{W}}}\left[\frac{1}{2}\left[\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{W}}) - \log|\boldsymbol{\Sigma}_{\boldsymbol{W}}|\right] + \frac{\gamma}{2}\mathrm{Tr}(\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{W}})\right]$$

$$= \frac{1}{2}\left(\mathbf{I}_m - \boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\right) + \frac{\gamma}{2}\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}$$

$$= \frac{1}{2}\left[\mathbf{I}_m - \boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1} + \gamma\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\right].$$

□

By setting the gradient $\nabla\boldsymbol{\Gamma}_1(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$ in **Lemma** 14 to $\langle \mathbf{0}_{n\times m}, \mathbf{0}_{m\times n}, \mathbf{0}_{n\times n}, \mathbf{0}_{m\times m}\rangle$, we get the

optimal solution for the γ-VAE loss function (7) as follows:

$$\begin{aligned}
\frac{\partial \mathbf{\Gamma}_1}{\partial \mathbf{A}}(\mathbf{A},\mathbf{B},\mathbf{\Sigma_Z},\mathbf{\Sigma_W}) = \mathbf{0}_{n\times m} &\Leftrightarrow \mathbf{\Sigma_Z}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W}) = \mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} \\
&\Leftrightarrow \mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W}) = \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} \\
&\Leftrightarrow \mathbf{A} = \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W})^{-1} \\
\frac{\partial \mathbf{\Gamma}_1}{\partial \mathbf{B}}(\mathbf{A},\mathbf{B},\mathbf{\Sigma_Z},\mathbf{\Sigma_W}) = \mathbf{0}_{m\times n} &\Leftrightarrow \mathbf{B}\mathbf{\Sigma_Y} - \gamma[\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y} - \mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A}\mathbf{B}\mathbf{\Sigma_Y}] = \mathbf{0}_{m\times n} \\
&\Leftrightarrow \mathbf{B} - \gamma[\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1} - \mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A}\mathbf{B}] = \mathbf{0}_{m\times n} \\
&\Leftrightarrow (\mathbf{I}_m + \gamma\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A})\mathbf{B} = \gamma\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1} \\
&\Leftrightarrow \mathbf{B} = \gamma(\mathbf{I}_m + \gamma\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1} \\
&\Leftrightarrow \mathbf{B} = (\mathbf{I}_m + \mathbf{A}^\mathsf{T}[\mathbf{\Sigma_Z}/\gamma]^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}[\mathbf{\Sigma_Z}/\gamma]^{-1} \\
\frac{\partial \mathbf{\Gamma}_1}{\partial \mathbf{\Sigma_Z}}(\mathbf{A},\mathbf{B},\mathbf{\Sigma_Z},\mathbf{\Sigma_W}) = \mathbf{0}_{n\times n} &\Leftrightarrow \mathbf{A}\mathbf{B}\mathbf{\Sigma_Y} + \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T} - \mathbf{\Sigma_Y} - \mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W})\mathbf{A}^\mathsf{T} + \mathbf{\Sigma_Z} = \mathbf{0}_{n\times n} \\
&\Leftrightarrow \mathbf{\Sigma_Z} = \mathbf{\Sigma_Y} + \mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W})\mathbf{A}^\mathsf{T} - \mathbf{A}\mathbf{B}\mathbf{\Sigma_Y} - \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T} \\
\frac{\partial \mathbf{\Gamma}_1}{\partial \mathbf{\Sigma_W}}(\mathbf{A},\mathbf{B},\mathbf{\Sigma_Z},\mathbf{\Sigma_W}) = \mathbf{0}_{m\times m} &\Leftrightarrow \mathbf{I}_m - \mathbf{\Sigma_W}^{-1} + \gamma\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A} = \mathbf{0}_{m\times m} \\
&\Leftrightarrow \mathbf{\Sigma_W}^{-1} = \mathbf{I}_m + \gamma\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A} \\
&\Leftrightarrow \mathbf{\Sigma_W} = (\mathbf{I}_m + \gamma\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A})^{-1} \\
&\Leftrightarrow \mathbf{\Sigma_W} = (\mathbf{I}_m + \mathbf{A}^\mathsf{T}[\mathbf{\Sigma_Z}/\gamma]^{-1}\mathbf{A})^{-1}.
\end{aligned}$$

Substituting $\mathbf{A} = \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W})^{-1}$ into the expression for $\mathbf{\Sigma_Z}$ and utilizing the formula for the inverse of a sum of matrices [3], we arrive at the following result:

$$\begin{aligned}
\mathbf{\Sigma_Z} &= \mathbf{\Sigma_Y} - \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W})^{-1}\mathbf{B}\mathbf{\Sigma_Y} \\
&= (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})^{-1}.
\end{aligned}$$

We now prove that matrix $\mathbf{A}$ can be reformulated as

$$\mathbf{A} = (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})^{-1}\mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}.$$

To demonstrate this, it is sufficient to verify that

$$\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W})^{-1} = (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})^{-1}\mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}. \tag{39}$$

To achieve this goal, we perform both the left and right multiplication of the matrix on the left-hand side (LHS) of equation (39) by $(\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})$ and $(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W})$, respectively. We apply an analogous process to the right-hand side (RHS) of equation (39). This yields:

$$\begin{aligned}
(\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})(\text{LHS})(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W}) &= (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} \\
&= \mathbf{B}^\mathsf{T} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} \\
(\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})(\text{RHS})(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W}) &= \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}+\mathbf{\Sigma_W}) \\
&= \mathbf{B}^\mathsf{T} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}.
\end{aligned}$$

As the outcomes obtained from the multiplications on both the left-hand side (LHS) and the right-hand side (RHS) of equation (39) coincide, equation (39) holds true.

