Marginalization is not Marginal: No Bad VAE Local Minima when Learning Optimal Sparse Representations

David Wipf
Amazon Web Services

ICML, July 2023

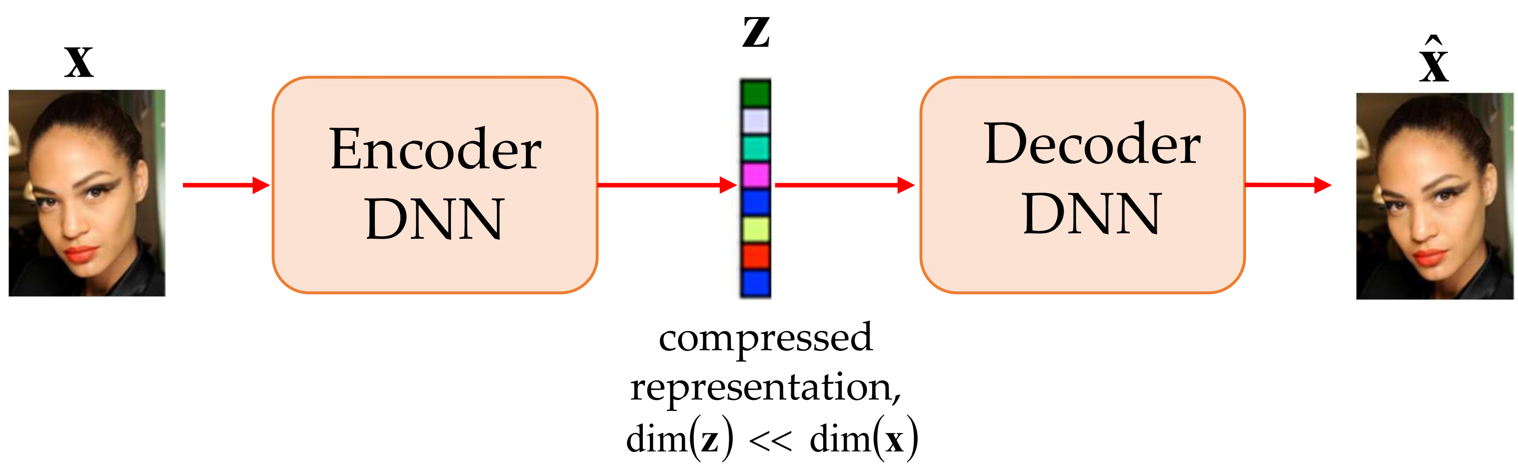
Variational Autoencoders (VAE)

- Popular generative model, probabilistic AE extension
- 33573+ citations [Kingma & Welling, 2014; Rezende et al., 2014]

Variational Autoencoders (VAE)

- Popular generative model, probabilistic AE extension
- 33573+ citations [Kingma & Welling, 2014; Rezende et al., 2014]

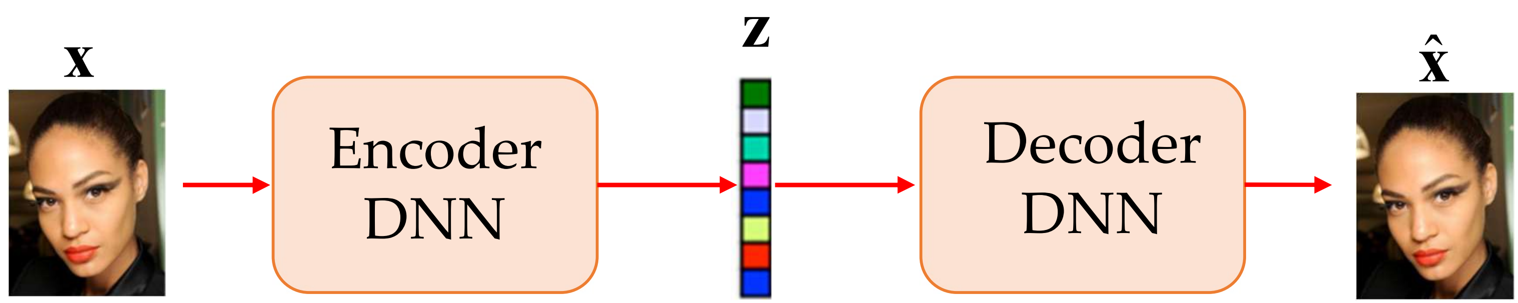
Autoencoder (AE):



Variational Autoencoders (VAE)

