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

Minh Vu[1] Xiaoliang Wan[2] Shuangqing Wei[3]

August 12, 2023

Abstract

Variational Autoencoders (VAEs) have garnered substantial attention as generative models for producing lower-dimensional representations of high-dimensional data. The β-VAE model employs the hyperparameter β to strike a balance between reconstruction accuracy and disentanglement. This study exclusively targets the enhancement of reconstruction accuracy in the linear Gaussian β-VAE model by introducing three variants: γ-VAE with both arbitrary and diagonalized $\mathbf{\Sigma}_{\boldsymbol{Z}}$, as well as $\gamma\lambda$-VAE with diagonalized $\mathbf{\Sigma}_{\boldsymbol{Z}}$. We commence by deriving closed-form solutions for all three proposed frameworks using gradient-based and iterative methods. This demonstration of consistency between approaches highlights the robustness of our findings. Subsequently, we perform comprehensive numerical experiments employing the Blahut-Arimoto algorithm. These experiments underscore the benefits of utilizing a diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$ over an arbitrary one, leading to more informative numerical outcomes and augmented control over reconstruction accuracy. Furthermore, the introduction of an additional hyperparameter λ offers an avenue for further refining reconstruction accuracy control. In conclusion, the introduction of these three variants to the β-VAE model, combined with analytical and numerical analyses, underscores the potential for improved reconstruction accuracy through the strategic incorporation of additional hyperparameters and nuanced adjustments to the foundational framework.

1 Introduction

1.1 Main Objectives

The β-VAE model [8], an extension of the VAE framework [11], 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.

Student[1]: Minh Vu (Department of Mathematics, Louisiana State University).
Advisor[2]: Xiaoliang Wan (Department of Mathematics, Louisiana State University)
Co-Project Advisor[3]: Shuangqing Wei (Division of Electrical & Computer Engineering, Louisiana State University).

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 [10]. 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$, γ-VAE with diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$, and $\gamma\lambda$-VAE with diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

- 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ instead of an arbitrary positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$. 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 $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{W}})$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{Z}})$, 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ to investigate their impact on reconstruction accuracy. Consequently, the choice of structure for $\mathbf{\Sigma}_{\boldsymbol{W}}$ 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 $\mathbf{\Sigma}_{\boldsymbol{W}}$.

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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ accounting for noise added to the reconstructed output, in conjunction with an arbitrary positive definite matrix $\mathbf{\Sigma}_{\boldsymbol{W}}$ representing the noise introduced during encoding. The second variation corresponds to the γ-VAE employing a diagonalized positive definite matrix $\mathbf{\Sigma}_{\boldsymbol{Z}}$ along with an arbitrary positive definite matrix $\mathbf{\Sigma}_{\boldsymbol{W}}$.

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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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 $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{Y}})$ with a positive definite covariance matrix $\mathbf{\Sigma}_{\boldsymbol{Y}}$. 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 $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{W}})$, where $\mathbf{\Sigma}_{\boldsymbol{W}}$ 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 $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{Z}})$, where $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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:

$$\begin{aligned}
(\boldsymbol{\mu}_{\boldsymbol{X}}, \mathbf{\Sigma}_{\boldsymbol{X}}) &= (\mathbf{0}, \mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\mathsf{T} + \mathbf{\Sigma}_{\boldsymbol{W}}) \\
(\boldsymbol{\mu}_{\hat{\boldsymbol{Y}}}, \mathbf{\Sigma}_{\hat{\boldsymbol{Y}}}) &= (\mathbf{0}, \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{\Sigma}_{\boldsymbol{Z}}).
\end{aligned} \tag{4}$$

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.

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

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:

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

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 $\{\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}\}$. This section presents closed-form solutions for these frameworks, leveraging the linear Gaussian assumption. Specifically, we reformulate the loss function in terms of $\{\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}\}$ 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:*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Y}}[D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]]\\ &= \frac{1}{2}\left[\mathrm{Tr}(\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\top} + \mathbf{\Sigma}_{\boldsymbol{W}}) - \log|\mathbf{\Sigma}_{\boldsymbol{W}}| - m\right].\end{aligned}$$

Proof. Appendix B.1.1. □

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

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

$$= \frac{1}{2}\left(\mathrm{Tr}\left[\mathbf{A}^{\top}\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\top} + \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}} - \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^{\top}\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\top} + \mathbf{\Sigma}_{\boldsymbol{W}})\right] - n\log(2\pi) - \log|\mathbf{\Sigma}_{\boldsymbol{Z}}|\right).$$

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 $\{\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}\}$, denoted as $\mathbf{\Gamma}_1(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})$, as follows:*

$$\begin{aligned}&\mathbf{\Gamma}_1(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})\\ &= \frac{1}{2}\left[\mathrm{Tr}(\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\top} + \mathbf{\Sigma}_{\boldsymbol{W}}) - \log|\mathbf{\Sigma}_{\boldsymbol{W}}| - m\right]\\ &\quad - \frac{\gamma}{2}\Bigg(\mathrm{Tr}\Big[\mathbf{A}^{\top}\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\top} + \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}} - \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}}\\ &\quad - \mathbf{A}^{\top}\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\top} + \mathbf{\Sigma}_{\boldsymbol{W}})\Big] - n\log(2\pi) - \log|\mathbf{\Sigma}_{\boldsymbol{Z}}|\Bigg).\end{aligned}$$

