# Investigating Disentanglement in $\beta$-VAE within a Linear Gaussian Setting

**Minh Vu**
(Joint work with **Shuangqing Wei** and **Xiaoliang Wan**)

January 19, 2024

# Overview

**$\beta$-VAE Model:**

- $\beta$-VAE (Higgins et al. 2017) integrates encoder and decoder components for efficient dimension reduction and data compression.
- The encoder maps input data into a lower-dimensional latent space, and the decoder reconstructs the original data.
- Aims for a balance between compression efficiency and reconstruction accuracy, crucial for applications like image or signal processing.

**Role of Disentanglement:**

- Ensures each dimension in the latent space corresponds to a specific and independent factor of variation in the data.
- Enhances understanding of the latent space dynamics, enabling precise control over individual factors without affecting others.

# Linear Gaussian Framework

$$\boldsymbol{V} \xrightarrow{\;g\;} \boldsymbol{Y} \xrightarrow[\text{Encoder}]{\;\phi\;} \boldsymbol{X} \xrightarrow[\text{Decoder}]{\;\boldsymbol{\theta}\;} \hat{\boldsymbol{Y}}$$

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

Consider a linear Gaussian setting represented by the generative model $(\{v_i\}_{i=1}^{s}, \boldsymbol{Y})$, where the input $\boldsymbol{Y} \in \mathbb{R}^n$ is defined as:

$$\begin{aligned} \boldsymbol{Y} &= \sum_{i=1}^{s} v_i \Gamma_i + \tilde{\boldsymbol{Z}} \\ &= \boldsymbol{\Gamma} \boldsymbol{V} + \tilde{\boldsymbol{Z}} \end{aligned} \tag{1}$$

- $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times s}$: A matrix formed by concatenating $s$ independent eigenvectors $\Gamma_i$, each corresponding to a standard basis vector in $\mathbb{R}^n$.
- $\tilde{\mathbf{Z}} \in \mathbb{R}^n$: The noise follows a Gaussian distribution of $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 < 1$.

# Linear Gaussian Framework

The encoding and decoding processes are defined as follows:

$$\begin{aligned} \boldsymbol{X} &= \mathbf{B}\boldsymbol{Y} + \boldsymbol{W} \\ \hat{\boldsymbol{Y}} &= \mathbf{A}\hat{\boldsymbol{X}} + \boldsymbol{Z} \end{aligned} \tag{2}$$

Here,

- $\mathbf{B} \in \mathbb{R}^{m \times n}$: The encoding matrix.
- $\boldsymbol{W} \in \mathbb{R}^{m}$: The encoder noise follows a Gaussian distribution of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ is a positive definite matrix.
- $\mathbf{A} \in \mathbb{R}^{n \times m}$: The decoding matrix.
- $\hat{\boldsymbol{X}} \in \mathbb{R}^{m}$: A sample drawn from the latent variable distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$.
- $\boldsymbol{Z} \in \mathbb{R}^{n}$: The decoder noise follows a Gaussian distribution of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$ is a diagonalized positive definite matrix.

# $\gamma\lambda$-VAE Loss Function

The loss function for the $\gamma\lambda$-VAE in a linear Gaussian setting is presented as follows:

$$\begin{aligned}
\mathcal{L}_{\gamma\lambda\text{-VAE}} &= \mathbb{E}_{\boldsymbol{Y}}[D_{KL}[q_{\boldsymbol{X}|\boldsymbol{Y},\phi}(\boldsymbol{x}|\boldsymbol{y})\|p_{\boldsymbol{X}}(\boldsymbol{x})]] \\
&\quad - \gamma\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\phi}[\log p_{\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}},\boldsymbol{\theta}}(\hat{\boldsymbol{Y}} = \boldsymbol{y}|\hat{\boldsymbol{X}} = \boldsymbol{x})] \\
&\quad + \lambda\mathbb{E}_{\boldsymbol{Y}}[\|\boldsymbol{Y} - \mathbf{A}\boldsymbol{X}\|^2] \\
&= \frac{1}{2}\Big[\mathrm{Tr}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{W}}) - \log|\boldsymbol{\Sigma}_{\boldsymbol{W}}| - m\Big] \\
&\quad - \frac{\gamma}{2}\Big(\mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}} \\
&\quad - \mathbf{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathbf{A}(\mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\mathbf{B}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\boldsymbol{W}})\Big] - n\log(2\pi) - \log|\boldsymbol{\Sigma}_{\boldsymbol{Z}}|\Big) \\
&\quad + \lambda\,\mathrm{Tr}[(\mathbf{I}_n - \mathbf{A}\mathbf{B})\boldsymbol{\Sigma}_{\boldsymbol{Y}}(\mathbf{I}_n - \mathbf{A}\mathbf{B})^{\mathsf{T}} + \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{W}}\mathbf{A}^{\mathsf{T}}].
\end{aligned} \tag{3}$$

# Problem Formulation for $(s, n, m) = (3, 4, 3)$

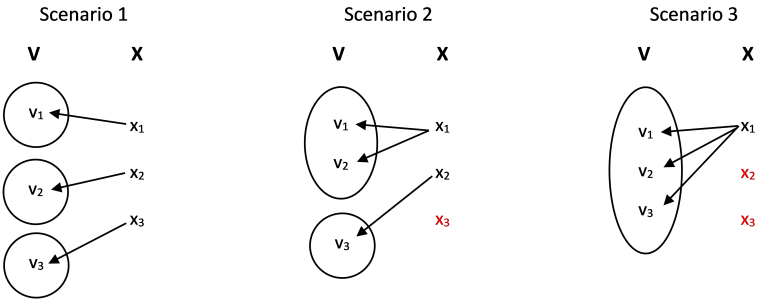


Figure: **Scenarios Illustrating Generative Factor Relationships for Disentanglement Study:** Independence, Linear Dependence of $v_1$ and $v_2$ with Independence of $v_3$, and Linear Dependence of $v_2$ and $v_3$ on $v_1$.

# Configurations of Generative Factors

1. **Independence of Generative Factors:**
   We independently sample each of the three generative factors, $v_1$, $v_2$, and $v_3$, from their respective Gaussian distributions: $\mathcal{N}(0, \sigma_{v_1}^2)$, $\mathcal{N}(0, \sigma_{v_2}^2)$, and $\mathcal{N}(0, \sigma_{v_3}^2)$.
2. **Linear Dependence of $v_1$ and $v_2$, with Independence of $v_3$:**
   We independently sample $v_1$ and $v_3$ from Gaussian distributions, specifically $\mathcal{N}(0, \sigma_{v_1}^2)$ and $\mathcal{N}(0, \sigma_{v_3}^2)$, respectively. Introducing a scaling factor, denoted as $\alpha$, we calculate $v_2$ using the equation $v_2 = \alpha v_1 + z_2$, where $z_2 \sim \mathcal{N}(0, \sigma_{z_2}^2)$.
3. **Linear Dependence of $v_2$ and $v_3$ on $v_1$:**
   We sample $v_1$ from a Gaussian distribution: $v_1 \sim \mathcal{N}(0, \sigma_{v_1}^2)$. Introducing scaling factors $\alpha$ and $\beta$, we calculate $v_2$ and $v_3$ using the equations $v_2 = \alpha v_1 + z_2$ and $v_3 = \beta v_1 + z_3$, where $z_2 \sim \mathcal{N}(0, \sigma_{z_2}^2)$ and $z_3 \sim \mathcal{N}(0, \sigma_{z_3}^2)$.