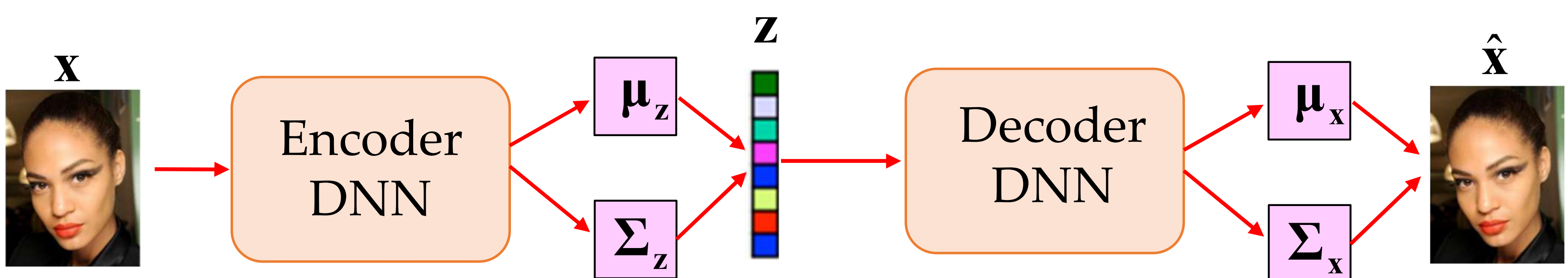
- Popular generative model, probabilistic AE extension
- 33573+ citations [Kingma & Welling, 2014; Rezende et al., 2014]

Autoencoder (AE):



compressed representation, $\dim(\mathbf{z}) \ll \dim(\mathbf{x})$

VAE (continuous data):



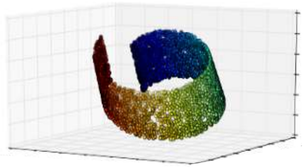
stochastic representation

$$\mathbf{z} \sim q_\varphi(\mathbf{z} \mid \mathbf{x}) = N(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

$$\hat{\mathbf{x}} \sim p_\theta(\mathbf{x} \mid \mathbf{z}) = N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$$

Why Variational Autoencoders?

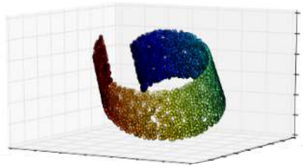
- Beyond generating samples, VAE global minima can be used to:
  - Compute the dimension of data manifolds.
  - Model/Exploit low-dimensional structure in data.



[Zheng et al., 2022]

Why Variational Autoencoders?

- Beyond generating samples, VAE global minima can be used to:
  - Compute the dimension of data manifolds.
  - Model/Exploit low-dimensional structure in data.



[Zheng et al., 2022]

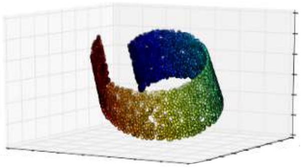
- Simple representative example:

With linear decoder, VAE minima learn PCA dimensions.

[Dai et al., 2019; Lucas et al., 2019]

Why Variational Autoencoders?

- Beyond generating samples, VAE global minima can be used to:
  - Compute the dimension of data manifolds.
  - Model/Exploit low-dimensional structure in data.



[Zheng et al., 2022]

- Simple representative example:

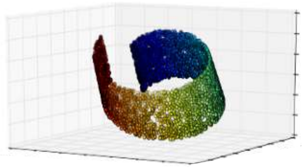
With linear decoder, VAE minima learn PCA dimensions.

[Dai et al., 2019; Lucas et al., 2019]

- More broadly:
  - VAE global minima generalize many classical linear methods.
  - Useful for feature learning, interpretability, solving underdetermined inverse problems … [Dai et al., 2021]

Why Variational Autoencoders?

- Beyond generating samples, VAE global minima can be used to:
  - Compute the dimension of data manifolds.
  - Model/Exploit low-dimensional structure in data.



[Zheng et al., 2022]

- Simple representative example:

With linear decoder, VAE minima learn PCA dimensions.

[Dai et al., 2019; Lucas et al., 2019]

- More broadly:
  - VAE global minima generalize many classical linear methods.
  - Useful for feature learning, interpretability, solving underdetermined inverse problems … [Dai et al., 2021]

- These VAE capabilities can be formalized through the notion of optimal sparse representations.

Optimal Sparse Representations

- **Two VAE criteria**

1) Optimal reconstruction: $$\sum_i \left\{ \mathrm{E}_{q_\varphi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_x\left[\mathbf{z},\boldsymbol{\theta}\right] \right\|_2^2 \right] \right\} = 0$$

2) Maximal # of latent dimensions set to prior: $$q_\varphi\left(z_j \mid \mathbf{x}^{(i)}\right) = N\left(0,1\right), \quad \forall i$$ no benefit to reconstructions

[Dai et al., 2021]

Optimal Sparse Representations

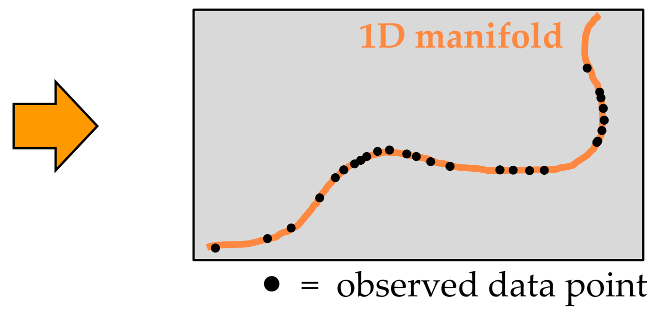
- **Two VAE criteria**

1) Optimal reconstruction: $\sum_i \left\{ \mathrm{E}_{q_\varphi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right] \right\|_2^2 \right] \right\} = 0$

2) Maximal # of latent dimensions set to prior: $q_\varphi\left(z_j \mid \mathbf{x}^{(i)}\right) = N\left(0,1\right), \quad \forall i$ no benefit to reconstructions

[Dai et al., 2021]

- Note: Second criteria determines data manifold dimension.



