Identifiable deep generative models via sparse decoding

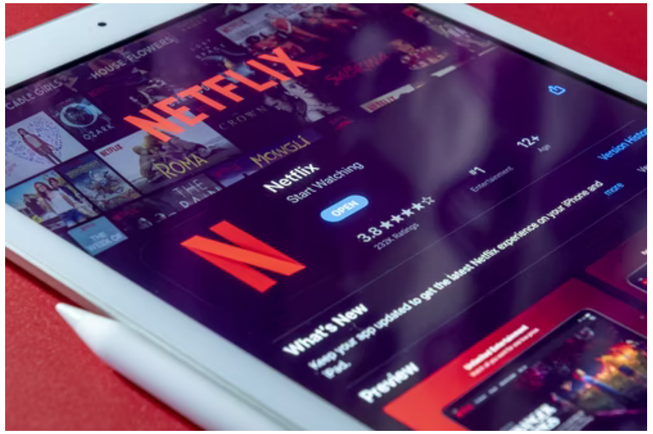
Gemma Moran
Columbia University

Collaborators

Dhanya Sridhar
University of Montreal and Mila

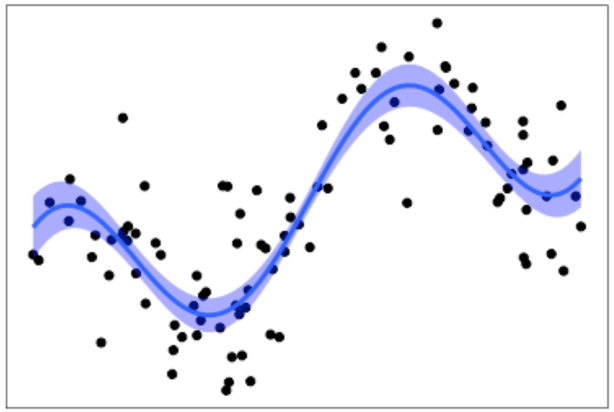
Yixin Wang
University of Michigan

Dave Blei
Columbia University



We have large, complicated data; we want **meaningful representations** of it.

What is a **meaningful representation**?



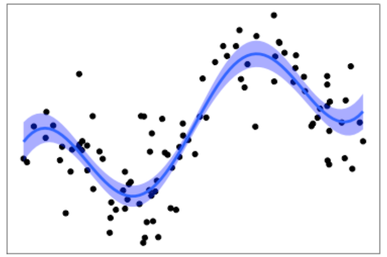
Flexible: capture nonlinear associations

Interpretable: what features are associated?

Identifiable: one optimal representation

Past work (non-exhaustive)

Interpretable



Flexible

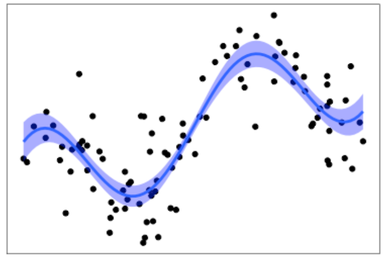
Linear methods

- Nonnegative matrix factorization
  Donoho and Stodden, 2003
- Anchored topic models *Arora et al. 2013*
- Anchored linear factor analysis *Bing et al. 2020*
- PCA+Varimax *Rohe and Zeng, 2020*

Identifiable

Past work (non-exhaustive)

Interpretable



Flexible

Linear methods

- Nonnegative matrix factorization *Donoho and Stodden, 2003*
- Anchored topic models *Arora et al. 2013*
- Anchored linear factor analysis *Bing et al. 2020*
- PCA+Varimax *Rohe and Zeng, 2020*

Identifiable

Weakly supervised methods

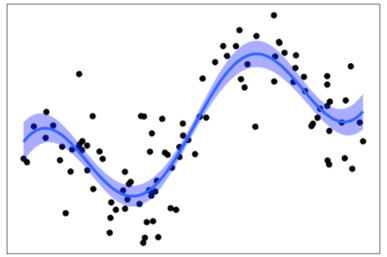
- Auxiliary data *Khemakhem et al. 2020*
- Data augmentation *von Kügelgen et al. 2021*
- Temporal information *Locatello et al. 2020, Hälvä et al. 2021, Lippe et al. 2022*