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

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

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

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Y}}[\|\boldsymbol{Y} - \mathbf{A}\boldsymbol{X}\|^2]\\ &= \mathbb{E}_{\boldsymbol{Y}}[\|(\mathbf{I}_n - \mathbf{A}\mathbf{B})\boldsymbol{Y} - \mathbf{A}\boldsymbol{W}\|^2]\\ &= \mathbb{E}_{\boldsymbol{Y}}[[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\boldsymbol{Y} - \mathbf{A}\boldsymbol{W}]^{\top}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\boldsymbol{Y} - \mathbf{A}\boldsymbol{W}]]\\ &= \mathbb{E}_{\boldsymbol{Y}}[\mathrm{Tr}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\boldsymbol{Y}\boldsymbol{Y}^{\top}(\mathbf{I}_n - \mathbf{A}\mathbf{B})^{\top} + \mathbf{A}\boldsymbol{W}\boldsymbol{W}^{\top}\mathbf{A}^{\top}]]\\ &= \mathrm{Tr}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\mathbf{\Sigma}_{\boldsymbol{Y}}(\mathbf{I}_n - \mathbf{A}\mathbf{B})^{\top} + \mathbf{A}\mathbf{\Sigma}_{\boldsymbol{W}}\mathbf{A}^{\top}].\end{aligned}$$

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

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

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

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

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 [10] 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

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

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

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

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 $\left(\mathbf{A}^{(t)}\right.$ along with a diagonalized $\left.\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$, 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

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

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 $\mathbf{\Sigma}_{\boldsymbol{W}}$
- 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$
- switch $flag$ to 0.

4. Initialize iteration counter $t = 1$.

5. Iteration step:

(a) If $flag = 0$,

- first, update the encoder inputs $\boldsymbol{\phi}^{(t+1)} = \left(\mathbf{B}^{(t+1)}, \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$ given decoder $\boldsymbol{\theta}^{(t)} = \left(\mathbf{A}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ 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 $\mathbf{\Gamma}_1$ 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 $\mathbf{\Sigma}_{\boldsymbol{W}}$ and itself in the previous iteration to get $Sigma\_W\_norm\_diff$.
  iii. compute the difference between the cost function $\mathbf{\Gamma}_1$ 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 $\boldsymbol{\theta}^{(t)} = \left(\mathbf{A}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$ given encoder $\boldsymbol{\phi}^{(t)} = \left(\mathbf{B}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right)$ 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 $\mathbf{\Gamma}_1$ 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ and itself in the previous iteration to get $Sigma\_Z\_norm\_diff$.
  iii. compute the difference between the cost function $\mathbf{\Gamma}_1$ 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 $\mathbf{\Sigma}_{\boldsymbol{X}}$ and $\mathbf{\Sigma}_{\hat{\boldsymbol{Y}}}$.
9. Display results
   - display the optimal solution $(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})$.
   - display the minimum value of γ-VAE loss function.
   - display the mutual information of both encoder and decoder.
   - display the values of $\mathbf{\Sigma}_{\boldsymbol{X}}$ and $\mathbf{\Sigma}_{\hat{\boldsymbol{Y}}}$.
   - display the value of reconstruction error.
   - move to step 10.
10. Stop.

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

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

5.3 $\gamma\lambda$-VAE given Diagonalized $\mathbf{\Sigma}_{\boldsymbol{Z}}$

We can revise the algorithm for determining the optimal solution $(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})$ 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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 $(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})$ 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 $\mathbf{\Gamma}_2$ 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$

In our numerical investigations, our primary objective is to determine the optimal solution for γ-VAE, considering an arbitrary positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$. 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 $\mathbf{\Sigma}_{\boldsymbol{Y}}$ as follows:

$$\mathbf{\Sigma}_{\boldsymbol{Y}} = \begin{bmatrix} 1 & 0.18 & 0.12 \\ 0.18 & 1 & 0.06 \\ 0.12 & 0.06 & 1 \end{bmatrix}.$$

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
$(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})$	$(\mathbf{0}, \mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{Y}}, \mathbf{I})$
$\mathbf{\Sigma}_{\boldsymbol{X}}$	$\mathbf{I}$
$\mathbf{\Sigma}_{\hat{\boldsymbol{Y}}}$	$\mathbf{\Sigma}_{\boldsymbol{Y}}$
$I_{\boldsymbol{\phi}}(\boldsymbol{Y};\boldsymbol{X})$	0
$I_{\boldsymbol{\theta}}(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}})$	0
Minimum value of $\mathbf{\Gamma}_1$	4.1477
Reconstruction error	0

Table 1: Optimal solution for $\gamma = 0.98$ given arbitrary positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

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

Table 2: Optimal solutions for $\gamma = 1$ given arbitrary positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