Optimal Sparse Representations

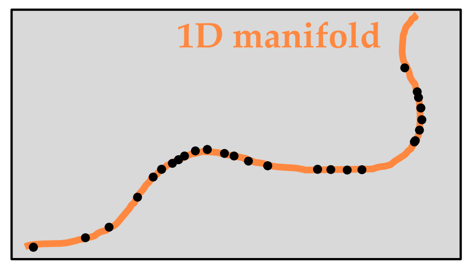
- **Two VAE criteria**

1) Optimal reconstruction: $$\sum_i \left\{ \mathrm{E}_{q_\varphi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\left\|\mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right]\right\|_2^2 \right] \right\} = 0$$

2) Maximal # of latent dimensions set to prior: $$q_\varphi\left(z_j \mid \mathbf{x}^{(i)}\right) = N\left(0,1\right), \quad \forall i$$ no benefit to reconstructions

[Dai et al., 2021]

- Note: Second criteria determines data manifold dimension.



● = observed data point

- Under broad conditions, VAE global minima provably satisfy both criteria.

[Zheng et al., 2022]

Optimal Sparse Representations

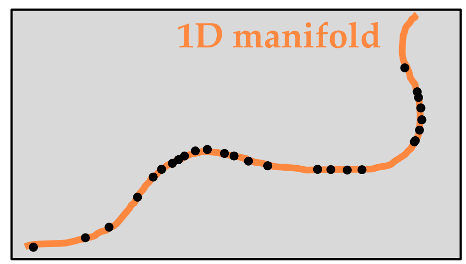
- **Two VAE criteria**

1) Optimal reconstruction: $$\sum_i \left\{ \mathrm{E}_{q_\varphi\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\left\|\mathbf{x}^{(i)} - \boldsymbol{\mu}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right]\right\|_2^2 \right] \right\} = 0$$

2) Maximal # of latent dimensions set to prior: $$q_\varphi\left(z_j \mid \mathbf{x}^{(i)}\right) = N(0,1), \quad \forall i$$ no benefit to reconstructions

[Dai et al., 2021]

- Note: Second criteria determines data manifold dimension.



● = observed data point

- Under broad conditions, VAE global minima provably satisfy both criteria.

[Zheng et al., 2022]

- Open question: What about bad VAE local minima?

Practical Example:

Multiple-Response Sparse Regression

- Observed data: $\mathbf{X} = \left\{ \mathbf{x}^{(i)} \right\}_{i=1}^{n} \in \mathbb{R}^{d \times n}$

Practical Example:
Multiple-Response Sparse Regression

- Observed data: $\mathbf{X} = \left\{ \mathbf{x}^{(i)} \right\}_{i=1}^{n} \in \mathbb{R}^{d \times n}$

- Assumed ground-truth model:

$$\mathbf{X} = \mathbf{\Phi} \mathbf{U}_0, \quad \underbrace{\mathbf{\Phi} \in \mathbb{R}^{d \times \kappa}}_{\text{known feature dictionary}}, \quad \underbrace{\mathbf{U}_0 \in \mathbb{R}^{\kappa \times n}}_{\text{unknown row-sparse coefficients}}$$

⇒ optimal sparse representation

Practical Example:
Multiple-Response Sparse Regression

- Observed data: $\mathbf{X} = \left\{ \mathbf{x}^{(i)} \right\}_{i=1}^{n} \in \mathbb{R}^{d \times n}$

- Assumed ground-truth model:

$$\mathbf{X} = \mathbf{\Phi} \mathbf{U}_0, \quad \underbrace{\mathbf{\Phi} \in \mathbb{R}^{d \times \kappa}}_{\text{known feature dictionary}}, \quad \underbrace{\mathbf{U}_0 \in \mathbb{R}^{\kappa \times n}}_{\text{unknown row-sparse coefficients}}$$

⇒ optimal sparse representation

- NP-hard inverse problem:

$$\mathbf{U}_0 = \arg\min_{\mathbf{U}} \ \rho_0(\mathbf{U}), \quad \text{s.t. } \mathbf{X} = \mathbf{\Phi}\mathbf{U}, \quad \rho_0(\mathbf{U}) \triangleq \sum_{j=1}^{\kappa} \mathbf{1}\left[\left\| \mathbf{u}_{(j)} \right\|_2 \neq 0 \right]$$