Constraints on function classes

- Independent mechanisms *Gresele et al. 2022*
- Sparse structure *Zheng et al. 2022*

Past work (non-exhaustive)

- Output-Interpretable VAE
  Ainsworth et al. 2018



Interpretable

Flexible

Linear methods

- Nonnegative matrix factorization
  Donoho and Stodden, 2003
- Anchored topic models *Arora et al. 2013*
- Anchored linear factor analysis *Bing et al. 2020*
- PCA+Varimax *Rohe and Zeng, 2020*

Identifiable

Weakly supervised methods

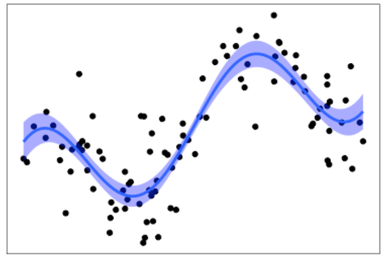
- Auxiliary data *Khemakhem et al. 2020*
- Data augmentation *von Kügelgen et al. 2021*
- Temporal information *Locatello et al. 2020, Hälvä et al. 2021, Lippe et al. 2022*

Constraints on function classes

- Independent mechanisms *Gresele et al. 2022*
- Sparse structure *Zheng et al. 2022*

Past work (non-exhaustive)

- Output-Interpretable VAE
  Ainsworth et al. 2018



Interpretable

Identifiable deep generative models via sparse decoding *Moran et al. 2022*

Flexible

Linear methods

- Nonnegative matrix factorization *Donoho and Stodden, 2003*
- Anchored topic models *Arora et al. 2013*
- Anchored linear factor analysis *Bing et al. 2020*
- PCA+Varimax *Rohe and Zeng, 2020*

Identifiable

Weakly supervised methods

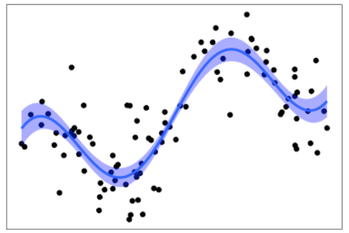
- Auxiliary data *Khemakhem et al. 2020*
- Data augmentation *von Kügelgen et al. 2021*
- Temporal information *Locatello et al. 2020, Hälvä et al. 2021, Lippe et al. 2022*

Constraints on function classes

- Independent mechanisms *Gresele et al. 2022*
- Sparse structure *Zheng et al. 2022*

This work

The Sparse VAE: introduce sparsity into a flexible factor-to-feature mapping

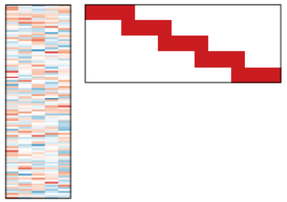


- **Flexible:** fits a neural network model with a variational autoencoder (VAE)

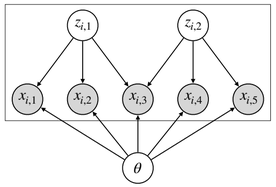
- **Interpretable:** can inspect which features are important for the representation

- **Identifiable:** a single optimal representation (under certain assumptions)