# Configurations of Generative Factors

1. **Independence of Generative Factors:**

$$\mathbf{\Sigma}_{\boldsymbol{V}} = \begin{bmatrix} \sigma_{v_1}^2 & 0 & 0 \\ 0 & \sigma_{v_2}^2 & 0 \\ 0 & 0 & \sigma_{v_3}^2 \end{bmatrix}$$

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

$$\mathbf{\Sigma}_{\boldsymbol{V}} = \begin{bmatrix} \sigma_{v_1}^2 & \alpha\sigma_{v_1}^2 & 0 \\ \alpha\sigma_{v_1}^2 & \alpha^2\sigma_{v_1}^2 + \sigma_{z_2}^2 & 0 \\ 0 & 0 & \sigma_{v_3}^2 \end{bmatrix}$$

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

$$\mathbf{\Sigma}_{\boldsymbol{V}} = \begin{bmatrix} \sigma_{v_1}^2 & \alpha\sigma_{v_1}^2 & \beta\sigma_{v_1}^2 \\ \alpha\sigma_{v_1}^2 & \alpha^2\sigma_{v_1}^2 + \sigma_{z_2}^2 & \alpha\beta\sigma_{v_1}^2 \\ \beta\sigma_{v_1}^2 & \alpha\beta\sigma_{v_1}^2 & \beta^2\sigma_{v_1}^2 + \sigma_{z_3}^2 \end{bmatrix}$$

# Disentanglement Metric $\mathcal{I}_m$

**Objective:** Introduce a disentanglement metric based on mutual information, denoted as $\mathcal{I}_m$, to evaluate disentanglement in the three specified scenarios, where $m$ is the latent variable dimension.

**Formulation of metric $\mathcal{I}_3$:**

1. For $(s, n, m) = (3, 4, 3)$, partition the set of 3 generative factors, $\mathcal{V}_3 = \{v_1, v_2, v_3\}$, into three distinct groups: $v_{s_1}$, $v_{s_2}$, and $v_{s_3}$.

$$\begin{aligned} & v_{s_1} \subset \mathcal{V}_3, v_{s_2} \subset \mathcal{V}_3, v_{s_3} \subset \mathcal{V}_3 \\ & v_{s_1} \cup v_{s_2} \cup v_{s_3} = \mathcal{V}_3 \\ & v_{s_1} \cap v_{s_2} \cap v_{s_3} = \varnothing \end{aligned} \tag{4}$$

   There are a total of 27 partitions that satisfy the conditions outlined in system (4), considering scenarios where a group of generative factors may be empty.

   - $v_{s_1} = \{v_1\}$, $v_{s_2} = \{v_2\}$, and $v_{s_3} = \{v_3\}$
   - $v_{s_1} = \{v_1\}$, $v_{s_2} = \{v_2, v_3\}$, and $v_{s_3} = \varnothing$
   - $v_{s_1} = \{v_1, v_2, v_3\}$, $v_{s_2} = \varnothing$, and $v_{s_3} = \varnothing$

# Disentanglement Metric $\mathcal{I}_3$ Formulation

2. i. Evaluate latent variables' effectiveness in capturing and representing generative factors by aggregating mutual information $I(x_i; v_{s_i})$ for $i \in \{1, 2, 3\}$.
   ii. Quantify the mutual information between two groups among the three (e.g., $I(v_{s_1}; v_{s_2})$, $I(v_{s_1}; v_{s_3})$, $I(v_{s_2}; v_{s_3})$) to address potential correlations among generative factor groups.
   iii. Subtract mutual information values between groups from the sum to formulate metric $\mathcal{I}_3$, preventing overlapping information among generative factor groups.

**Note:** If the group of generative factors $v_{s_i}$ is empty, any mutual information involving $v_{s_i}$ with another group or latent variable should be excluded from the computation.

# Disentanglement Metric $\mathcal{I}_3$ Formulation

For distinct values of $i$, $j$, and $k$ chosen from the set $\{1, 2, 3\}$, we consider 3 following cases:

- **None of the groups are empty:** $v_{s_i}, v_{s_j}, v_{s_k} \neq \varnothing$
- **One group is empty:** $v_{s_i}, v_{s_j} \neq \varnothing$ and $v_{s_k} = \varnothing$
- **Two groups are empty:** $v_{s_i} \neq \varnothing$ and $v_{s_j} = v_{s_k} = \varnothing$