} counts # of nonzero rows

[Cotter et al., 2005; Lee et al., 2012; Adcock et al., 2019]

Practical Example:

Multiple-Response Sparse Regression

- Observed data: $\mathbf{X} = \left\{ \mathbf{x}^{(i)} \right\}_{i=1}^{n} \in \mathbb{R}^{d \times n}$

- Assumed ground-truth model:

$$\mathbf{X} = \mathbf{\Phi} \mathbf{U}_0, \quad \underbrace{\mathbf{\Phi} \in \mathbb{R}^{d \times \kappa}}_{\text{known feature dictionary}}, \quad \underbrace{\mathbf{U}_0 \in \mathbb{R}^{\kappa \times n}}_{\text{unknown row-sparse coefficients}}$$

⇒ optimal sparse representation

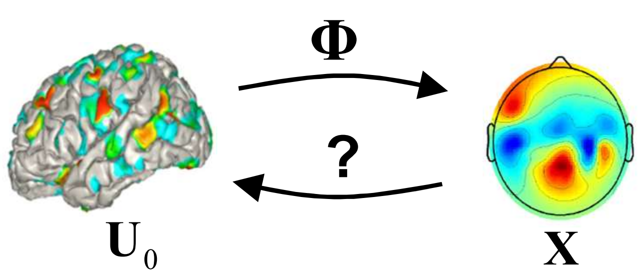
- NP-hard inverse problem:

$$\mathbf{U}_0 = \arg\min_{\mathbf{U}} \ \rho_0(\mathbf{U}), \quad \text{s.t. } \mathbf{X} = \mathbf{\Phi} \mathbf{U}, \quad \rho_0(\mathbf{U}) \triangleq \sum_{j=1}^{\kappa} \mathbf{1}\left[\left\| \mathbf{u}_{(j)} \right\|_2 \neq 0 \right]$$

} counts # of nonzero rows

[Cotter et al., 2005; Lee et al., 2012; Adcock et al., 2019]

- Applications:



[Bannier et al., 2021; Bhutada et al., 2022; Cai et al., 2018]

Corresponding VAE Design

- VAE components for multiple sparse regression:

Decoder: $\boldsymbol{\mu}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right] = \boldsymbol{\Phi}\,\text{diag}\left[\mathbf{w}_{\mathbf{x}}\right]\mathbf{z}, \quad \boldsymbol{\Sigma}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right] = \lambda\mathbf{I}, \qquad \boldsymbol{\theta} = \left\{\mathbf{w}_{\mathbf{x}}, \lambda\right\}$

Corresponding VAE Design

- VAE components for multiple sparse regression:

Decoder: $\boldsymbol{\mu}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right] = \boldsymbol{\Phi}\,\mathrm{diag}\left[\mathbf{w}_{\mathbf{x}}\right]\mathbf{z}, \quad \boldsymbol{\Sigma}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right] = \lambda\mathbf{I}, \quad \boldsymbol{\theta} = \left\{\mathbf{w}_{\mathbf{x}}, \lambda\right\}$

Encoder: $\boldsymbol{\mu}_{\mathbf{z}}\left[\mathbf{x},\boldsymbol{\varphi}\right] = \mathbf{W}_{\mathbf{z}}\mathbf{x}, \quad \boldsymbol{\Sigma}_{\mathbf{z}}\left[\mathbf{x},\boldsymbol{\varphi}\right] = \mathbf{S}\mathbf{S}^{T}, \quad \boldsymbol{\varphi} = \left\{\mathbf{W}_{\mathbf{z}}, \mathbf{S}\right\}$

Corresponding VAE Design

- VAE components for multiple sparse regression:

Decoder: $\boldsymbol{\mu}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right] = \boldsymbol{\Phi}\,\mathrm{diag}\left[\mathbf{w}_{\mathbf{x}}\right]\mathbf{z}, \quad \boldsymbol{\Sigma}_{\mathbf{x}}\left[\mathbf{z},\boldsymbol{\theta}\right] = \lambda\mathbf{I}, \quad \boldsymbol{\theta} = \left\{\mathbf{w}_{\mathbf{x}},\lambda\right\}$

Encoder: $\boldsymbol{\mu}_{\mathbf{z}}\left[\mathbf{x},\boldsymbol{\varphi}\right] = \mathbf{W}_{\mathbf{z}}\mathbf{x}, \quad \boldsymbol{\Sigma}_{\mathbf{z}}\left[\mathbf{x},\boldsymbol{\varphi}\right] = \mathbf{S}\mathbf{S}^{T}, \quad \boldsymbol{\varphi} = \left\{\mathbf{W}_{\mathbf{z}},\mathbf{S}\right\}$