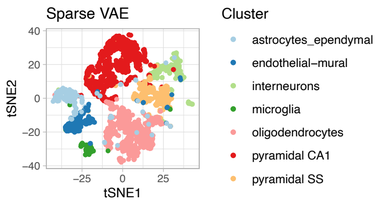
Roadmap



Identifiability and the role of sparsity



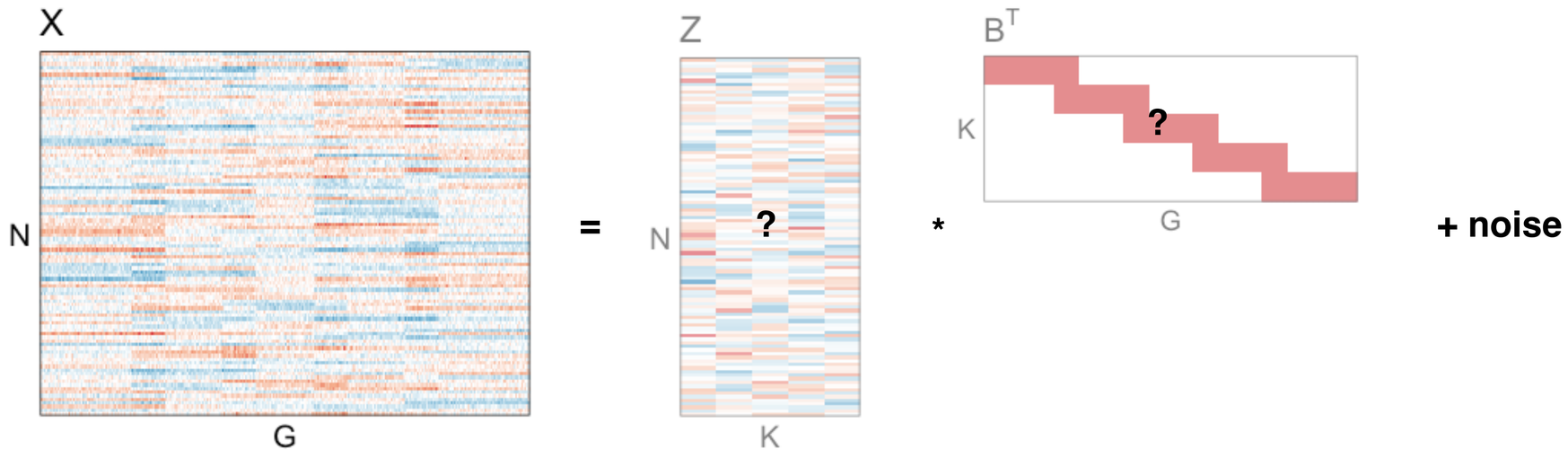
The Sparse VAE



Experimental results

The linear case

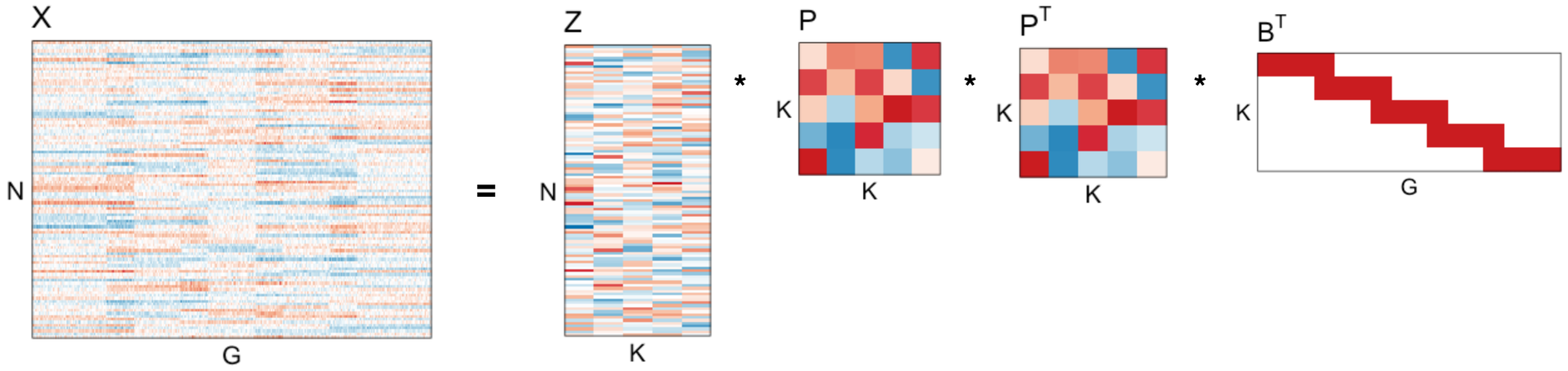
Linear factor analysis model: $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{B}^T + \boldsymbol{E}$