To update the encoder $\left(\mathbf{B}^{(t+1)}, \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$ 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}$ 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 $\mathbf{\Sigma}_{\boldsymbol{W}} = \mathbf{I}$. 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 $\mathbf{\Gamma}_1$ can be computed as:

  $$\mathbf{\Gamma}_1(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) = \frac{\gamma}{2} \left[\log |\mathbf{\Sigma}_{\boldsymbol{Y}}| + n + n \log(2\pi) \right].$$

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

Table 3: Optimal solutions for $\gamma = 1.02$ given arbitrary positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

- 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 $\mathbf{\Gamma}_1$ 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 $\mathbf{\Gamma}_1$, 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$

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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

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

Table 4: Optimal solutions for $\gamma = 0.98$ given diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

7.1.2 Case 2: $\gamma = 1$

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

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

Table 5: Optimal solutions for $\gamma = 1$ given diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

7.1.3 Case 3: $\gamma = 1.02$

Similar to the case of arbitrary positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$ as discussed in Section 6.1.3, the algorithm encounters convergence issues attributed to NaN or singular matrices when a diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$

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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$. 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

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

Table 6: Optimal solutions for $\gamma = 1.02$ given diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ in lieu of an arbitrary positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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.

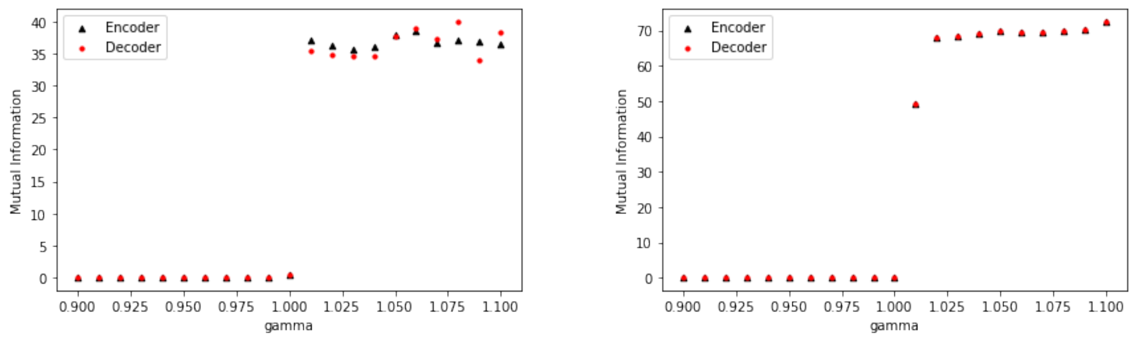

Figure 1: Plot of mutual information w.r.t. γ, given arbitrary $\mathbf{\Sigma_Z}$ (left) and diagonalized $\mathbf{\Sigma_Z}$ (right).

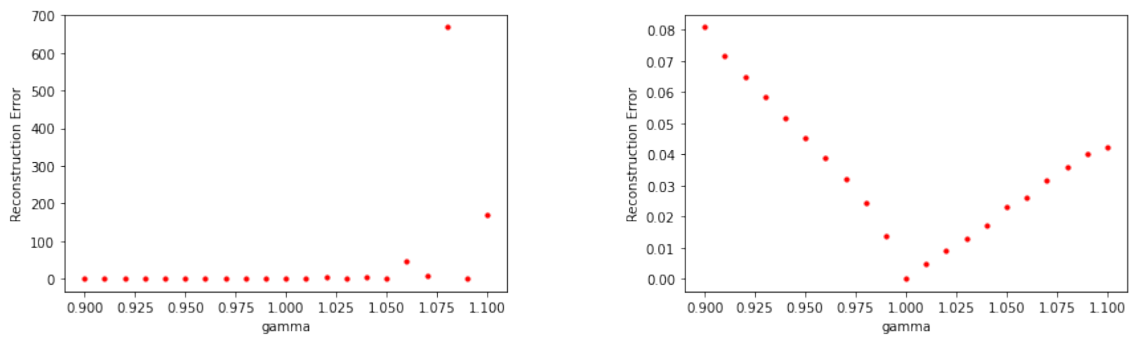

Figure 2: Plot of reconstruction error w.r.t. γ, given arbitrary $\mathbf{\Sigma_Z}$ (left) and diagonalized $\mathbf{\Sigma_Z}$ (right).

- 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.

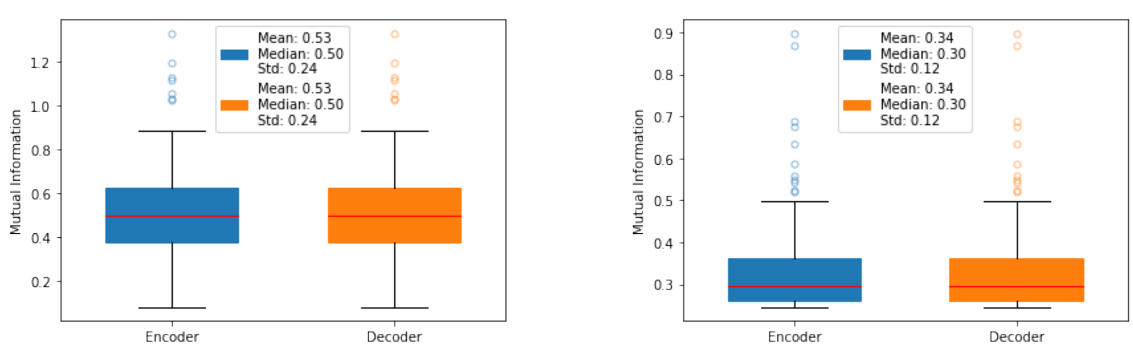

Figure 3: Plot of mutual information at $\gamma = 1$, given arbitrary $\mathbf{\Sigma_Z}$ (left) and diagonalized $\mathbf{\Sigma_Z}$ (right).