B.1.4 Lemma 2

Proof. The proof of **Lemma** 2 follows a similar structure to that of **Lemma** 1. By [6], the gradient of $\gamma\lambda$-VAE cost function $\mathbf{\Gamma}_2(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W})$ can be computed as

$$\nabla\mathbf{\Gamma}_2(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}) = \Big\langle \frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{A}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}), \frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{B}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}), \\ \frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{\Sigma_Z}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}), \frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{\Sigma_W}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}) \Big\rangle,$$

where

$$\frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{A}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}) = \gamma[\mathbf{\Sigma_Z}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W}) - \mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}] + 2\lambda[\mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W}) - \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}]$$
$$\frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{B}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}) = \mathbf{B}\mathbf{\Sigma_Y} - \gamma[\mathbf{B}\mathbf{\Sigma_Y} + \mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y} - (\mathbf{I}_m + \mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A})\mathbf{B}\mathbf{\Sigma_Y}] + 2\lambda(\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{B}\mathbf{\Sigma_Y} - \mathbf{A}^\mathsf{T}\mathbf{\Sigma_Y})$$
$$\frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{\Sigma_W}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}) = \frac{1}{2}[\mathbf{I}_m - \mathbf{\Sigma_W}^{-1} + \gamma\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1}\mathbf{A}] + \lambda\mathbf{A}^\mathsf{T}\mathbf{A}$$

$$\frac{\partial\mathbf{\Gamma}_2}{\partial\mathbf{\Sigma_Z}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W})$$
$$= \frac{\gamma}{2}\left[\mathbf{\Sigma_Z}^{-1}\mathbf{A}\mathbf{B}\mathbf{\Sigma_Y}\mathbf{\Sigma_Z}^{-1} + \mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1} - \mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Y}\mathbf{\Sigma_Z}^{-1} - \mathbf{\Sigma_Z}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W})\mathbf{A}^\mathsf{T}\mathbf{\Sigma_Z}^{-1} + \mathbf{\Sigma_Z}^{-1}\right].$$

By letting $\nabla\mathbf{\Gamma}_2(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma_Z}, \mathbf{\Sigma_W}) = \langle \mathbf{0}_{n\times m}, \mathbf{0}_{m\times n}, \mathbf{0}_{n\times n}, \mathbf{0}_{m\times m}\rangle$, we get the optimal solution to $\gamma\lambda$-VAE loss function (8) as follows

$$\begin{aligned}
\mathbf{A} &= \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W})^{-1} \\
&= (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})^{-1}\mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1} \\
\mathbf{B} &= [\mathbf{I}_m + \mathbf{A}^\mathsf{T}(\gamma\mathbf{\Sigma_Z}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}]^{-1}\mathbf{A}^\mathsf{T}(\gamma\mathbf{\Sigma_Z}^{-1} + 2\lambda\mathbf{I}_n) \\
\mathbf{\Sigma_Z} &= \mathbf{\Sigma_Y} - \mathbf{A}\mathbf{B}\mathbf{\Sigma_Y} - \mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T} + \mathbf{A}(\mathbf{B}\mathbf{\Sigma_Y}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma_W})\mathbf{A}^\mathsf{T} \\
&= (\mathbf{\Sigma_Y}^{-1} + \mathbf{B}^\mathsf{T}\mathbf{\Sigma_W}^{-1}\mathbf{B})^{-1} \\
\mathbf{\Sigma_W} &= [\mathbf{I}_m + \mathbf{A}^\mathsf{T}(\gamma\mathbf{\Sigma_Z}^{-1} + 2\lambda\mathbf{I}_n)\mathbf{A}]^{-1}.
\end{aligned}$$

□

B.2 Proofs in Section 4

B.2.1 Lemma 3

Proof. By the Blahut-Arimoto algorithm [1, 2], for the fixed decoder $\boldsymbol{Y}^{(t)}$, the encoder $\boldsymbol{X}^{(t+1)}$ can be updated by performing the following equation:

$$q^{(t+1)}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p^{(t)}_{\boldsymbol{X}}(\boldsymbol{x})e^{-\gamma[d(\boldsymbol{X},\boldsymbol{Y})]^{(t)}}}{\int p^{(t)}_{\boldsymbol{X}}(\boldsymbol{x})e^{-\gamma[d(\boldsymbol{X},\boldsymbol{Y})]^{(t)}}d\boldsymbol{x}}, \tag{40}$$

where $[d(\boldsymbol{X}, \boldsymbol{Y})]^{(t)} = -\log p^{(t)}_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})$. Since the conditional distribution of decoder given encoder $p^{(t)}_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x}) \sim \mathcal{N}\left(\mathbf{A}^{(t)}\hat{\boldsymbol{x}}^{(t)}, \mathbf{\Sigma}^{(t)}_{\boldsymbol{Z}}\right)$, by using equation (40), the encoder $\boldsymbol{X}^{(t+1)}$ can be updated as follows:

$$q^{(t+1)}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})$$

$$
\begin{aligned}
&= \frac{p_{\boldsymbol{X}}^{(t)}(\boldsymbol{x}) e^{\gamma \log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t)}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x})}}{\int p_{\boldsymbol{X}}^{(t)}(\boldsymbol{x}) e^{\gamma \log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t)}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x})} d\boldsymbol{x}} \\
&= \frac{p_{\boldsymbol{X}}^{(t)}(\boldsymbol{x}) \left[p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t)}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x})\right]^{\gamma}}{\int p_{\boldsymbol{X}}^{(t)}(\boldsymbol{x}) \left[p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t)}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x})\right]^{\gamma} d\boldsymbol{x}} \\
&= \frac{\mathcal{N}(\mathbf{0}, \mathbf{I}_m) \left[\mathcal{N}\left(\mathbf{A}^{(t)}\boldsymbol{x}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)\right]^{\gamma}}{\int \mathcal{N}(\mathbf{0}, \mathbf{I}_m) \left[\mathcal{N}\left(\mathbf{A}^{(t)}\boldsymbol{x}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)\right]^{\gamma} d\boldsymbol{x}} \\
&= \frac{\frac{1}{(2\pi)^{(m+\gamma n)/2}\left|\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right|^{\gamma/2}} \exp\left(-\frac{1}{2}\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)} - \frac{\gamma}{2}\left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]\right)}{\int \frac{1}{(2\pi)^{(m+\gamma n)/2}\left|\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right|^{\gamma/2}} \exp\left(-\frac{1}{2}\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)} - \frac{\gamma}{2}\left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]\right) d\boldsymbol{x}} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)} + \left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]\right]\right)}{\int \exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)} + \left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]\right]\right) d\boldsymbol{x}}.
\end{aligned}
$$

Suppose $\mathbf{M}$ is a symmetric and non-singular matrix. By the completion-of-squares formula

$$\boldsymbol{x}^{\mathsf{T}}\mathbf{M}\boldsymbol{x} + 2\boldsymbol{b}^{\mathsf{T}}\boldsymbol{x} = (\boldsymbol{x} + \mathbf{M}^{-1}\boldsymbol{b})^{\mathsf{T}}\mathbf{M}(\boldsymbol{x} + \mathbf{M}^{-1}\boldsymbol{b}) - \boldsymbol{b}^{\mathsf{T}}\mathbf{M}^{-1}\boldsymbol{b},$$

this gives

$$
\begin{aligned}
&\left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)} + \left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right] \\
&= \left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\boldsymbol{x}^{(t)} + \left(\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}} - \left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\right)\left(\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\boldsymbol{y}^{(t)} - \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\boldsymbol{x}^{(t)}\right) \\
&= \left[\boldsymbol{x}^{(t)}\right]^{\mathsf{T}}\left(\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right)\boldsymbol{x}^{(t)} + 2\left(-\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right)\boldsymbol{x}^{(t)} \\
&\quad + \left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\boldsymbol{y}^{(t)} \\
&= \left[\boldsymbol{x}^{(t)} - \mathbf{D}^{-1}\boldsymbol{u}\right]^{\mathsf{T}}\mathbf{D}\left[\boldsymbol{x}^{(t)} - \mathbf{D}^{-1}\boldsymbol{u}\right] - \boldsymbol{u}^{\mathsf{T}}\mathbf{D}^{-1}\boldsymbol{u} + \left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\boldsymbol{y}^{(t)},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{D} &= \mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)} \\
\boldsymbol{u}^{\mathsf{T}} &= \left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}.
\end{aligned}
$$

So, the updated encoder at time $t+1$ can be reformulated as

$$q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}^{(t+1)}(\boldsymbol{x}|\boldsymbol{y})$$

$$
\begin{aligned}
&= \frac{\exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]^{\mathsf{T}}\mathbf{D}\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]-\boldsymbol{u}^{\mathsf{T}}\mathbf{D}^{-1}\boldsymbol{u}+\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\boldsymbol{y}^{(t)}\right]\right)}{\int \exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]^{\mathsf{T}}\mathbf{D}\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]-\boldsymbol{u}^{\mathsf{T}}\mathbf{D}^{-1}\boldsymbol{u}+\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\boldsymbol{y}^{(t)}\right]\right)d\boldsymbol{x}} \\
&= \frac{\exp\left(\frac{1}{2}\boldsymbol{u}^{\mathsf{T}}\mathbf{D}^{-1}\boldsymbol{u}-\frac{1}{2}\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\boldsymbol{y}^{(t)}\right)\exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]^{\mathsf{T}}\mathbf{D}\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]\right]\right)}{\exp\left(\frac{1}{2}\boldsymbol{u}^{\mathsf{T}}\mathbf{D}^{-1}\boldsymbol{u}-\frac{1}{2}\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\boldsymbol{y}^{(t)}\right)\int \exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]^{\mathsf{T}}\mathbf{D}\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]\right]\right)d\boldsymbol{x}} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]^{\mathsf{T}}\mathbf{D}\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]\right]\right)}{\int \exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]^{\mathsf{T}}\mathbf{D}\left[\boldsymbol{x}^{(t)}-\mathbf{D}^{-1}\boldsymbol{u}\right]\right]\right)d\boldsymbol{x}} \\
&= \mathcal{N}\left(\mathbf{D}^{-1}\boldsymbol{u},\mathbf{D}^{-1}\right).
\end{aligned}
$$