- Data: $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times G}$
- Latent factors: $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^T \in \mathbb{R}^{N \times K}$
- Latent loadings: $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K] \in \mathbb{R}^{G \times K}$
- Noise: $\boldsymbol{E} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N] \in \mathbb{R}^{N \times G}$

Lack of identifiability

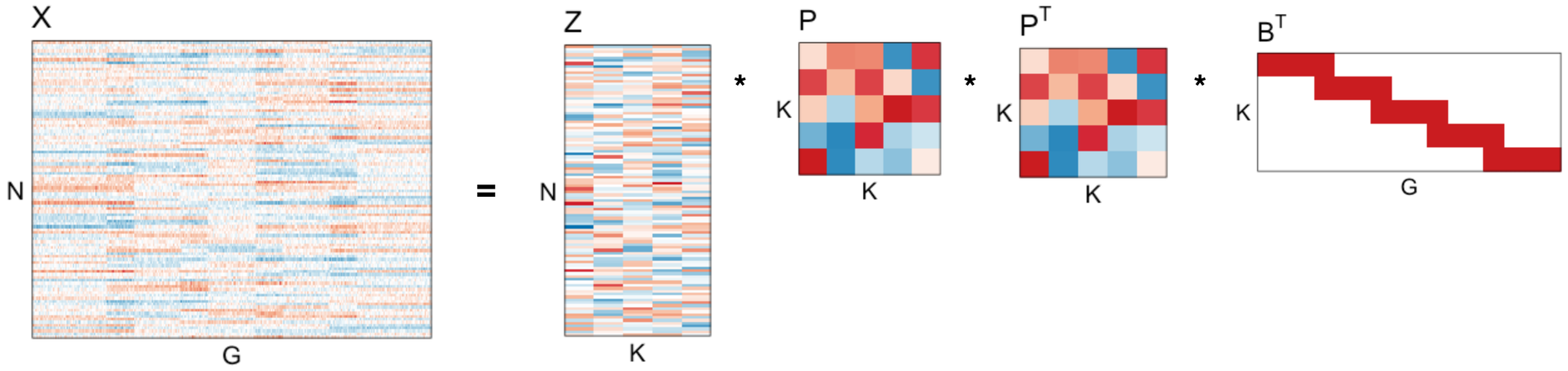
Linear factor analysis model: $\boldsymbol{X} = (\boldsymbol{ZP})(\boldsymbol{BP})^T + \boldsymbol{E}, \quad \boldsymbol{PP}^T = \boldsymbol{I}$



