THE POLARISED REGIME OF IDENTIFIABLE VARIATIONAL AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

The polarised regime — the capacity of variational autoencoders (VAEs) to discard superfluous latent variables — is well-studied in the context of "classical" VAEs with a standard Gaussian prior. In this paper, we extend these results to the case of identifiable VAEs (iVAEs).

1 BACKGROUND

Variational autoencoders VAEs (Kingma & Welling, 2014; Rezende & Mohamed, 2015) are non-identifiable deep probabilistic generative models which maximise $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right)$ where $p(\mathbf{z})$ is modelled as a standard multivariate Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$.

Identifiable Variational autoencoders iVAEs (Khemakhem et al., 2020) are identifiable versions of VAEs with a conditionally factorial prior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ where $\mathbf{u}$ is observed. They maximise $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{u}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})\right)$.

The polarised regime The polarised regime is the capacity of any well-behaved VAE to discard superfluous (passive) latent variables while learning the remaining active variables with high precision (Rolinek et al., 2019; Dai & Wipf, 2018; Dai et al., 2018). When this is extended to multiple inputs, variables can be active or passive if they are active (resp. passive) over all $(\mathbf{x}^{(i)}, \mathbf{u}^{(i)})$, and mixed if they are passive for some $(\mathbf{x}^{(i)}, \mathbf{u}^{(i)})$ and active otherwise (Bonheme & Grzes, 2021).

Notational considerations The mean, variance and sampled representations of the encoder will be defined as $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\mathbf{x}, \mathbf{u}; \boldsymbol{\phi})$, $\boldsymbol{\sigma} \triangleq diag[\boldsymbol{\Sigma}(\mathbf{x}, \mathbf{u}; \boldsymbol{\phi})]$, and $\mathbf{z} \triangleq \boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}^{1/2}$, respectively, such that $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}, diag[\boldsymbol{\sigma}])$. Similarly, the mean and variance representations of the prior are defined as $\boldsymbol{\mu}^{\dagger} \triangleq \boldsymbol{\mu}^{\dagger}(\mathbf{u}; \boldsymbol{\theta})$ and $\boldsymbol{\sigma}^{\dagger} \triangleq diag[\boldsymbol{\Sigma}^{\dagger}(\mathbf{u}; \boldsymbol{\theta})]$, such that $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}^{\dagger}, diag[\boldsymbol{\sigma}^{\dagger}])$. Moreover, any representations coming from specific samples of the observations $\boldsymbol{X} \triangleq \{\mathbf{x}^{(i)}\}_{i=1}^{n}$ and $\boldsymbol{U} \triangleq \{\mathbf{u}^{(i)}\}_{i=1}^{n}$ will be indicated by the superscript $^{(i)}$. For example, $\boldsymbol{\mu}^{(i)} \triangleq \boldsymbol{\mu}(\mathbf{x}^{(i)}, \mathbf{u}^{(i)}; \boldsymbol{\phi})$.

2 EXTENSION OF THE POLARISED REGIME TO IVAES

Assumptions We consider iVAEs with a Gaussian prior and posterior, similarly to Khemakhem et al. (2020), where the mean and variance of the prior are learned (e.g., by a deep neural network). We also assume that the number of samples n is sufficiently large for sample mean and variance to be good approximations of the true mean and variance.

The proofs of this section can be found in Appendix A.

2.1 IVAES LEARN IN A POLARISED REGIME

Extending the work of Dai et al. (2018) to iVAEs, we have:

Theorem 1 (Polarised regime of iVAEs)**.** *Any well-behaved iVAE learns in a polarised regime.*

Given Theorem 1, one can then readily extend the definition of Rolinek et al. (2019) to iVAEs.