Hence, through the utilization of the Blahut-Arimoto algorithm with a fixed decoder, we can iteratively fine-tune the encoder until convergence is achieved. This iterative process enables us to identify the optimal encoder that minimizes the γ-VAE loss function (7), as expressed by the following equation:

$$
q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}^{(t+1)}(\boldsymbol{x}|\boldsymbol{y}) \sim \mathcal{N}\left(\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)},\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right),
$$

where

$$
\begin{aligned}
\mathbf{B}^{(t+1)} &= \left[\mathbf{I}_m+\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right]^{-1}\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \\
\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)} &= \left[\mathbf{I}_m+\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right]^{-1}.
\end{aligned}
$$

□

B.2.2 Lemma 4

Proof. According to [9], the conditional distribution of the decoder given the encoder is defined as follows:

$$
p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x}) = \frac{p_{\boldsymbol{Y}}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})}{\int p_{\boldsymbol{Y}}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{y}}. \tag{41}
$$

Notice that the conditional distribution of the encoder given the decoder is denoted as $q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}^{(t+1)}(\boldsymbol{x}|\boldsymbol{y}) \sim \mathcal{N}\left(\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)},\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$. Combining this result with equation (41), we can iteratively update the decoder until convergence is reached, following this procedure:

$$
\begin{aligned}
&p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t+1)}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x}) \\
&= \frac{p_{\boldsymbol{Y}}^{(t)}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}^{(t+1)}(\boldsymbol{x}|\boldsymbol{y})}{\int p_{\boldsymbol{Y}}^{(t)}(\boldsymbol{y})q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}^{(t+1)}(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{y}} \\
&= \frac{\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{Y}})\mathcal{N}\left(\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)},\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)}{\int \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{Y}})\mathcal{N}\left(\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)},\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)d\boldsymbol{y}} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{y}^{(t)}-\frac{1}{2}\left[\boldsymbol{x}^{(t+1)}-\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\left[\boldsymbol{x}^{(t+1)}-\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}\right]\right)}{\int \exp\left(-\frac{1}{2}\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{y}^{(t)}-\frac{1}{2}\left[\boldsymbol{x}^{(t+1)}-\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\left[\boldsymbol{x}^{(t+1)}-\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}\right]\right)d\boldsymbol{y}}.
\end{aligned}
$$

Similar to **Lemma** 3, employing the completion-of-squares formula yields:

$$
\left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{y}^{(t)}+\left[\boldsymbol{x}^{(t+1)}-\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\left[\boldsymbol{x}^{(t+1)}-\mathbf{B}^{(t+1)}\boldsymbol{y}^{(t)}\right]
$$

$$
\begin{aligned}
&= \left[\boldsymbol{y}^{(t)}\right]^{\mathsf{T}} \left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right) \boldsymbol{y}^{(t)} + 2\left(-\left[\boldsymbol{x}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right) \boldsymbol{y}^{(t)} \\
&\quad + \left[\boldsymbol{x}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \boldsymbol{x}^{(t+1)} \\
&= \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]^{\mathsf{T}} \mathbf{E} \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right] - \boldsymbol{v}^{\mathsf{T}}\mathbf{E}^{-1}\boldsymbol{v} + \left[\boldsymbol{x}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \boldsymbol{x}^{(t+1)},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{E} &= \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)} \\
\boldsymbol{v}^{\mathsf{T}} &= \left[\boldsymbol{x}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}.
\end{aligned}
$$

Thus, the updated decoder at time $t+1$ can be rewritten as

$$
\begin{aligned}
&p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t+1)}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x}) \\
&= \frac{\exp\left(-\frac{1}{2}\left[\left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]^{\mathsf{T}} \mathbf{E} \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right] - \boldsymbol{v}^{\mathsf{T}}\mathbf{E}^{-1}\boldsymbol{v} + \left[\boldsymbol{x}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \boldsymbol{x}^{(t+1)}\right]\right)}{\int \exp\left(-\frac{1}{2}\left[\left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]^{\mathsf{T}} \mathbf{E} \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right] - \boldsymbol{v}^{\mathsf{T}}\mathbf{E}^{-1}\boldsymbol{v} + \left[\boldsymbol{x}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \boldsymbol{x}^{(t+1)}\right]\right) d\boldsymbol{y}} \\
&= \frac{\exp\left(-\frac{1}{2}\left[\left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]^{\mathsf{T}} \mathbf{E} \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]\right]\right)}{\int \exp\left(-\frac{1}{2}\left[\left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]^{\mathsf{T}} \mathbf{E} \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]\right]\right) d\boldsymbol{y}} \\
&= \mathcal{N}\left(\mathbf{E}^{-1}\boldsymbol{v}, \mathbf{E}^{-1}\right).
\end{aligned}
$$

Therefore, by employing the Blahut-Arimoto algorithm with a fixed encoder, we can iteratively refine the decoder until convergence is attained. This iterative procedure allows us to ascertain the optimal decoder that minimizes the γ-VAE loss function (7), as described by the following equation:

$$
p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}^{(t+1)}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x}) \sim \mathcal{N}\left(\mathbf{A}^{(t+1)}\boldsymbol{x}^{(t+1)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right),
$$

where

$$
\begin{aligned}
\mathbf{A}^{(t+1)} &= \left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right)^{-1} \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \\
\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)} &= \left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right)^{-1}.
\end{aligned}
$$