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.

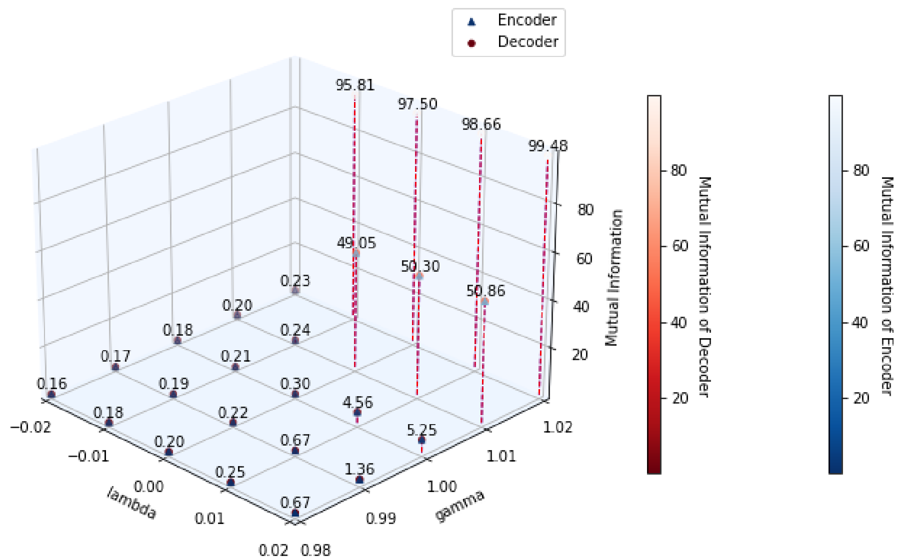

Figure 4: Plot of mutual information w.r.t. (γ, λ) pair.

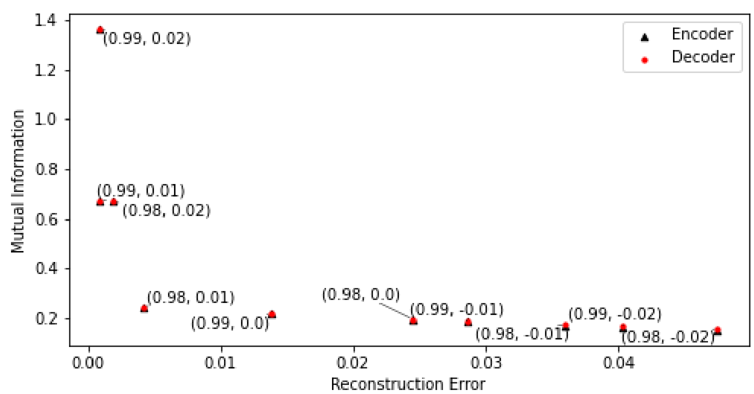

Figure 5: Plot of mutual information w.r.t. reconstruction error for each (γ, λ) pair, given $\gamma < 1$.

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

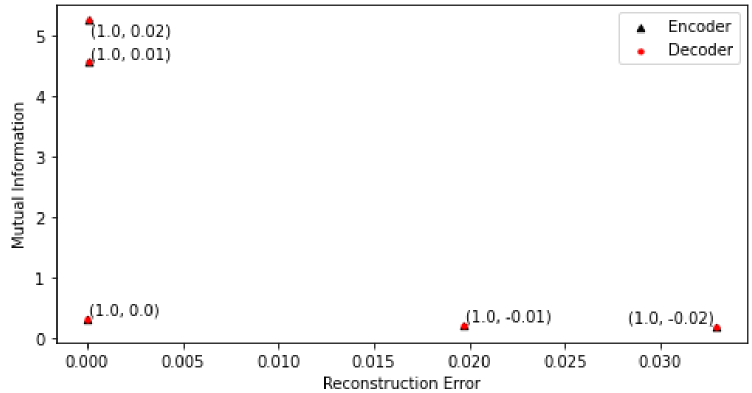

Figure 6: Plot of mutual information w.r.t. reconstruction error for each (γ, λ) pair, given $\gamma = 1$.