Proposition 1 (Polarised regime of $\boldsymbol{\mu}^{(i)}$, $\boldsymbol{\sigma}^{(i)}$, and $\mathbf{z}^{(i)}$)**.** *When an iVAE learns in a polarised regime, its mean, variance, and sampled representations, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\mathbf{z}$, are composed of a set of passive and active variables, $\mathbb{V}_p^{(i)} \cup \mathbb{V}_a^{(i)}$ such that, for each pair of $\mathbf{x}^{(i)}$ and $\mathbf{u}^{(i)}$:*

(i) $\boldsymbol{\mu}_j^{(i)} \approx \boldsymbol{\mu}_j^{\dagger(i)}$, $\boldsymbol{\sigma}_j^{(i)} \approx \boldsymbol{\sigma}_j^{\dagger(i)}$, *and* $\mathbf{z}_j^{(i)} \approx \mathbf{z}_j^{\dagger(i)} \quad \forall\, j \in \mathbb{V}_p^{(i)}$,

(ii) $\boldsymbol{\sigma}_j^{(i)} \ll 1$ *and* $\mathbf{z}_j^{(i)} \approx \boldsymbol{\mu}_j^{(i)} \quad \forall\, j \in \mathbb{V}_a^{(i)}$,

where $\mathbf{z}^{\dagger(i)} = \boldsymbol{\mu}^{\dagger(i)} + \boldsymbol{\epsilon}^{(i)} \left(\boldsymbol{\sigma}^{\dagger(i)}\right)^{1/2}$, $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(0, \boldsymbol{I})$ *and j is the jth variable of a representation.*

2.2 Polarised regime of iVAEs for multiple data examples

Based on Proposition 1 and Bonheme & Grzes (2021), we will now consider the properties of the mean, variance and sampled representations of iVAEs over multiple data examples.

Proposition 2 (Polarised regime of $\boldsymbol{\mu}$ over $\boldsymbol{X}$ and $\boldsymbol{U}$)**.** *When an iVAE learns in a polarised regime, its mean representation $\boldsymbol{\mu} \approx \boldsymbol{\mu}(\boldsymbol{X}, \boldsymbol{U}; \boldsymbol{\phi})$ is composed of a set of passive, active and mixed variables $\mathbb{V}_p \cup \mathbb{V}_a \cup \mathbb{V}_m$ such that, over $\boldsymbol{X}$ and $\boldsymbol{U}$:*

(i) $\bar{\boldsymbol{\mu}}_j \approx \bar{\boldsymbol{\mu}}_j^{\dagger}$ *and* $Var(\boldsymbol{\mu}_j) \approx Var(\boldsymbol{\mu}_j^{\dagger}) \quad \forall\, j \in \mathbb{V}_p$,

(ii) If the mean of the prior is fixed to some vector $\mu^{\dagger}$ and only its variance is learned, then $\bar{\boldsymbol{\mu}}_j \approx \mu_j^{\dagger}$ *and* $Var(\boldsymbol{\mu}_j) \ll 1 \quad \forall\, j \in \mathbb{V}_p$,

where $\boldsymbol{\mu}^{\dagger} \approx \boldsymbol{\mu}^{\dagger}(\boldsymbol{U}; \boldsymbol{\theta})$, $\bar{\boldsymbol{\mu}}_j \triangleq \mathbb{E}_{p(\mathbf{x},\mathbf{u})}[\boldsymbol{\mu}_j]$, $\bar{\boldsymbol{\mu}}_j^{\dagger} \triangleq \mathbb{E}_{p(\mathbf{u})}[\boldsymbol{\mu}_j^{\dagger}]$, *and $Var(\cdot)$ denotes the variance.*

Now, let us consider how the the change of prior in iVAE impacts the polarised regime of variance representations.

Proposition 3 (Polarised regime of $\boldsymbol{\sigma}$ over $\boldsymbol{X}$ and $\boldsymbol{U}$)**.** *When an iVAE learns in a polarised regime, its variance representation $\boldsymbol{\sigma} \approx diag[\boldsymbol{\Sigma}(\boldsymbol{X}, \boldsymbol{U}; \boldsymbol{\phi})]$ is composed of a set of passive, active and mixed variables $\mathbb{V}_p \cup \mathbb{V}_a \cup \mathbb{V}_m$ such that, over $\boldsymbol{X}$ and $\boldsymbol{U}$:*

(i) $\bar{\boldsymbol{\sigma}}_j \approx \bar{\boldsymbol{\sigma}}_j^{\dagger}$ *and* $Var(\boldsymbol{\sigma}_j) \approx Var(\boldsymbol{\sigma}_j^{\dagger}) \quad \forall j \in \mathbb{V}_p$,