- Resulting VAE energy function (i.e., – ELBO):

$$L_{VAE}(\boldsymbol{\theta},\boldsymbol{\varphi}) = \sum_{i=1}^{n}\Big\{ \mathrm{E}_{q_{\boldsymbol{\varphi}}\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\tfrac{1}{\lambda}\left\|\mathbf{x}^{(i)} - \boldsymbol{\Phi}\,\mathrm{diag}\left[\mathbf{w}_{\mathbf{x}}\right]\mathbf{z}\right\|_2^2\right] + d\log\lambda + \mathrm{tr}\left[\mathbf{S}\mathbf{S}^{T}\right] - \log\left|\mathbf{S}\mathbf{S}^{T}\right| + \left\|\mathbf{W}_{\mathbf{z}}\mathbf{x}^{(i)}\right\|_2^2\Big\}$$

⇨ nonconvex, potentially many local minima

Corresponding VAE Design

- VAE components for multiple sparse regression:

Decoder: $\boldsymbol{\mu}_{\mathbf{x}}[\mathbf{z},\boldsymbol{\theta}] = \boldsymbol{\Phi}\,\mathrm{diag}[\mathbf{w}_{\mathbf{x}}]\mathbf{z}, \quad \boldsymbol{\Sigma}_{\mathbf{x}}[\mathbf{z},\boldsymbol{\theta}] = \lambda\mathbf{I}, \quad \boldsymbol{\theta} = \{\mathbf{w}_{\mathbf{x}},\lambda\}$

Encoder: $\boldsymbol{\mu}_{\mathbf{z}}[\mathbf{x},\boldsymbol{\varphi}] = \mathbf{W}_{\mathbf{z}}\mathbf{x}, \quad \boldsymbol{\Sigma}_{\mathbf{z}}[\mathbf{x},\boldsymbol{\varphi}] = \mathbf{S}\mathbf{S}^T, \quad \boldsymbol{\varphi} = \{\mathbf{W}_{\mathbf{z}},\mathbf{S}\}$

- Resulting VAE energy function (i.e., – ELBO):

$$L_{VAE}(\boldsymbol{\theta},\boldsymbol{\varphi}) = \sum_{i=1}^{n}\Big\{\mathrm{E}_{q_{\boldsymbol{\varphi}}(\mathbf{z}|\mathbf{x}^{(i)})}\Big[\tfrac{1}{\lambda}\big\|\mathbf{x}^{(i)} - \boldsymbol{\Phi}\,\mathrm{diag}[\mathbf{w}_{\mathbf{x}}]\mathbf{z}\big\|_2^2\Big] + d\log\lambda$$
$$+ \mathrm{tr}\big[\mathbf{S}\mathbf{S}^T\big] - \log\big|\mathbf{S}\mathbf{S}^T\big| + \big\|\mathbf{W}_{\mathbf{z}}\mathbf{x}^{(i)}\big\|_2^2\Big\}$$

→ nonconvex, potentially many local minima

- Analogous AE for multiple sparse regression:

$$L_{AE}(\mathbf{w}_{\mathbf{x}},\mathbf{W}_{\mathbf{z}}) = \sum_{i=1}^{n}\tfrac{1}{\lambda}\big\|\mathbf{x}^{(i)} - \boldsymbol{\Phi}\,\mathrm{diag}[\mathbf{w}_{\mathbf{x}}]\mathbf{z}^{(i)}\big\|_2^2 + \underbrace{g\big(\|\mathbf{w}_{\mathbf{x}}\|_2\big)}_{\text{for avoiding scaling ambiguity}} + \underbrace{\sum_{j=1}^{\kappa} h\big(\|\mathbf{z}_{(j)}\|_2\big)}_{\text{promotes row sparsity}}$$

$$\text{s.t.}\quad \mathbf{Z} = \mathbf{W}_{\mathbf{z}}\mathbf{X} \in \mathbb{R}^{\kappa\times n}$$

Main Result

Assume: $\mathbf{X} = \mathbf{\Phi}\mathbf{U}_0 \in \mathbb{R}^{d \times n}, \quad \rho_0\left(\mathbf{U}_0\right) < d$ (+ other minor tech. cond. on $\mathbf{\Phi}$)

Main Result

Assume: $\mathbf{X} = \mathbf{\Phi U}_0 \in \mathbb{R}^{d \times n}, \quad \rho_0\left(\mathbf{U}_0\right) < d$ (+ other minor tech. cond. on $\mathbf{\Phi}$)

Denote: $\left\{\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*\right\} \equiv \left\{\mathbf{w}_{\mathbf{x}}^*, \mathbf{W}_{\mathbf{z}}^*, \mathbf{S}^*\right\} =$ any local minimum of $L_{VAE}\left(\boldsymbol{\theta}, \boldsymbol{\varphi}\right)$ as $\lambda \to 0$

Main Result

Assume: $\mathbf{X} = \mathbf{\Phi U}_0 \in \mathbb{R}^{d \times n}$, $\rho_0(\mathbf{U}_0) < d$ (+ other minor tech. cond. on $\mathbf{\Phi}$)