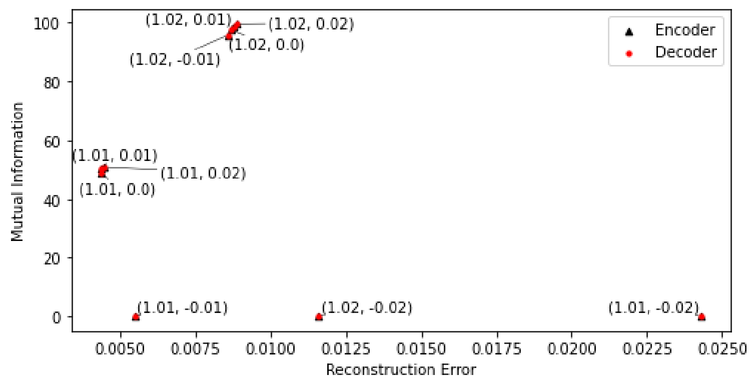

Figure 7: Plot of mutual information w.r.t. reconstruction error for each (γ, λ) pair, given $\gamma > 1$.

$\boldsymbol{\Sigma_Z}$ and diagonalized $\boldsymbol{\Sigma_Z}$, 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

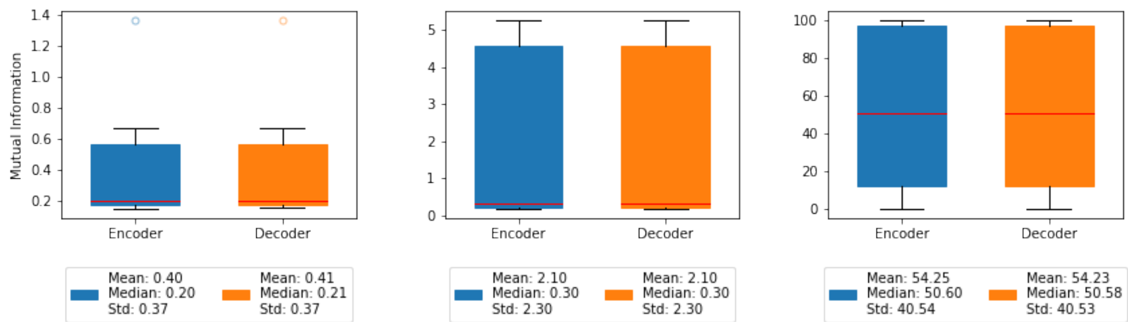

Figure 8: Plot of mutual information, given $\gamma < 1$ (left), $\gamma = 1$ (middle), $\gamma > 1$ (right).

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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$, γ-VAE with a diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$, and $\gamma\lambda$-VAE with a diagonalized positive definite $\mathbf{\Sigma}_{\boldsymbol{Z}}$.

- 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$, while numerical experiments revealed compromised uniqueness in the case of diagonalized $\mathbf{\Sigma}_{\boldsymbol{Z}}$. 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ significantly influenced numerical outcomes and reconstruction accuracy. Analytically, arbitrary $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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 $\mathbf{\Sigma}_{\boldsymbol{Z}}$ 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.

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

[11] 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*

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

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

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

□

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*

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

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

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

□

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

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

□

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*

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

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

$$\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{Y}}^{-1} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} \end{bmatrix}$$

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

□

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

$$D_{KL}[p_1\|p_2] = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} - n + \mathrm{Tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right], \tag{9}$$

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}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top & \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}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top & \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}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}})\right] \\
&= \frac{1}{2}\left[\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \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*

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

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

$$p_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right], \tag{10}$$

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

$$\begin{aligned}&\log p_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}},\boldsymbol{\theta}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})\\ &= \log\left[\mathcal{N}\left(\mathbf{0}_{(m+n)\times 1}, \begin{bmatrix}\mathbf{I}_m & \mathbf{A}^\top\\ \mathbf{A} & \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\boldsymbol{Z}}\end{bmatrix}\right)\right]\\ &= \log\left[\frac{1}{(2\pi)^{(m+n)/2}|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|^{1/2}}\exp\left(-\frac{1}{2}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}^\top\begin{bmatrix}\mathbf{I}_m & \mathbf{A}^\top\\ \mathbf{A} & \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\boldsymbol{Z}}\end{bmatrix}^{-1}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}\right)\right]\\ &= -\frac{m+n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{Z}}| - \frac{1}{2}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}^\top\begin{bmatrix}\mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\\ -\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{X}}\\ \hat{\boldsymbol{Y}}\end{bmatrix}.\end{aligned} \tag{11}$$

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}^\top\begin{bmatrix}\mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\top\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}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\top\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}^\top\right)\right]\\ &= \mathrm{Tr}\left(\begin{bmatrix}\mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\top\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}^\top\right]\right)\\ &= \mathrm{Tr}\left(\begin{bmatrix}\mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\top\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}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} & -\mathbf{A}^\top\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}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}} & \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\\ \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top & \boldsymbol{\Sigma}_{\boldsymbol{Y}}\end{bmatrix}\right)\\ &= \mathrm{Tr}[(\mathbf{I}_m + \mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A})(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}}) + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}].\end{aligned} \tag{12}$$

From equations (11) and (12), 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*

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

Proof. By equations (4) and (10), the expected log-likelihood of $p_{\boldsymbol{X}}(\boldsymbol{x})$ can be computed as

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

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

□

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

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

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

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:

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

$$
\begin{aligned}
&= -\frac{\gamma}{2}\left[\frac{\partial}{\partial \mathbf{A}}\left(\operatorname{Tr}[\mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}})]\right)\right] \\
&= -\frac{\gamma}{2}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top - 2\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}})\right] \\
&= \gamma[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}}) - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top]
\end{aligned}
$$

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

$$
\begin{aligned}
&= \frac{\partial}{\partial \mathbf{B}}\left[\frac{1}{2}\operatorname{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top) - \frac{\gamma}{2}\left(\operatorname{Tr}[\mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top]\right)\right] \\
&= \frac{1}{2}(2\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}) - \frac{\gamma}{2}\left[\mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} + \mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - 2\mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right] \\
&= \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \gamma[\mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}]
\end{aligned}
$$

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

$$
\begin{aligned}
&= -\frac{\gamma}{2}\left[\frac{\partial}{\partial \boldsymbol{\Sigma}_{\boldsymbol{Z}}}\left(\operatorname{Tr}[\mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}})] - \log|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|\right)\right] \\
&= \frac{\gamma}{2}\left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top\mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \boldsymbol{\Sigma}_{\boldsymbol{W}})\mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\right]
\end{aligned}
$$

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

$$
\begin{aligned}
&= \frac{\partial}{\partial \boldsymbol{\Sigma}_{\boldsymbol{W}}}\left[\frac{1}{2}\left[\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{W}}) - \log|\boldsymbol{\Sigma}_{\boldsymbol{W}}|\right] + \frac{\gamma}{2}\operatorname{Tr}(\mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{W}})\right] \\
&= \frac{1}{2}\left(\mathbf{I}_m - \boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\right) + \frac{\gamma}{2}\mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} \\
&= \frac{1}{2}\left[\mathbf{I}_m - \boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1} + \gamma\mathbf{A}^\top\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\right].
\end{aligned}
$$

□

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 \boldsymbol{\Gamma}_1}{\partial \mathbf{A}}(\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_{\boldsymbol{Z}},\boldsymbol{\Sigma}_{\boldsymbol{W}}) = \mathbf{0}_{n\times m} &\Leftrightarrow \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}}) = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal \\
&\Leftrightarrow \mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}}) = \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal \\
&\Leftrightarrow \mathbf{A} = \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}})^{-1} \\
\frac{\partial \boldsymbol{\Gamma}_1}{\partial \mathbf{B}}(\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_{\boldsymbol{Z}},\boldsymbol{\Sigma}_{\boldsymbol{W}}) = \mathbf{0}_{m\times n} &\Leftrightarrow \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \gamma[\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}] = \mathbf{0}_{m\times n} \\
&\Leftrightarrow \mathbf{B} - \gamma[\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} - \mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}] = \mathbf{0}_{m\times n} \\
&\Leftrightarrow (\mathbf{I}_m + \gamma\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A})\mathbf{B} = \gamma\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} \\
&\Leftrightarrow \mathbf{B} = \gamma(\mathbf{I}_m + \gamma\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A})^{-1}\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} \\
&\Leftrightarrow \mathbf{B} = (\mathbf{I}_m + \mathbf{A}^\intercal[\boldsymbol{\Sigma}_{\boldsymbol{Z}}/\gamma]^{-1}\mathbf{A})^{-1}\mathbf{A}^\intercal[\boldsymbol{\Sigma}_{\boldsymbol{Z}}/\gamma]^{-1} \\
\frac{\partial \boldsymbol{\Gamma}_1}{\partial \boldsymbol{\Sigma}_{\boldsymbol{Z}}}(\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_{\boldsymbol{Z}},\boldsymbol{\Sigma}_{\boldsymbol{W}}) = \mathbf{0}_{n\times n} &\Leftrightarrow \mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} + \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal\mathbf{A}^\intercal - \boldsymbol{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}})\mathbf{A}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{Z}} = \mathbf{0}_{n\times n} \\
&\Leftrightarrow \boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{\Sigma}_{\boldsymbol{Y}} + \mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}})\mathbf{A}^\intercal - \mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal\mathbf{A}^\intercal \\
\frac{\partial \boldsymbol{\Gamma}_1}{\partial \boldsymbol{\Sigma}_{\boldsymbol{W}}}(\mathbf{A},\mathbf{B},\boldsymbol{\Sigma}_{\boldsymbol{Z}},\boldsymbol{\Sigma}_{\boldsymbol{W}}) = \mathbf{0}_{m\times m} &\Leftrightarrow \mathbf{I}_m - \boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1} + \gamma\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} = \mathbf{0}_{m\times m} \\
&\Leftrightarrow \boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1} = \mathbf{I}_m + \gamma\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A} \\
&\Leftrightarrow \boldsymbol{\Sigma}_{\boldsymbol{W}} = (\mathbf{I}_m + \gamma\mathbf{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A})^{-1} \\
&\Leftrightarrow \boldsymbol{\Sigma}_{\boldsymbol{W}} = (\mathbf{I}_m + \mathbf{A}^\intercal[\boldsymbol{\Sigma}_{\boldsymbol{Z}}/\gamma]^{-1}\mathbf{A})^{-1}.
\end{aligned}$$