(ii) $\bar{\boldsymbol{\sigma}}_j \ll 1$ *and* $Var(\boldsymbol{\sigma}_j) \ll 1 \quad \forall j \in \mathbb{V}_a$,

where $\boldsymbol{\sigma}^{\dagger} \approx diag[\boldsymbol{\Sigma}^{\dagger}(\boldsymbol{U}; \boldsymbol{\theta})]$, $\bar{\boldsymbol{\sigma}}_j \triangleq \mathbb{E}_{p(\mathbf{x},\mathbf{u})}[\boldsymbol{\sigma}_j]$, *and* $\bar{\boldsymbol{\sigma}}_j^{\dagger} \triangleq \mathbb{E}_{p(\mathbf{u})}[\boldsymbol{\sigma}_j^{\dagger}]$.

Using Propositions 2 and 3, we obtain the following result for sampled representations of iVAEs, which extends the corresponding proof for standard VAEs (Bonheme & Grzes, 2021).

Theorem 2 (Polarised regime of $\mathbf{z}$ over $\boldsymbol{X}$ and $\boldsymbol{U}$)**.** *When an iVAE learns in a polarised regime, its sampled representation $\mathbf{z}$ is composed of a set of passive, active and mixed variables $\mathbb{V}_p \cup \mathbb{V}_a \cup \mathbb{V}_m$ such that, over $\boldsymbol{X}$ and $\boldsymbol{U}$:*

(i) $p(\mathbf{z}_j) \approx p(\mathbf{z}_j^{\dagger}) \quad \forall\, j \in \mathbb{V}_p$,

(ii) $p(\mathbf{z}_j) \approx p(\boldsymbol{\mu}_j) \quad \forall\, j \in \mathbb{V}_a$,

(iii) $p(\mathbf{z}_j) = c\, p(\mathbf{z}_j^{\dagger}) + (1 - c)\, p(\boldsymbol{\mu}_j) \quad \forall\, j \in \mathbb{V}_m$,

where $0 < c < 1$.

3 Conclusion

We have shown that iVAEs, whose priors are learned, behave in a polarised regime and display similar active variables as standard VAEs, whose priors are fixed. Moreover, the properties of passive variables of standard VAEs are a specific case of those of iVAEs where the mean and variance of the prior are fixed to 0 and $\boldsymbol{I}$.

URM STATEMENT

"The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track."

REFERENCES

Lisa Bonheme and Marek Grzes. Be More Active! Understanding the Differences between Mean and Sampled Representations of Variational Autoencoders. *arXiv e-prints*, 2021.

Bin Dai and David Wipf. Diagnosing and Enhancing VAE Models. In *International Conference on Learning Representations*, volume 6, 2018.

Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with Robust PCA and the Role of Emergent Sparsity in Variational Autoencoder Models. *Journal of Machine Learning Research*, 19(41):1–42, 2018.

Gene H. Golub and Charles F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Fourth edition, 2013. ISBN 9781421407944.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2020.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, volume 2, 2014.

Danilo Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 2015.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

A PROOFS

A.1 PROOF OF THEOREM 1

We want to show that a well-behaved iVAE — an iVAE converging to the lowest reconstruction error and KL divergence possible — learns in a polarised regime. That is, any superfluous (passive) latent dimension is discarded and relevant (active) variables are learned with high precision. Specifically, passive variables will only depend on the prior while active variables will have very low variance. Based on Dai & Wipf (2018, Theorem 5), we will show that iVAEs similarly behave in a polarised regime.