□

B.3 Proofs in Section 6

B.3.1 Case 1: $\gamma < 1$

Lemma 15. *For any integer $t \geq 1$, we have $\left\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right\| \leq \|\mathbf{I}_m\|$.*

Proof. By **Lemma 3**, the covariance matrix of the encoder noise at iteration $t+1$ is given by

$$
\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)} = \left(\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \mathbf{A}^{(t)}\right)^{-1},
$$

which is equivalent to

$$
\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} - \mathbf{I}_m = \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \mathbf{A}^{(t)}.
$$

Since $\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \mathbf{A}^{(t)}$ is positive semi-definite,

$$
\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \succeq \mathbf{I}_m \quad \Leftrightarrow \quad \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)} \preceq \mathbf{I}_m.
$$

So for all $i = 1, \cdots, m$, the i-th eigenvalue of $\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}$ is less than or equal to the i-th eigenvalue of $\mathbf{I}_m$, i.e.,

$$\lambda_i\left(\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right) \leq \lambda_i(\mathbf{I}_m) = 1.$$

Since $\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}$ is positive definite, its eigenvalues are positive. Then we have

$$\left\|\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right\| = \left|\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)\right| = \lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right) \leq 1 = \|\mathbf{I}_m\|.$$

□

Lemma 16. *When $\gamma < 1$, the algorithm returns a unique trivial optimal solution $(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) = (\mathbf{0}, \mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{Y}}, \mathbf{I}_m)$.*

Proof. We first prove that $\lim_{t\to\infty} \mathbf{B}^{(t)} = 0$. If we start with the encoder $\left(\mathbf{B}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right)$, by **Lemma 4**, the decoder $\left(\mathbf{A}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ can be updated using the following equations:

$$\begin{aligned}
\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)} &= \left(\mathbf{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\right)^{-1} \\
\mathbf{A}^{(t)} &= \left(\mathbf{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\right)^{-1}\left[\mathbf{B}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \\
&= \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\left[\mathbf{B}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}.
\end{aligned}$$

Then, we have

$$\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} = \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}. \tag{42}$$

Using the decoder inputs from iteration t, the encoder at iteration $t+1$ can be updated as

$$\begin{aligned}
\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)} &= \left(\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right)^{-1} \\
\mathbf{B}^{(t+1)} &= \left[\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}\mathbf{A}^{(t)}\right]^{-1}\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \\
&= \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1}.
\end{aligned}$$

This gives

$$\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)} = \gamma\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}. \tag{43}$$

From equations (42) and (43), for $\gamma < 1$,

$$\left\|\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\right\| = \left\|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\right\| \geq \gamma\left\|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\right\| = \left\|\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)}\right\| \geq 0.$$

Hence, the l^2-norm $\left\|\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\right\|$ tends to decrease as t increases. We now prove that $\left\|\mathbf{B}^{(t)}\right\|$ is pushed to 0 as t increases. Notice that

$$\begin{aligned}
\mathbf{B}^{(t+2)} &= \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+2)}\left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}/\gamma\right]^{-1} \\
&= \gamma\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+2)}\left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \\
&= \gamma\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+2)}\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)} \\
&= \gamma^2\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+2)}\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \\
&= \gamma^2\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+2)}\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}.
\end{aligned}$$

So, for any integer $n \geq 1$,

$$\mathbf{B}^{(t+n)} = \gamma^n \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}.$$

For $\gamma < 1$, by **Lemma 15**, the limit of l^2-norm of $\mathbf{B}^{(t+n)}$ as n goes to infinity is then given by

$$\begin{aligned}
0 \leq \lim_{n\to\infty} \left\|\mathbf{B}^{(t+n)}\right\| &= \lim_{n\to\infty} \left\|\gamma^n \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right\| \\
&= \left(\lim_{n\to\infty} \gamma^n\right) \lim_{n\to\infty} \left\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right\| \\
&\leq \left(\lim_{n\to\infty} \gamma^n\right) \left(\lim_{n\to\infty} \left\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right\|\right) \left\|\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right\| \\
&\leq \left(\lim_{n\to\infty} \gamma^n\right) \left(\lim_{n\to\infty} \|\mathbf{I}_m\|\right) \left\|\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right\| \\
&= 0 \cdot 1 \cdot \left\|\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right\| \\
&= 0.
\end{aligned}$$

By the Squeeze theorem, $\lim_{n\to\infty} \left\|\mathbf{B}^{(t+n)}\right\| = 0$. Thus,

$$\lim_{t\to\infty} \left\|\mathbf{B}^{(t)}\right\| = 0 \quad \Leftrightarrow \quad \lim_{t\to\infty} \mathbf{B}^{(t)} = \mathbf{0}.$$

Since $\lim_{t\to\infty} \mathbf{B}^{(t)} = \mathbf{0}$, we have

$$\begin{aligned}
\lim_{t\to\infty} \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} &= \lim_{t\to\infty} \left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \lim_{t\to\infty} \left(\left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{0} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}.
\end{aligned}$$