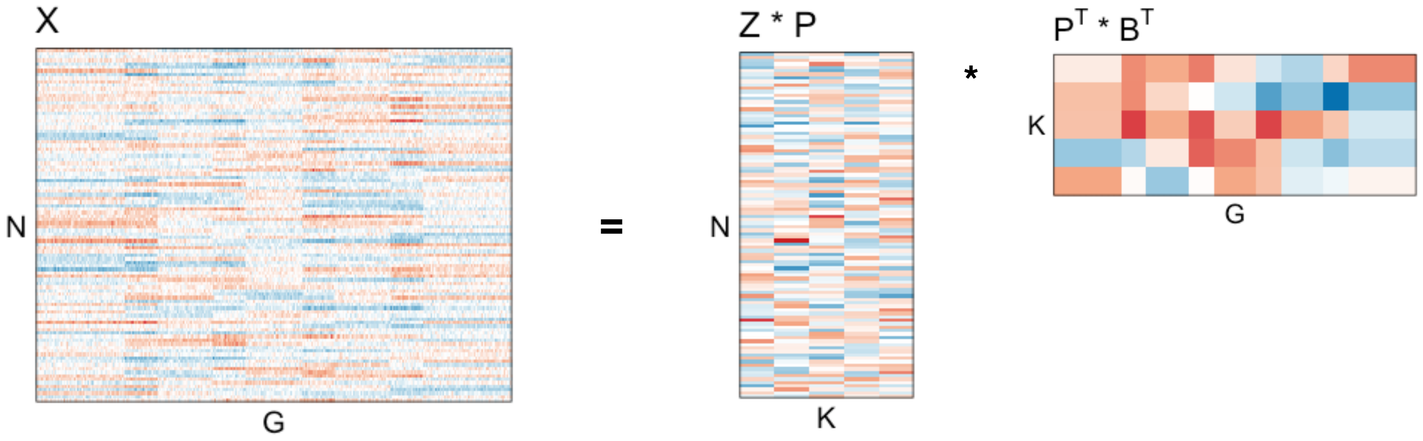
+ noise

Lack of identifiability

Linear factor analysis model: $\boldsymbol{X} = (\boldsymbol{Z}\boldsymbol{P})(\boldsymbol{B}\boldsymbol{P})^T + \boldsymbol{E}, \quad \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}$