Proof. Let us consider the following multivariate Gaussian distributions with diagonal covariance

$$q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}^{\dagger}, \boldsymbol{\Sigma}^{\dagger}), \tag{2}$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}^{\ddagger}, \gamma \boldsymbol{I}), \tag{3}$$

where $\boldsymbol{\mu}^{\ddagger} \stackrel{\text{def}}{=} \boldsymbol{\mu}^{\ddagger}[\mathbf{z};\theta]$ and $\gamma > 0$. Given k latent dimensions, the learning objective to minimise is

$$-\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{u}) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] + D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})\right). \tag{4}$$

Plugging Equation 3 into Equation 4, and dropping the dependencies on model parameters for readability we get

$$-\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{u}) = \frac{1}{2\gamma}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[\|\mathbf{x} - \boldsymbol{\mu}^{\ddagger}[\mathbf{z}]\|_2^2\right] + \frac{k}{2}\log(2\pi\gamma) + D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})\right) \tag{5}$$

$$\geqslant \frac{1}{2\gamma}\mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\|\mathbf{x} - \boldsymbol{\mu}^{\ddagger}[\boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}^{1/2}]\|_2^2\right] + \frac{k}{2}\log(2\pi\gamma). \tag{6}$$

Now suppose that the reconstruction is highly precise (i.e., $\gamma \to 0$), and we have

$$\lim_{\gamma\to 0}\mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\|\mathbf{x} - \boldsymbol{\mu}^{\ddagger}[\boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}^{1/2}]\|_2^2\right] = \Delta.$$

If $\Delta \neq 0$, we have $\lim_{\gamma\to 0} -\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) \geqslant \lim_{\gamma\to 0}\frac{\Delta}{2\gamma} + \frac{k}{2}\log(2\pi\gamma) = +\infty$ which contradicts the fact that $-\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x})$ converges to $-\infty$ (i.e., the model is not well-behaved). Thus we must have $\Delta = 0$ and, given that $\|\mathbf{x} - \boldsymbol{\mu}^{\ddagger}[\boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}^{1/2}]\|_2^2 \geqslant 0$, it means that

$$\lim_{\gamma\to 0}\boldsymbol{\mu}^{\ddagger}[\boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}^{1/2}] = \mathbf{x}.$$

Furthermore, if $\boldsymbol{\epsilon} = 0$, this becomes

$$\lim_{\gamma\to 0}\boldsymbol{\mu}^{\ddagger}[\boldsymbol{\mu}] = \mathbf{x}. \tag{7}$$

We will now see that this is achieved by setting $\boldsymbol{\sigma}$ to very low values on active variables.

Recalling Equation 7, let us derive the Taylor approximation of $\boldsymbol{\mu}^{\ddagger}[\mathbf{z}]$ at $\mathbf{z} = \boldsymbol{\mu}$

$$\boldsymbol{\mu}^{\ddagger}[\mathbf{z}] \approx \boldsymbol{\mu}^{\ddagger}[\boldsymbol{\mu}] + \boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}](\mathbf{z} - \boldsymbol{\mu}) \approx \mathbf{x} + \boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}](\mathbf{z} - \boldsymbol{\mu}). \tag{8}$$

Plugging Equation 8 into Equation 5, and letting $C \stackrel{\text{def}}{=} \frac{k}{2}\log(2\pi\gamma) + D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})\right)$, we get

$$\begin{aligned}
-\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{u}) &\approx \frac{1}{2\gamma}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[\|\mathbf{x} - \mathbf{x} - \boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}](\mathbf{z} - \boldsymbol{\mu})\|_2^2\right] + C, \\
&= \frac{1}{2\gamma}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[\left(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}](\mathbf{z} - \boldsymbol{\mu})\right)^T \boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}](\mathbf{z} - \boldsymbol{\mu})\right] + C, \\
&= \frac{1}{2\gamma}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[Tr\left((\mathbf{z} - \boldsymbol{\mu})^T(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}](\mathbf{z} - \boldsymbol{\mu})\right)\right] + C, \\
&= \frac{1}{2\gamma}Tr\left(\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[(\mathbf{z} - \boldsymbol{\mu})^T(\mathbf{z} - \boldsymbol{\mu})](\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]\right) + C, \\
&= \frac{1}{2\gamma}Tr\left(\boldsymbol{\Sigma}(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]\right) + C.
\end{aligned} \tag{9}$$