So, the formula for $\mathcal{I}_3$ is defined as follows:

$$\mathcal{I}_3 = \begin{cases} \sum_{m=1}^{3} I(x_m; v_{s_m}) - \sum_{i<j} I(v_{s_i}; v_{s_j}) & \text{if Case 1} \\ I(x_i; v_{s_i}) + I(x_j; v_{s_j}) - I(v_{s_i}; v_{s_j}) & \text{if Case 2} \\ I(x_i; v_{s_i}) & \text{if Case 3} \end{cases} \tag{5}$$

# Criteria for Disentanglement using Metric $\mathcal{I}_3$

3. For each partition, we compute $\mathcal{I}_3$. After evaluating all 27 partitions, the $\mathcal{I}_3$ score is determined by selecting the highest among the 27 computed scores:

$$\mathcal{I}_3 \text{ Score} = \max\left\{\mathcal{I}_3^{(i)} \middle| 1 \leq i \leq 27\right\} \tag{6}$$

The successful disentanglement, as measured by the $\mathcal{I}_3$ metric, is achieved when the highest $\mathcal{I}_3$ score corresponds to the partition that accurately characterizes the relationships among the given generative factors.

- **Independence of Generative Factors:**
  $v_{s_1} = \{v_1\}$, $v_{s_2} = \{v_2\}$, and $v_{s_3} = \{v_3\}$
- **Linear Dependence of $v_1$ and $v_2$, with Independence of $v_3$:**
  $v_{s_1} = \{v_1, v_2\}$, $v_{s_2} = \{v_3\}$, and $v_{s_3} = \varnothing$
- **Linear Dependence of $v_2$ and $v_3$ on $v_1$:**
  $v_{s_1} = \{v_1, v_2, v_3\}$, $v_{s_2} = \varnothing$, and $v_{s_3} = \varnothing$

# SAP Score (Kumar et al. 2018)

**Calculation Steps:**

1. Construct a score matrix $\mathbf{S}$ of size $m \times s$, where $m$ represents the latent variables, and $s$ denotes the generative factors.
2. Compute each $S_{i,j}$, the $ij$-th entry of matrix $\mathbf{S}$, with the given formula:
$$S_{i,j} = \left[ \frac{\mathrm{cov}(x_i, v_j)}{\sqrt{\mathrm{var}(x_i)}\sqrt{\mathrm{var}(v_j)}} \right]^2 \quad (7)$$
3. Identify the two highest-scoring entries for each generative factor.
4. Calculate the mean difference between these top two entries across all generative factors:
$$\text{SAP score} = \frac{1}{s} \sum_{j=1}^{s} \left( S_{i^{(j)},j} - S_{i'^{(j)},j} \right) \quad (8)$$

Here, $i^{(j)} = \arg\max_i S_{i,j}$ and $i'^{(j)} = \arg\max_{i \neq i^{(j)}} S_{i,j}$.

# Numerical Simulation Configuration Across Scenarios

1. Utilize two arrays of hyperparameters: $\gamma = [0.98, 1.02]$ and $\lambda = [-0.02, 0.02]$, each incremented by 0.01. For each $(\gamma, \lambda)$ pair, seek the optimal solution $(\mathbf{A}_{\text{opt}}, \mathbf{B}_{\text{opt}}, \mathbf{\Sigma}_{\boldsymbol{Z}\text{opt}}, \mathbf{\Sigma}_{\boldsymbol{W}\text{opt}})$ for the $\gamma\lambda$-VAE loss function, adhering to a constraint of a 5% reconstruction error tolerance. Employ the Blahut-Arimoto algorithm to iteratively adjust the encoder and decoder until convergence is achieved.
   - Update the encoder $\phi^{(t+1)} = \left(\mathbf{B}^{(t+1)}, \mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right)$

$$
\begin{aligned}
\mathbf{B}^{(t+1)} &= \left[\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\gamma \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} + 2\lambda \mathbf{I}_n\right) \mathbf{A}^{(t)}\right]^{-1} \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\gamma \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} + 2\lambda \mathbf{I}_n\right) \\
\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)} &= \left[\mathbf{I}_m + \left[\mathbf{A}^{(t)}\right]^{\mathsf{T}} \left(\gamma \left[\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t)}\right]^{-1} + 2\lambda \mathbf{I}_n\right) \mathbf{A}^{(t)}\right]^{-1}
\end{aligned} \tag{9}
$$