Denote: $\{\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*\} \equiv \{\mathbf{w}_\mathbf{x}^*, \mathbf{W}_\mathbf{z}^*, \mathbf{S}^*\} =$ any local minimum of $L_{VAE}(\boldsymbol{\theta}, \boldsymbol{\varphi})$ as $\lambda \to 0$

Theorem (informal version)

If $\mathbf{U}_0$ has nonzero row norms with sufficiently different *scales*, then we have that:

i) $\mathbf{U}_0$ ⇨ unique optimal sparse representation

ii) $\{\mathbf{w}_\mathbf{x}^*, \mathbf{W}_\mathbf{z}^*, \mathbf{S}^*\}$ is a VAE global minimum

iii) $\text{diag}\left[\mathbf{w}_\mathbf{x}^*\right] \mathbf{W}_\mathbf{z}^* \mathbf{X} = \mathbf{U}_0$

iv) No analogous AE can satisfy equivalent recovery result

bad local and/or global minima cannot be ruled out

Main Result

Assume: $\mathbf{X} = \mathbf{\Phi U}_0 \in \mathbb{R}^{d \times n}, \quad \rho_0(\mathbf{U}_0) < d$ (+ other minor tech. cond. on $\mathbf{\Phi}$)

Denote: $\{\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*\} \equiv \{\mathbf{w}_\mathbf{x}^*, \mathbf{W}_\mathbf{z}^*, \mathbf{S}^*\} =$ any local minimum of $L_{VAE}(\boldsymbol{\theta}, \boldsymbol{\varphi})$ as $\lambda \to 0$

Theorem (informal version)

If $\mathbf{U}_0$ has nonzero row norms with sufficiently different *scales*, then we have that:

i) $\mathbf{U}_0$ ⇨ unique optimal sparse representation

ii) $\{\mathbf{w}_\mathbf{x}^*, \mathbf{W}_\mathbf{z}^*, \mathbf{S}^*\}$ is a VAE global minimum

iii) $\text{diag}\left[\mathbf{w}_\mathbf{x}^*\right] \; \mathbf{W}_\mathbf{z}^* \mathbf{X} = \mathbf{U}_0$

iv) No analogous AE can satisfy equivalent recovery result

bad local and/or global minima cannot be ruled out

How is this possible? ⇨ Marginalization over VAE latent space introduces selective smoothing effect

Empirical Examples

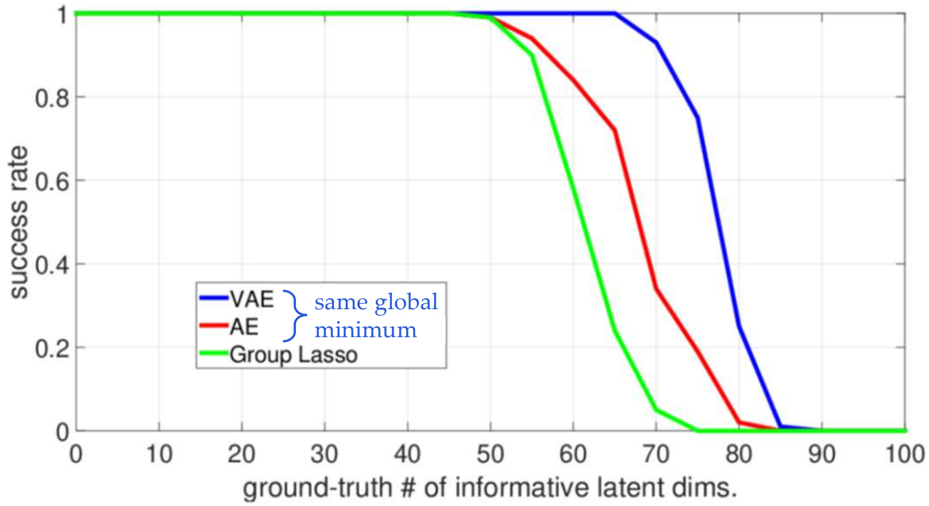
<u>Goal</u>: Recover ground-truth $\mathbf{U}_0$ from observations $\mathbf{X} = \mathbf{\Phi}\mathbf{U}_0$

Empirical Examples

Goal: Recover ground-truth $\mathbf{U}_0$ from observations $\mathbf{X} = \mathbf{\Phi}\mathbf{U}_0$