Plugging Equations 1 and 2 into the KL divergence term of Equation 5, we get

$$\begin{aligned}
-\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}, \mathbf{u}) &= \frac{1}{2\gamma}Tr\left(\boldsymbol{\Sigma}(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]\right) + \frac{1}{2}Tr(\boldsymbol{\Sigma}^{\dagger^{-1}}\boldsymbol{\Sigma}) + \frac{k}{2}\log(2\pi\gamma) + \\
&\quad \frac{1}{2}\left((\boldsymbol{\mu}^{\dagger} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{\dagger^{-1}}(\boldsymbol{\mu}^{\dagger} - \boldsymbol{\mu}) - k + \log|\boldsymbol{\Sigma}^{\dagger}| - \log|\boldsymbol{\Sigma}|\right), \\
&= \frac{1}{2}Tr\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^{\dagger^{-1}} + \frac{1}{\gamma}(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]\right)\right) + \frac{k}{2}\log(2\pi\gamma) + \\
&\quad \frac{1}{2}\left((\boldsymbol{\mu}^{\dagger} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{\dagger^{-1}}(\boldsymbol{\mu}^{\dagger} - \boldsymbol{\mu}) - k + \log|\boldsymbol{\Sigma}^{\dagger}| - \log|\boldsymbol{\Sigma}|\right).
\end{aligned} \tag{10}$$

$$\tag{11}$$

The optimal value of $\boldsymbol{\Sigma}$ must thus satisfy

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}^{\dagger^{-1}} + \frac{1}{\gamma}(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]\right)^{-1}. \tag{12}$$

As $\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]$ is a tangent space of the r-dimensional manifold $\mathcal{X}$ at $\boldsymbol{\mu}^{\ddagger}[\boldsymbol{\mu}]$, it has a rank of r. It follows that $(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]$ can be decomposed as $\boldsymbol{U}^{\ddagger}\boldsymbol{S}^{\ddagger}\boldsymbol{S}^{\ddagger^T}\boldsymbol{U}^{\ddagger^T}$ where the first r elements of the diagonal matrix $\boldsymbol{S}^{\ddagger}\boldsymbol{S}^{\ddagger^T} \in \mathbb{R}^{k\times k}$ are nonzero, such that $diag[\boldsymbol{S}^{\ddagger}\boldsymbol{S}^{\ddagger^T}] = [\lambda_1^{\ddagger}, \lambda_2^{\ddagger}, \cdots, \lambda_r^{\ddagger}, 0, \cdots, 0]$. If $r = k$, given that $\boldsymbol{\Sigma}$ and $(\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}])^T\boldsymbol{\mu}^{\ddagger\prime}[\boldsymbol{\mu}]$ are k-by-k symmetric matrices, by Golub & Van Loan (2013, Theorem 8.1.5), we have

$$\frac{1}{\frac{1}{\lambda^{\dagger}_{min}} + \frac{\lambda_i^{\ddagger}}{\gamma}} \leqslant \lambda_i \leqslant \frac{1}{\frac{1}{\lambda^{\dagger}_{max}} + \frac{\lambda_i^{\ddagger}}{\gamma}} \qquad \forall i = 1, \cdots, k. \tag{13}$$

We can directly see that as $\gamma \to 0$, both sides of the inequality converge to 0, thus the eigenvalues of $\boldsymbol{\Sigma}$ become arbitrarily small at a rate proportional to γ. Thus similarly to Dai & Wipf (2018), $\frac{1}{\sqrt{\gamma}}\boldsymbol{\Sigma}^{1/2} \approx O(1)$ under mild conditions and around the optimal solution, we have

$$-2\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] = 2\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[\frac{1}{\gamma}\|\mathbf{x} - \boldsymbol{\mu}^{\ddagger}[\mathbf{z}]\|_2^2\right] + k\log(2\pi\gamma), \tag{14}$$

$$\approx \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[O(1)\right] + k\log(2\pi\gamma), \tag{15}$$

$$= O(1) + k\log\gamma. \tag{16}$$

Moreover, because $2\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] \geqslant 0$, one can see that the lower bound provided by Equation 16 cannot be further decreased. Thus, any additional variables introduced by increasing k will not improve the reconstruction but may have a negative impact on the KL divergence when $\gamma \to 0$. Thus, similarly to VAEs, any superfluous variable of iVAEs will seek to lower the cost of the KL divergence by remaining close to the prior. □

A.2 Proof of Proposition 1

Proof. Let us consider the classical case where the prior is a standard Gaussian distribution. Following Rolinek et al. (2019); Bonheme & Grzes (2021), we have