So, $\lim_{t\to\infty} \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)} = \boldsymbol{\Sigma}_{\boldsymbol{Y}}$. Since $\mathbf{A}^{(t)} = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)} \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}$ and $\lim_{t\to\infty} \mathbf{B}^{(t)} = \mathbf{0}$, the limit of $\mathbf{A}^{(t)}$ is given by

$$\begin{aligned}
\lim_{t\to\infty} \mathbf{A}^{(t)} &= \lim_{t\to\infty} \left(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)} \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\right) \\
&= \left(\lim_{t\to\infty} \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right) \cdot \mathbf{0} \cdot \left(\lim_{t\to\infty} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\right) \\
&= \mathbf{0}.
\end{aligned}$$

Using $\lim_{t\to\infty} \mathbf{A}^{(t)} = \mathbf{0}$, this gives

$$\begin{aligned}
\lim_{t\to\infty} \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} &= \lim_{t\to\infty} \left(\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \mathbf{A}^{(t)}\right) \\
&= \mathbf{I}_m + \lim_{t\to\infty} \left(\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \mathbf{A}^{(t)}\right) \\
&= \mathbf{I}_m + \mathbf{0} \\
&= \mathbf{I}_m,
\end{aligned}$$

which is equivalent to $\lim_{t\to\infty} \boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)} = \mathbf{I}_m$.

Thus, the algorithm converges to a trivial optimal solution $(\mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{W}}) = (\mathbf{0}, \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Y}}, \mathbf{I}_m)$ for the case of $\gamma < 1$. □

B.3.2 Case 3: $\gamma > 1$

Lemma 17. *For any positive semi-definite matrix* $\mathbf{M}$*, we have* $\|(\mathbf{I}_m + \mathbf{M})^{-1}\| \leq \|\mathbf{M}^{-1}\|$.

Proof. Since $\mathbf{I}_m + \mathbf{M} \succeq \mathbf{M}$, its inverse

$$(\mathbf{I}_m + \mathbf{M})^{-1} \preceq \mathbf{M}^{-1}.$$

So for all $i = 1, \cdots, m$, the i-th eigenvalue of $(\mathbf{I}_m + \mathbf{M})^{-1}$ is less than or equal to the i-th eigenvalue of $\mathbf{M}^{-1}$, i.e.,

$$\lambda_i\left((\mathbf{I}_m + \mathbf{M})^{-1}\right) \leq \lambda_i(\mathbf{M}^{-1}).$$

Since $\mathbf{I}_m + \mathbf{M}$ and $\mathbf{M}$ are symmetric and the eigenvalues of both matrices are non-negative,

$$\begin{aligned}
\left\|(\mathbf{I}_m + \mathbf{M})^{-1}\right\| &= \left|\lambda_{\max}\left((\mathbf{I}_m + \mathbf{M})^{-1}\right)\right| \\
&= \lambda_{\max}\left((\mathbf{I}_m + \mathbf{M})^{-1}\right) \\
&\leq \lambda_{\max}\left(\mathbf{M}^{-1}\right) \\
&= \left|\lambda_{\max}\left(\mathbf{M}^{-1}\right)\right| \\
&= \left\|\mathbf{M}^{-1}\right\|.
\end{aligned}$$

□

Lemma 18. *When* $\gamma > 1$*, the algorithm fails to converge.*

Proof. We aim to demonstrate that as t approaches infinity, the covariance matrix of the encoder noise $\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}$ tends toward $\mathbf{0}$. However, this pattern can lead to a blow-up issue resulting in a singular matrix error. From **Lemma 16**, we obtained the system

$$\begin{aligned}
\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} &= \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)} \\
\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)} &= \gamma\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}.
\end{aligned}$$

For $\gamma > 1$, we have

$$\left\|\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)}\right\| = \gamma\left\|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\right\| \geq \left\|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\right\| = \left\|\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\right\| \geq 0.$$

Hence, the l^2-norm $\left\|\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\right\|$ tends to increase as t increases. We now prove that $\left\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right\|$ tends to be pushed to 0 as the iteration increases. Notice that matrix $\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+2)}$ can be described by

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+2)} &= \left(\mathbf{I}_m + \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}/\gamma\right]^{-1}\mathbf{A}^{(t+1)}\right)^{-1} \\
&= \left(\mathbf{I}_m + \gamma\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\mathbf{B}^{(t+1)}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\right)^{-1} \\
&= \left(\mathbf{I}_m + \gamma^3\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}\mathbf{A}^{(t)}\right)^{-1} \\
&= \left(\mathbf{I}_m + \gamma^3\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\left[\mathbf{B}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\right)^{-1}.
\end{aligned}$$

So, for any integer $n \geq 1$,

$$\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)} = \left(\mathbf{I}_m + \gamma^{2n-1}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\mathbf{B}^{(t)}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)}\left[\mathbf{B}^{(t)}\right]^{\mathsf{T}}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}\right)^{-1}.$$

For $\gamma > 1$, by **Lemma 17**, the limit of l^2-norm of $\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+n)}$ as n goes to infinity is then computed as

$$\begin{aligned}
&\lim_{n\to\infty} \left\|\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right\| \\
&= \lim_{n\to\infty} \left\| \left(\mathbf{I}_m + \gamma^{2n-1} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)} \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)} \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \right)^{-1} \right\| \\
&\leq \lim_{n\to\infty} \left\| \left(\gamma^{2n-1} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)} \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)} \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \right)^{-1} \right\| \\
&= \left(\lim_{n\to\infty} \frac{1}{\gamma^{2n-1}} \right) \lim_{n\to\infty} \left\| \left(\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)} \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)} \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \right)^{-1} \right\| \\
&= 0 \cdot \lim_{n\to\infty} \left\| \left(\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)} \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)} \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \right)^{-1} \right\| \\
&= 0.
\end{aligned}$$