poor approx to theory

$\mathbf{\Phi} \sim N(0,1)$, $\mathbf{U}_0$ has unit-norm nonzero rows

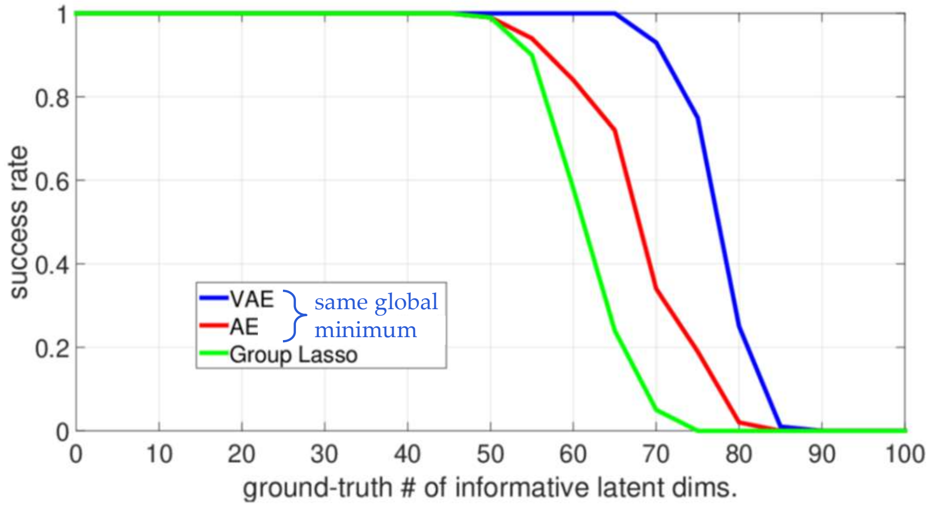


Empirical Examples

Goal: Recover ground-truth $\mathbf{U}_0$ from observations $\mathbf{X} = \mathbf{\Phi}\mathbf{U}_0$

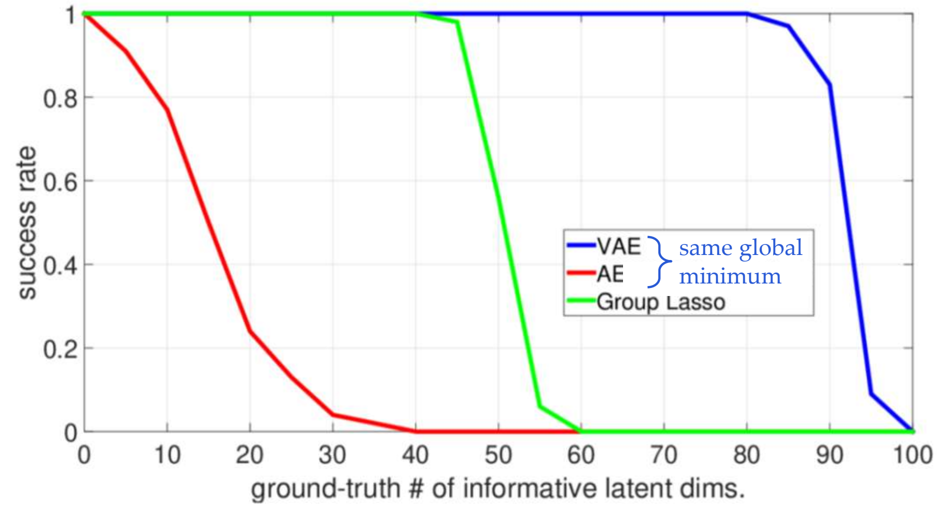
poor approx to theory

$\mathbf{\Phi} \sim N(0,1)$, $\mathbf{U}_0$ has unit-norm nonzero rows



close approx to theory

$\mathbf{\Phi} \sim N(0,1)$, $\mathbf{U}_0$ has scaled nonzero rows

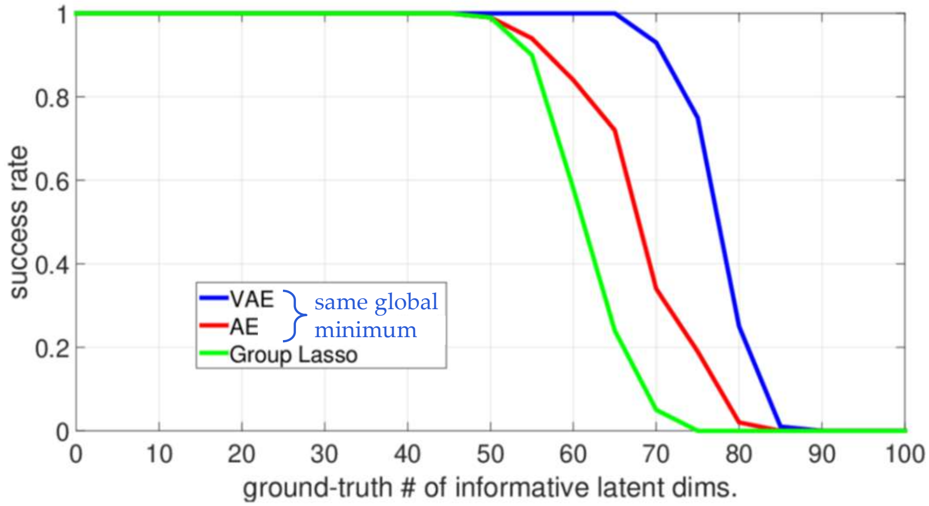


Empirical Examples

<u>Goal</u>: Recover ground-truth $\mathbf{U}_0$ from observations $\mathbf{X} = \mathbf{\Phi}\mathbf{U}_0$