Definition 1 (Polarised regime of VAEs)**.** *When a VAE learns in a polarised regime, its mean, variance, and sampled representations, $\boldsymbol{\mu}^{(i)}$, $\boldsymbol{\sigma}^{(i)}$, and $\mathbf{z}^{(i)}$, are composed of a set of passive and active variables, $\mathbb{V}_p^{(i)} \cup \mathbb{V}_a^{(i)}$ such that, for each data example $\mathbf{x}^{(i)}$:*

(i) $\boldsymbol{\mu}_j^{(i)} \approx 0$, $\boldsymbol{\sigma}_j^{(i)} \approx 1$, *and* $\mathbf{z}_j^{(i)} \approx \boldsymbol{\epsilon}_j^{(i)} \quad \forall\, j \in \mathbb{V}_p^{(i)}$,

(ii) $\boldsymbol{\sigma}_j^{(i)} \ll 1$ *and* $\mathbf{z}_j^{(i)} \approx \boldsymbol{\mu}_j^{(i)} \quad \forall\, j \in \mathbb{V}_a^{(i)}$,

where $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and j is the jth variable of a representation.

Statement (i) comes from the fact that the prior is $\mathcal{N}(0, \boldsymbol{I})$ and passive variables are as close as possible of the prior to decrease the KL divergence, while statement (ii) shows that active variables have high precision (i.e., low variance) while increasing the KL divergence. For iVAEs, statement (ii) remains unchanged but statement (i) needs to be updated to take into account the new prior distribution $\mathcal{N}(\boldsymbol{\mu}^{\dagger}, diag[\boldsymbol{\sigma}^{\dagger}])$.

Because the prior is Gaussian, $\mathbf{z}^{\dagger} = \boldsymbol{\mu}^{\dagger} + \boldsymbol{\epsilon}\left(\boldsymbol{\sigma}^{\dagger}\right)^{1/2}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ will be distributed according to $\mathcal{N}(\boldsymbol{\mu}^{\dagger}, diag[\boldsymbol{\sigma}^{\dagger}])$. Thus, for passive variables to be as close as possible to the prior, we need $\boldsymbol{\mu}_j^{(i)} \approx \boldsymbol{\mu}_j^{\dagger(i)}$ and $\boldsymbol{\sigma}_j^{(i)} \approx \boldsymbol{\sigma}_j^{\dagger(i)}$ for all $j \in \mathbb{V}_p^{(i)}$. We can thus generalise statement (i) to $\boldsymbol{\mu}_j^{(i)} \approx \boldsymbol{\mu}_j^{\dagger}$, $\boldsymbol{\sigma}_j^{(i)} \approx \boldsymbol{\sigma}_j^{\dagger}$, and $\mathbf{z}_j^{(i)} \approx \mathbf{z}_j^{\dagger(i)} \quad \forall\, j \in \mathbb{V}_p^{(i)}$, as required. □

A.3 Proof of Proposition 2

Proof. Let us first consider statement (i) of Proposition 2, where the mean of the prior is learned. We know from Proposition 1 that for all $j \in \mathbb{V}_p^{(i)}$, $\boldsymbol{\mu}_j^{(i)} \approx \boldsymbol{\mu}_j^{\dagger(i)}$. Thus, $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mu}_j^{(i)} \approx \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mu}_j^{\dagger(i)}$ and

$\bar{\boldsymbol{\mu}}_j \approx \bar{\boldsymbol{\mu}}_j^{\dagger}$. Similarly $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\mu}_j^{(i)} - \bar{\boldsymbol{\mu}}_j^{(i)})^2 \approx \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\mu}_j^{\dagger(i)} - \bar{\boldsymbol{\mu}}_j^{\dagger(i)})^2$ and $Var(\boldsymbol{\mu}_j) \approx Var(\boldsymbol{\mu}_j^{\dagger})$, as required.