Since $0 \leq \lim_{n\to\infty} \left\|\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right\| \leq 0$, by the Squeeze theorem, $\lim_{n\to\infty} \left\|\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right\| = \mathbf{0}$. Thus, $\lim_{t\to\infty} \mathbf{\Sigma}_{\boldsymbol{W}}^{(t)} = 0$. However, this contradicts the assumption of $\mathbf{\Sigma}_{\boldsymbol{W}}$ being a positive definite matrix. In addition, since the decoder inputs $\left(\mathbf{A}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ must be updated using $\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}$ while this inverse does not exist, the algorithm will fail to converge due to the singular matrix error. □

B.4 Proofs in Section 7

B.4.1 Case 3: $\gamma > 1$

Lemma 19. *For any integer $t \geq 1$, we have the following inequality:*

$$\begin{aligned}
&1 \Big/ \left| \mathbf{I}_m + \gamma \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)} \right| \\
&\leq \frac{1}{\gamma^m} \left[1 \Big/ \left| \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)} \right| \right].
\end{aligned}$$

Proof. By the Minkowski determinant theorem, if $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ positive semi-definite Hermitian matrices, then

$$(\det(\mathbf{A} + \mathbf{B}))^{1/n} \geq (\det \mathbf{A})^{1/n} + (\det \mathbf{B})^{1/n}.$$

For $n = 1$, the theorem becomes

$$|\mathbf{A} + \mathbf{B}| \geq |\mathbf{A}| + |\mathbf{B}|.$$

Since $\mathbf{I}_m$ and $\gamma \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)}$ are $m \times m$ positive semi-definite matrices, this gives

$$\begin{aligned}
&\left| \mathbf{I}_m + \gamma \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)} \right| \\
&\geq |\mathbf{I}_m| + \left| \gamma \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)} \right| \\
&= 1 + \left| \gamma \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)} \right| \\
&= 1 + \gamma^m \left| \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)} \right| \\
&\geq \gamma^m \left| \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n \right) \mathbf{A}^{(t+1)} \right|.
\end{aligned}$$

□

Lemma 20. *When $\gamma > 1$, the algorithm fails to converge.*

Proof. Similar to **Lemma 18**, we want to prove that the covariance matrix of the encoder noise $\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}$ is pushed to a singular matrix as t goes to infinity, which contradicts the assumption of $\mathbf{\Sigma}_{\boldsymbol{W}}$ being a positive definite matrix. Given the encoder $\left(\mathbf{B}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right)$, from **Lemma 16**, we obtained that the decoder $\left(\mathbf{A}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ can be updated as follows:

$$\begin{aligned}
\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)} &= \left(\mathbf{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}\right)^{-1} \\
\mathbf{A}^{(t)} &= \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)} \left[\mathbf{B}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}.
\end{aligned}$$

As $\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}$ is diagonalized, the subsequent update of the encoder at iteration $t+1$ can be expressed as:

$$\begin{aligned}
\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)} &= \left(\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\left(\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)} \circ \mathbf{I}_n\right)/\gamma\right]^{-1} \mathbf{A}^{(t)}\right)^{-1} \\
&= \left[\mathbf{I}_m + \gamma \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{A}^{(t)}\right]^{-1} \\
\mathbf{B}^{(t+1)} &= \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)} \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\left(\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)} \circ \mathbf{I}_n\right)/\gamma\right]^{-1} \\
&= \gamma \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)} \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right).
\end{aligned}$$

From the two aforementioned systems, we can deduce the following system:

$$\begin{aligned}
\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} &= \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)} \\
\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)} &= \gamma \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right).
\end{aligned} \tag{44}$$

Using system (44) and by **Lemma 19**, the determinant of matrix $\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+2)}$ can be described by

$$\begin{aligned}
&\left|\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+2)}\right| \\
&= \left|\left[\mathbf{I}_m + \gamma \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{A}^{(t+1)}\right]^{-1}\right| \\
&= 1 \Big/ \left|\mathbf{I}_m + \gamma \left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{A}^{(t+1)}\right| \\
&\leq \frac{1}{\gamma^m} \left[1 \Big/ \left|\left[\mathbf{A}^{(t+1)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{A}^{(t+1)}\right|\right] \\
&= \frac{1}{\gamma^m} \left[1 \Big/ \left|\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)} \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)} \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1}\right|\right] \\
&= \frac{1}{\gamma^m} \left[1 \Big/ \left|\gamma^2 \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{A}^{(t)}\right|\right] \\
&= \frac{1}{\gamma^m} \frac{1}{\gamma^{2m}} \left[1 \Big/ \left|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right) \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right] \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right]^{-1} \circ \mathbf{I}_n\right) \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right)\right] \mathbf{A}^{(t)}\right|\right].
\end{aligned}$$