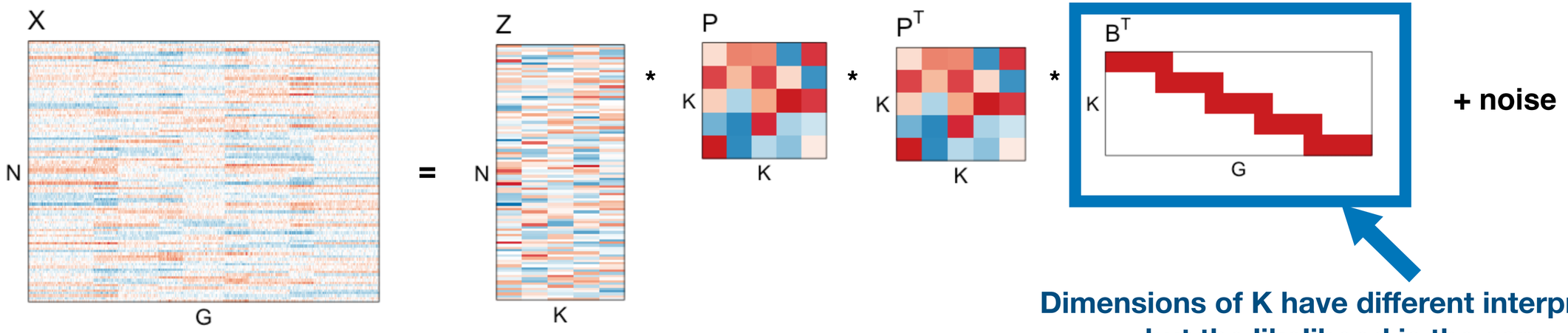
+ noise



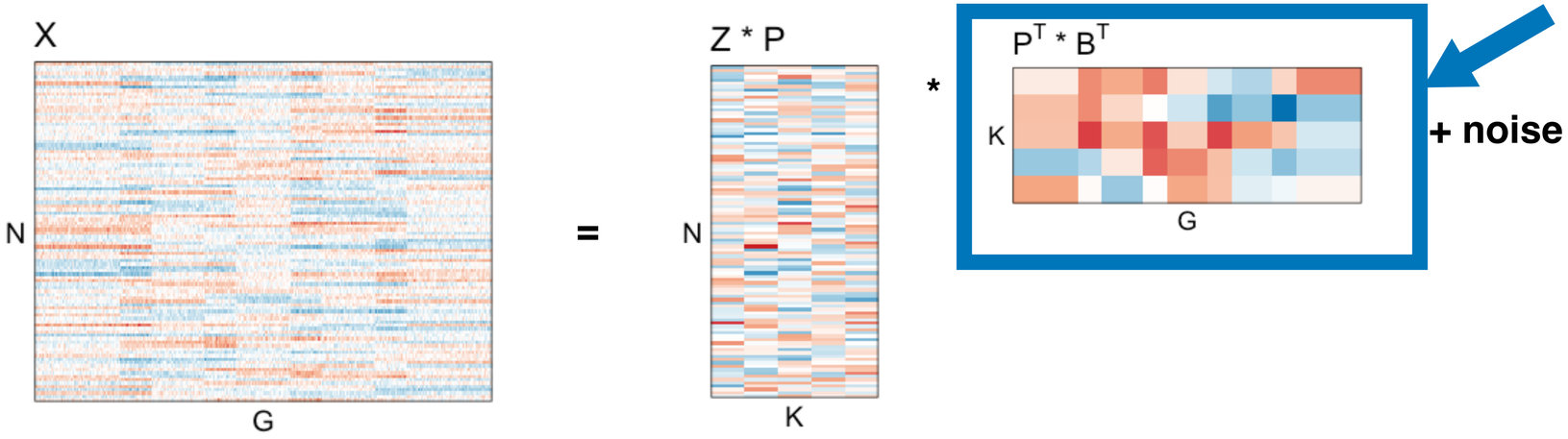
+ noise

Lack of identifiability

Linear factor analysis model: $\boldsymbol{X} = (\boldsymbol{ZP})(\boldsymbol{BP})^T + \boldsymbol{E}, \quad \boldsymbol{PP}^T = \boldsymbol{I}$



Dimensions of K have different interpretations but the likelihood is the same... which to choose?

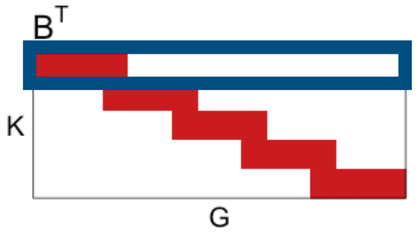


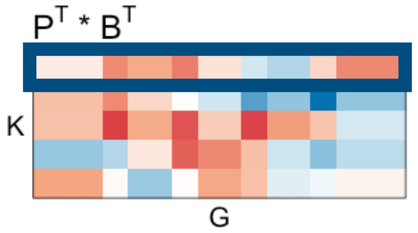
Lack of identifiability

Some problems:

- **Exploratory data analysis:** may select wrong features for further hypothesis testing

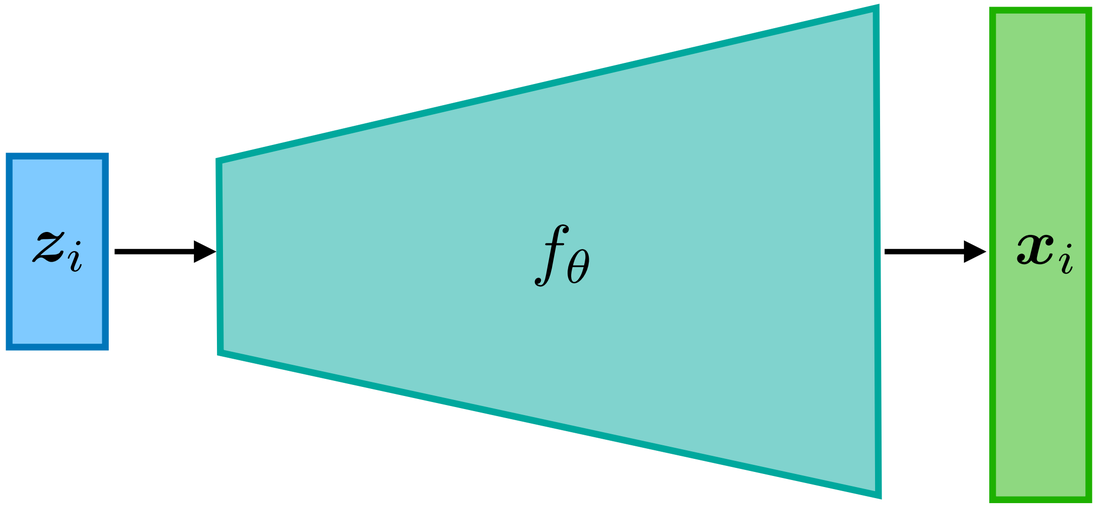
- **Generalization:** if factor-to-feature mapping mixes dimensions, may not generalize well to differently distributed factors