Substituting $\mathbf{A} = \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}})^{-1}$ into the expression for $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ and utilizing the formula for the inverse of a sum of matrices [3], we arrive at the following result:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{Z}} &= \boldsymbol{\Sigma}_{\boldsymbol{Y}} - \boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}})^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\
&= (\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})^{-1}.
\end{aligned}$$

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

$$\mathbf{A} = (\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})^{-1}\mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}.$$

To demonstrate this, it is sufficient to verify that

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}})^{-1} = (\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})^{-1}\mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}. \tag{13}$$

To achieve this goal, we perform both the left and right multiplication of the matrix on the left-hand side (LHS) of equation (13) by $(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})$ and $(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}})$, respectively. We apply an analogous process to the right-hand side (RHS) of equation (13). This yields:

$$\begin{aligned}
(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})(\text{LHS})(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}}) &= (\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal \\
&= \mathbf{B}^\intercal + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal \\
(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B})(\text{RHS})(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}}) &= \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{W}}) \\
&= \mathbf{B}^\intercal + \mathbf{B}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{W}}^{-1}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\intercal.
\end{aligned}$$

As the outcomes obtained from the multiplications on both the left-hand side (LHS) and the right-hand side (RHS) of equation (13) coincide, equation (13) 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 $\mathbf{\Gamma}_2(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}})$ can be computed as

$$\begin{aligned}\nabla \mathbf{\Gamma}_2(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) = \Big\langle & \frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{A}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}), \frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{B}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}), \\ & \frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{\Sigma}_{\boldsymbol{Z}}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}), \frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{\Sigma}_{\boldsymbol{W}}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) \Big\rangle,\end{aligned}$$

where

$$\begin{aligned}\frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{A}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) &= \gamma[\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \mathbf{\Sigma}_{\boldsymbol{W}}) - \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top] + 2\lambda[\mathbf{A}(\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \mathbf{\Sigma}_{\boldsymbol{W}}) - \mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top] \\ \frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{B}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) &= \mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}} - \gamma[\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}} + \mathbf{A}^\top\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}} - (\mathbf{I}_m + \mathbf{A}^\top\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A})\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}] + 2\lambda(\mathbf{A}^\top\mathbf{A}\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}} - \mathbf{A}^\top\mathbf{\Sigma}_{\boldsymbol{Y}}) \\ \frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{\Sigma}_{\boldsymbol{W}}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) &= \frac{1}{2}[\mathbf{I}_m - \mathbf{\Sigma}_{\boldsymbol{W}}^{-1} + \gamma\mathbf{A}^\top\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}] + \lambda\mathbf{A}^\top\mathbf{A}\end{aligned}$$

$$\begin{aligned}&\frac{\partial \mathbf{\Gamma}_2}{\partial \mathbf{\Sigma}_{\boldsymbol{Z}}}(\mathbf{A}, \mathbf{B}, \mathbf{\Sigma}_{\boldsymbol{Z}}, \mathbf{\Sigma}_{\boldsymbol{W}}) \\ &= \frac{\gamma}{2}\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1} + \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top\mathbf{A}^\top\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1} - \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1} - \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\mathbf{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^\top + \mathbf{\Sigma}_{\boldsymbol{W}})\mathbf{A}^\top\mathbf{\Sigma}_{\boldsymbol{Z}}^{-1} + \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}\right].\end{aligned}$$

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

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

□

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:

$$q^{(t+1)}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p^{(t)}_{\boldsymbol{X}}(\boldsymbol{x})e^{-\gamma[d(\boldsymbol{X},\boldsymbol{Y})]^{(t)}}}{\int p^{(t)}_{\boldsymbol{X}}(\boldsymbol{x})e^{-\gamma[d(\boldsymbol{X},\boldsymbol{Y})]^{(t)}}d\boldsymbol{x}}, \tag{14}$$