   - Update the decoder $\theta^{(t)} = \left(\mathbf{A}^{(t+1)}, \mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)}\right)$

$$
\begin{aligned}
\mathbf{A}^{(t+1)} &= \left(\mathbf{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right)^{-1} \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \\
\mathbf{\Sigma}_{\boldsymbol{Z}}^{(t+1)} &= \left(\mathbf{\Sigma}_{\boldsymbol{Y}}^{-1} + \left[\mathbf{B}^{(t+1)}\right]^{\mathsf{T}} \left[\mathbf{\Sigma}_{\boldsymbol{W}}^{(t+1)}\right]^{-1} \mathbf{B}^{(t+1)}\right)^{-1}
\end{aligned} \tag{10}
$$

# Numerical Simulation Configuration Across Scenarios

2. Determine the covariance of the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{V}$, denoted as $\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{V}}$ as follows:

$$\boldsymbol{\Sigma}_{\boldsymbol{X},\boldsymbol{V}} = \mathbf{B}_{\text{opt}}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\boldsymbol{V}}$$

3. Compute $\mathcal{I}_3$ and SAP scores.
4. Evaluate if $\mathcal{I}_3$ and SAP successfully capture disentanglement for the given pair of $(\gamma, \lambda)$.
5. After considering all pairs of $(\gamma, \lambda)$, calculate the **disentanglement success rates** for both metrics. The success rate is determined by the number of successful disentanglements over the total 25 $(\gamma, \lambda)$ pairs, expressed as a percentage.

# Numerical Results for $(s, n, m) = (3, 4, 3)$

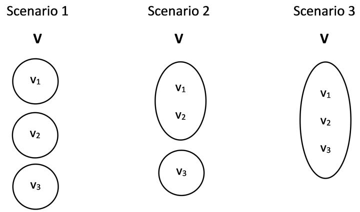


Figure: **Independent Sub-groups in Generative Factors:** 3 for Scenario 1, 2 for Scenario 2, and 1 for Scenario 3.

| Metrics | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| $\mathcal{I}_3$ | 16% | 64% | 100% |
| SAP | 40% | 92% | 92% |

Table: Disentanglement success rates for three scenarios.

# Hypothesis

## Hypothesis

*Let $G()$ denote the number of independent sub-groups hidden in the generative factors.*

i. *If $|G(V_j)| < m$, we anticipate that our metric $\mathcal{I}_m$ will outperform SAP. This expectation becomes more pronounced as the gap or difference between the two sides increases.*

ii. *Conversely, when $|G(V_j)| \geq m$, we expect a reversal in the performance order between SAP and our metric compared to the previous case.*

# Numerical Results for $(s, n, m) = (5, 5, 4)$

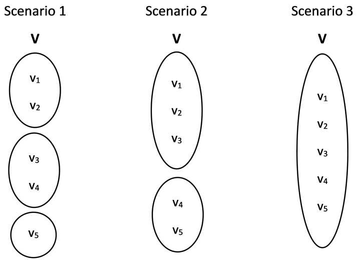


Figure: **Independent Sub-groups in Generative Factors:** 3 for Scenario 1, 2 for Scenario 2, and 1 for Scenario 3.

| Metrics | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| $\mathcal{I}_4$ | 44% | 60% | 56% |
| SAP | 52% | 20% | 4% |

Table: Disentanglement success rates for three scenarios.

# Conclusion

- If the number of independent sub-groups in the generative factors is significantly lower than the dimension of the latent space, the mutual information-based metric $\mathcal{I}_m$ is anticipated to outperform the correlation-based metric SAP.
- One limitation of metric $\mathcal{I}_m$ is its computational cost, particularly when the dimensions of generative and latent variables increase, resulting in a higher number of partitions.

# References


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

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