The nonlinear case

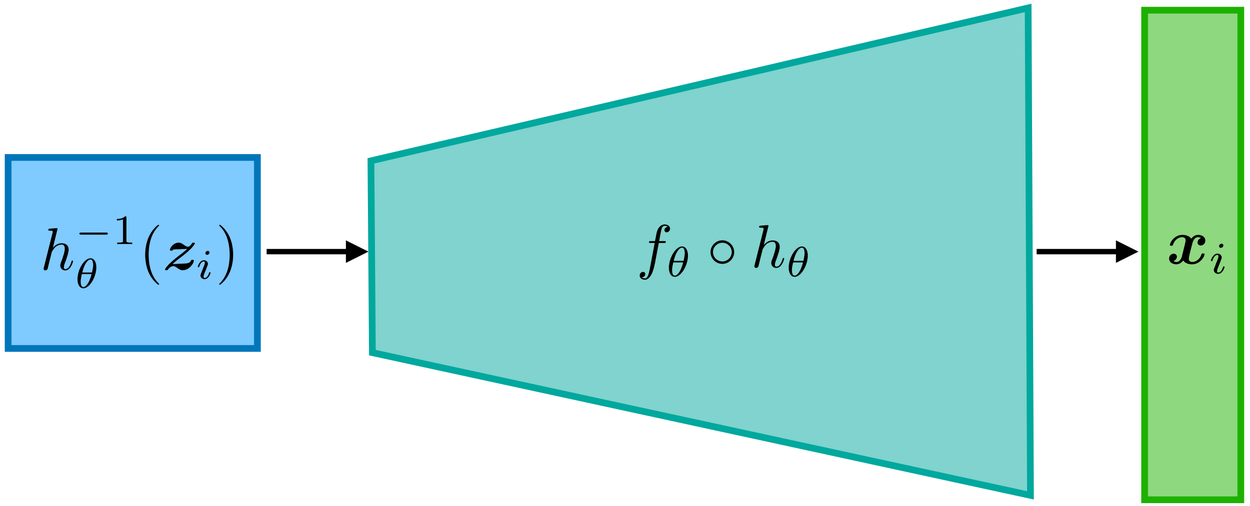
- Deep generative model: $\boldsymbol{x}_i = f_\theta(\boldsymbol{z}_i) + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_G(0, \Sigma)$



- Data: $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times G}$
- Latent factors: $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^T \in \mathbb{R}^{N \times K}$
- Neural network: $f_\theta : \mathbb{R}^K \to \mathbb{R}^G$