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 $p^{(t)}_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x}) \sim \mathcal{N}\left(\mathbf{A}^{(t)}\hat{\boldsymbol{x}}^{(t)}, \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right)$, by using equation (14), 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^{(t)}_{\boldsymbol{X}}(\boldsymbol{x}) e^{\gamma \log p^{(t)}_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x})}}{\int p^{(t)}_{\boldsymbol{X}}(\boldsymbol{x}) e^{\gamma \log p^{(t)}_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x})} d\boldsymbol{x}} \\
&= \frac{p^{(t)}_{\boldsymbol{X}}(\boldsymbol{x}) \left[p^{(t)}_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}}=\boldsymbol{y}|\hat{\boldsymbol{X}}=\boldsymbol{x}) \right]^{\gamma}}{\int p^{(t)}_{\boldsymbol{X}}(\boldsymbol{x}) \left[p^{(t)}_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\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}^{(t)}_{\boldsymbol{Z}} \right) \right]^{\gamma}}{\int \mathcal{N}(\mathbf{0}, \mathbf{I}_m) \left[\mathcal{N}\left(\mathbf{A}^{(t)} \boldsymbol{x}^{(t)}, \boldsymbol{\Sigma}^{(t)}_{\boldsymbol{Z}} \right) \right]^{\gamma} d\boldsymbol{x}} \\
&= \frac{\frac{1}{(2\pi)^{(m+\gamma n)/2} \left| \boldsymbol{\Sigma}^{(t)}_{\boldsymbol{Z}} \right|^{\gamma/2}} \exp\left(-\frac{1}{2} \left[\boldsymbol{x}^{(t)} \right]^{\top} \boldsymbol{x}^{(t)} - \frac{\gamma}{2} \left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)} \boldsymbol{x}^{(t)} \right]^{\top} \left[\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{Z}} \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}^{(t)}_{\boldsymbol{Z}} \right|^{\gamma/2}} \exp\left(-\frac{1}{2} \left[\boldsymbol{x}^{(t)} \right]^{\top} \boldsymbol{x}^{(t)} - \frac{\gamma}{2} \left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)} \boldsymbol{x}^{(t)} \right]^{\top} \left[\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{Z}} \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]^{\top} \boldsymbol{x}^{(t)} + \left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)} \boldsymbol{x}^{(t)} \right]^{\top} \left[\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{Z}} / \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]^{\top} \boldsymbol{x}^{(t)} + \left[\boldsymbol{y}^{(t)} - \mathbf{A}^{(t)} \boldsymbol{x}^{(t)} \right]^{\top} \left[\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{Z}} / \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

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

this gives

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

where

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

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

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

$$
\begin{aligned}
&= \frac{\exp\left(-\frac{1}{2}\left[\left[\boldsymbol{x}^{(t)} - \mathbf{D}^{-1}\boldsymbol{u}\right]^\top \mathbf{D}\left[\boldsymbol{x}^{(t)} - \mathbf{D}^{-1}\boldsymbol{u}\right] - \boldsymbol{u}^\top \mathbf{D}^{-1}\boldsymbol{u} + \left[\boldsymbol{y}^{(t)}\right]^\top \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]^\top \mathbf{D}\left[\boldsymbol{x}^{(t)} - \mathbf{D}^{-1}\boldsymbol{u}\right] - \boldsymbol{u}^\top \mathbf{D}^{-1}\boldsymbol{u} + \left[\boldsymbol{y}^{(t)}\right]^\top \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}^\top \mathbf{D}^{-1}\boldsymbol{u} - \frac{1}{2}\left[\boldsymbol{y}^{(t)}\right]^\top \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]^\top \mathbf{D}\left[\boldsymbol{x}^{(t)} - \mathbf{D}^{-1}\boldsymbol{u}\right]\right]\right)}{\exp\left(\frac{1}{2}\boldsymbol{u}^\top \mathbf{D}^{-1}\boldsymbol{u} - \frac{1}{2}\left[\boldsymbol{y}^{(t)}\right]^\top \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]^\top \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]^\top \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]^\top \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

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

□

B.2.2 Lemma 4

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

$$
p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x}) = \frac{p_{\boldsymbol{Y}}(\boldsymbol{y}) q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y})}{\int p_{\boldsymbol{Y}}(\boldsymbol{y}) q_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\phi}}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{y}}. \tag{15}
$$

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 (15), we can iteratively update the decoder until convergence is reached, following this procedure:

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

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

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

$$= \left[\boldsymbol{y}^{(t)}\right]^\top \left(\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^\top \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right) \boldsymbol{y}^{(t)} + 2\left(-\left[\boldsymbol{x}^{(t+1)}\right]^\top \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right) \boldsymbol{y}^{(t)}$$
$$+ \left[\boldsymbol{x}^{(t+1)}\right]^\top \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \boldsymbol{x}^{(t+1)}$$
$$= \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right]^\top \mathbf{E} \left[\boldsymbol{y}^{(t)} - \mathbf{E}^{-1}\boldsymbol{v}\right] - \boldsymbol{v}^\top \mathbf{E}^{-1}\boldsymbol{v} + \left[\boldsymbol{x}^{(t+1)}\right]^\top \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \boldsymbol{x}^{(t+1)},$$

where

$$\mathbf{E} = \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^\top \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}$$
$$\boldsymbol{v}^\top = \left[\boldsymbol{x}^{(t+1)}\right]^\top \left[\boldsymbol{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}.$$

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

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

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:

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

where

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

□

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

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

which is equivalent to

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

Since $\left[\mathbf{A}^{(t)}\right]^\top \left[\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{(t)}/\gamma\right]^{-1} \mathbf{A}^{(t)}$ 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

$$\left[\mathbf{A}^{(t)}\right]^\mathsf{T} \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} = \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)}. \tag{16}$$

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

$$\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)} = \gamma \left[\mathbf{A}^{(t)}\right]^\mathsf{T} \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1}. \tag{17}$$

From equations (16) and (17), 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 $n \geq 1$,

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

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

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

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

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

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

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

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

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

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

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

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:

$$\begin{aligned}
\left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} &= \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t)}\right]^{-1} \mathbf{B}^{(t)} \\
\left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)} &= \gamma \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} \circ \mathbf{I}_n\right).
\end{aligned} \tag{18}$$

Using system (18) 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.