In statement (ii), the prior has a fixed mean $\mu^{\dagger}$. Thus, $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mu}_j^{(i)} \approx \frac{1}{n}\sum_{i=1}^{n}\mu_j^{\dagger} = \mu_j^{\dagger}$ and $\bar{\boldsymbol{\mu}}_j \approx \mu_j^{\dagger}$. Similarly $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\mu}_j^{(i)} - \bar{\boldsymbol{\mu}}_j^{(i)})^2 \approx \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\mu}_j^{\dagger} - \boldsymbol{\mu}_j^{\dagger})^2 \ll 1$ and $Var(\boldsymbol{\mu}_j) \ll 1$, as required. □

A.4 Proof of Proposition 3

Proof. Let us first consider statement (i) of Proposition 3 which concerned the passive variables of the variance representation. We know from Proposition 1 that for all $j \in \mathbb{V}_p^{(i)}$, $\boldsymbol{\sigma}_j^{(i)} \approx \boldsymbol{\sigma}_j^{\dagger(i)}$. Thus, $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_j^{(i)} \approx \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_j^{\dagger(i)}$ and $\bar{\boldsymbol{\sigma}}_j \approx \bar{\boldsymbol{\sigma}}_j^{\dagger}$. Similarly $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\sigma}_j^{(i)} - \bar{\boldsymbol{\sigma}}_j^{(i)})^2 \approx \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\sigma}_j^{\dagger(i)} - \bar{\boldsymbol{\sigma}}_j^{\dagger(i)})^2$ and $Var(\boldsymbol{\sigma}_j) \approx Var(\boldsymbol{\sigma}_j^{\dagger})$, as required.

Statement (ii) of Proposition 3 is concerned with the active variables of the variance representation. We know from Proposition 1 that for all $j \in \mathbb{V}_a^{(i)}$, $\boldsymbol{\sigma}_j^{(i)} \ll 1$. Thus, $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\sigma}_j^{(i)} \ll 1$ and $\bar{\boldsymbol{\sigma}}_j \ll 1$. Similarly $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\sigma}_j^{(i)} - \bar{\boldsymbol{\sigma}}_j^{(i)})^2 \ll 1$ and $Var(\boldsymbol{\sigma}_j) \ll 1$, as required.

□

A.5 Proof of Theorem 2

Proof. Let $\mathbf{z}_j$ be the sampled representation variable at index j. There are three cases:

(i) If $j \in \mathbb{V}_p$, then, from statement (i) of Proposition 2, $\boldsymbol{\mu}_j \approx \boldsymbol{\mu}_j^{\dagger}$. Moreover, from statement (i) of Proposition 2, $\boldsymbol{\sigma}_j \approx \boldsymbol{\sigma}_j^{\dagger}$. Thus, $\mathbf{z}_j \approx \boldsymbol{\mu}_j^{\dagger} + \boldsymbol{\epsilon}_j \left(\boldsymbol{\sigma}_j^{\dagger}\right)^{1/2}$. Recall from Proposition 1 that $\mathbf{z}_j^{\dagger}$ is distributed according to $N(\boldsymbol{\mu}^{\dagger}, diag[\boldsymbol{\sigma}^{\dagger}])$. It follows that $p(\mathbf{z}_j) \approx p(\mathbf{z}_j^{\dagger})$, which proves statement (i).

(ii) If $j \in \mathbb{V}_a$, then, from statement (ii) of Proposition 3, $\boldsymbol{\sigma}_j$ is almost constant with a value close to 0. Thus, $\mathbf{z}_j \approx \boldsymbol{\mu}_j$. It follows that $p(\mathbf{z}_j) \approx p(\boldsymbol{\mu}_j)$, which proves statement (ii).

(iii) If $j \in \mathbb{V}_m$, we know that $\mathbf{z}_j$ is composed of a subset of active components and a subset of passive components. Thus, $\mathbf{z}_j$ is distributed according to a mixture distribution. Using step (i) and (ii) of the proof, we know that $p(\mathbf{z}_j) \approx p(\mathbf{z}_j^{\dagger})$ for passive variables and $p(\mathbf{z}_j) \approx p(\boldsymbol{\mu}_j)$ for active variables. It follows that for mixed variables $p(\mathbf{z}_j) = c\ p(\mathbf{z}_j^{\dagger}) + (1 - c)\ p(\boldsymbol{\mu}_j)$ where $0 < c < 1$. This concludes the proof.

□