So for any integer $n \geq 1$, the determinant of $\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+n)}$ satisfies the following inequality:

$$\left|\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right| \leq \frac{1}{\gamma^m} \left(\frac{1}{\gamma^{2m}}\right)^{n-1} \left[1 \Big/ \left|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \mathbf{M}^{(n)} \mathbf{A}^{(t)}\right|\right],$$

where

$$\mathbf{M}^{(n)} = \left[\left(\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right)\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right] \cdots \left[\left(\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+n-2)}\right]^{-1} \circ \mathbf{I}_n\right)\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)}\right] \cdot \left(\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)}\right]^{-1} \circ \mathbf{I}_n\right) \cdot \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+n-1)}\left(\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+n-2)}\right]^{-1} \circ \mathbf{I}_n\right)\right] \cdots \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\left(\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right)\right].$$

For $\gamma > 1$, the limit of the determinant of $\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)}$ as n goes to infinity can be described as follows:

$$\begin{aligned}
&\lim_{n\to\infty}\left|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right| \\
&\leq \lim_{n\to\infty}\left(\frac{1}{\gamma^m}\left(\frac{1}{\gamma^{2m}}\right)^{n-1}\left[1\Big/\left|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\mathbf{M}^{(n)}\mathbf{A}^{(t)}\right|\right]\right) \\
&= \frac{1}{\gamma^m}\left[\lim_{n\to\infty}\left(\frac{1}{\gamma^{2m}}\right)^{n-1}\right]\lim_{n\to\infty}\left[1\Big/\left|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\mathbf{M}^{(n)}\mathbf{A}^{(t)}\right|\right] \\
&= \frac{1}{\gamma^m}\cdot 0\cdot\lim_{n\to\infty}\left[1\Big/\left|\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}}\mathbf{M}^{(n)}\mathbf{A}^{(t)}\right|\right] \\
&= 0.
\end{aligned}$$

Given that $0 \leq \lim_{n\to\infty}\left|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right| \leq 0$, the Squeeze theorem implies that $\lim_{n\to\infty}\left|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+n)}\right| = \mathbf{0}$. Consequently, this causes $\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}$ to converge towards a singular matrix. However, this contradicts the initial assumption that $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ is nonsingular. Moreover, the update of decoder inputs $\left(\mathbf{A}^{(t)}, \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ necessitates the utilization of $\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1}$. Because this inverse does not exist for a singular matrix, the algorithm's convergence fails due to the occurrence of a singular matrix error. □

B.5 Maximizing Mutual Information and VAE

We now argue that it is not robust to induce the maximization of mutual information in the setting of VAE by simply letting $\gamma > 1$. Consider the decoder and encoder:

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{\mu}_{de}(\boldsymbol{X}) + \boldsymbol{\sigma}_{de}(\boldsymbol{X}) \odot \boldsymbol{\xi}_1 \\
\boldsymbol{X} &= \boldsymbol{\mu}_{en}(\boldsymbol{Y}) + \boldsymbol{\sigma}_{en}(\boldsymbol{Y}) \odot \boldsymbol{\xi}_2,
\end{aligned}$$

where $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are standard Gaussians. We have

$$\begin{aligned}
-\log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) =& \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_{de}(\boldsymbol{x}))^{\mathsf{T}}\operatorname{diag}(\boldsymbol{\sigma}_{de}(\boldsymbol{x}))^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{de}(\boldsymbol{x})) \\
&+ \frac{1}{2}\log|\operatorname{diag}(\boldsymbol{\sigma}_{de}(\boldsymbol{x}))| + \text{cnst}.
\end{aligned}$$

To minimize $-\log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$, there are two ways: For fixed $\boldsymbol{\sigma}_{de}(\boldsymbol{x})$, the model of $\boldsymbol{\mu}_{de}(\boldsymbol{x})$ need to approach $\boldsymbol{y}$; and for fixed $\boldsymbol{\mu}_{de}(\boldsymbol{x})$, the model of $\boldsymbol{\sigma}_{de}(\boldsymbol{x})$ may go to zero. When the second case dominates, the problem may become illposed. So we consider the following case when we want to introduce the maximization of mutual information:

$$\text{loss of VAE} : -\text{ELBO} + \lambda\mathbb{E}_{q_{\boldsymbol{Y}}}[\|\boldsymbol{Y} - \mu_{de}(\boldsymbol{X}_{en}(\boldsymbol{Y}))\|^2],$$

where in the second term $\boldsymbol{X}$ depends on $\boldsymbol{Y}$ through the encoder. The second term corresponds to a simple model of $p_{\boldsymbol{Y}|\boldsymbol{X}}$ as

$$\boldsymbol{Y} = \boldsymbol{\mu}_{de}(\boldsymbol{X}) + \sigma^2\boldsymbol{\xi},$$

with σ being a constant, which yields that

$$-\log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu}_{de}(\boldsymbol{x}))^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{\mu}_{de}(\boldsymbol{x})) + n\log\sigma + \text{cnst}.$$

For this case, the maximization of the mutual information will be achieved by improving the model $\boldsymbol{\mu}_{de}(\boldsymbol{x})$, and $\boldsymbol{\sigma}_{de}(\boldsymbol{x})$ will not be used for the maximization of mutual information. The problem is always wellposed and the degree of the maximization of mutual information will be controlled by the value of λ, where a larger λ corresponds to a smaller σ, meaning that the maximization of mutual information is in favor of a smaller $\boldsymbol{\sigma}_{de}(\boldsymbol{x})$ in the VAE. This way, $\boldsymbol{\sigma}_{de}(\boldsymbol{x})$ will only be used for the density estimation through the maximization of ELBO and only $\mu_{de}(\boldsymbol{x})$ is needed for the maximization of the mutual information.