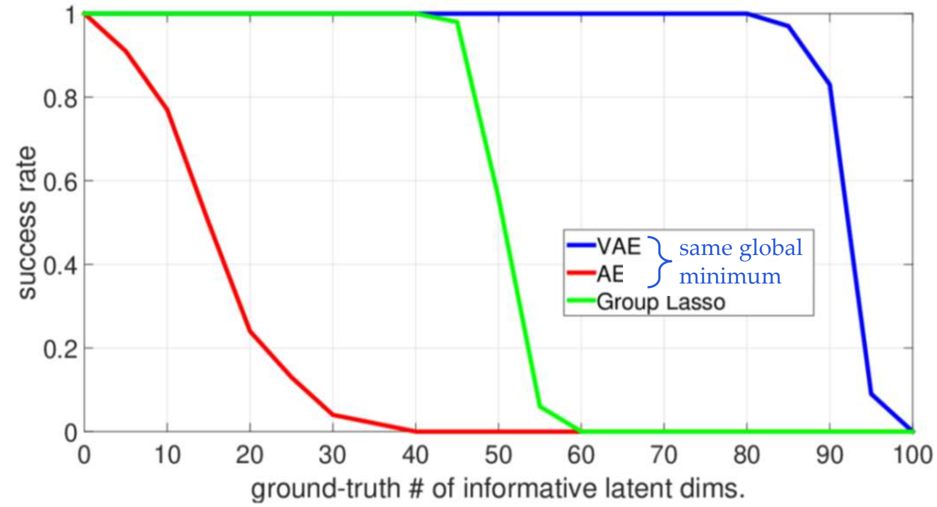
poor approx to theory

$\mathbf{\Phi} \sim N(0,1)$, $\mathbf{U}_0$ has unit-norm nonzero rows

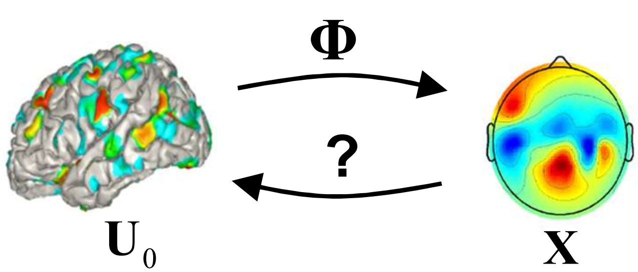


close approx to theory

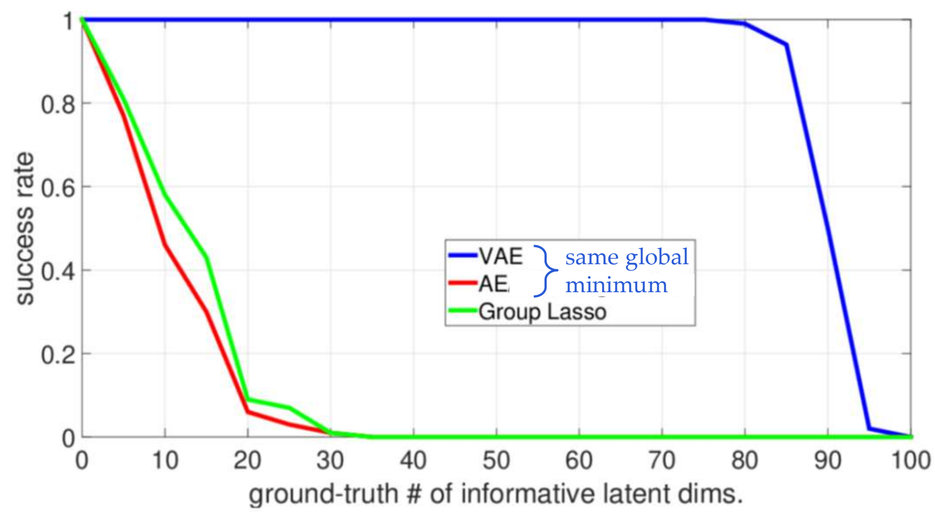
$\mathbf{\Phi} \sim N(0,1)$, $\mathbf{U}_0$ has scaled nonzero rows



Neuromagnetic inverse modeling



$\mathbf{\Phi}$ = leadfield matrix, highly correlated columns



Summary

- Prior work has shown that the VAE global minima can provably recover low-dimensional structure in data.
- But previously no strict guarantees (outside of cases that reduce to PCA) regarding bad local minima.
- We demonstrate a challenging regime whereby all VAE local minima produce optimal sparse representations.
- Made possible by VAE marginalization.
- Practical relevance:
  - Helps to explain the effectiveness of VAEs in modeling low-dimensional structure in data.
  - Motivates diverse VAE use cases beyond generating samples.
  - Theory makes accurate predictions regarding empirical VAE behavior in broader regimes of interest, e.g., more complex decoders.

Thank You

Links:

- http://www.davidwipf.com/
- http://www.dgl.ai/

If interested in graph neural networks or deep generative models, AWS Shanghai AI Lab is hiring interns.