The nonlinear case

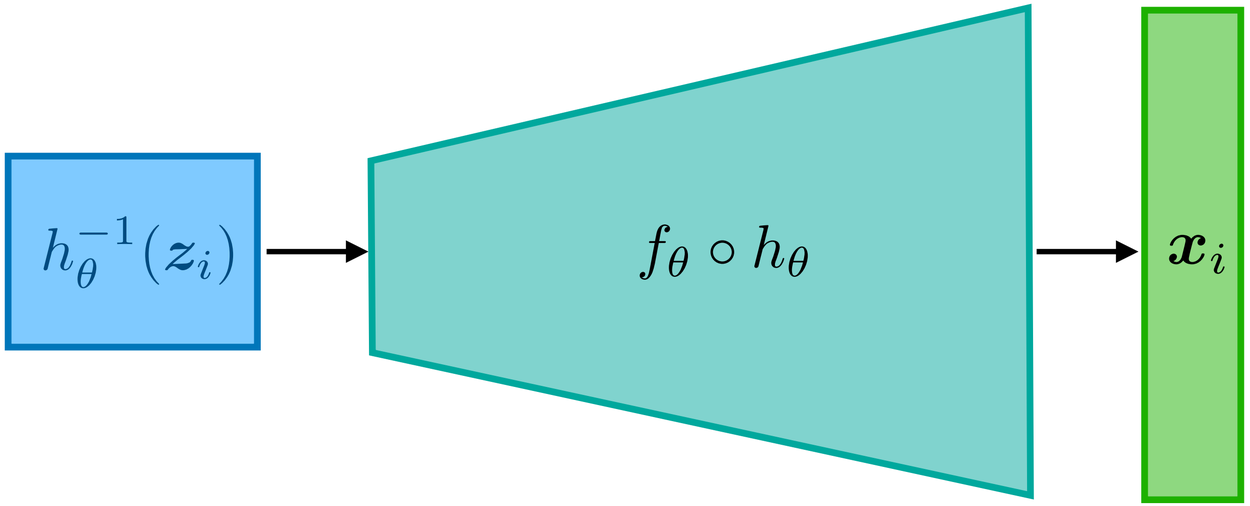
- Deep generative model: $\boldsymbol{x}_i = f_\theta(h_\theta(h_\theta^{-1}(\boldsymbol{z}_i))) + \boldsymbol{\varepsilon}_i$



- Data: $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times G}$
- Latent factors: $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^T \in \mathbb{R}^{N \times K}$
- Neural network: $f_\theta : \mathbb{R}^K \to \mathbb{R}^G$

The nonlinear case

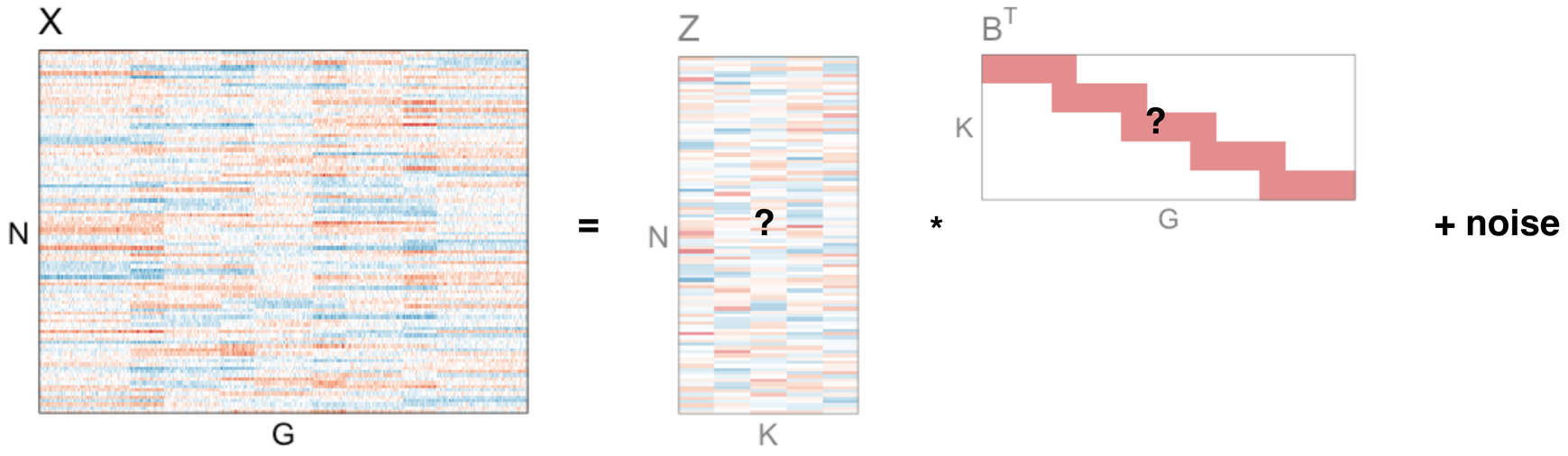
- Deep generative model: $\boldsymbol{x}_i = f_\theta(h_\theta(h_\theta^{-1}(\boldsymbol{z}_i))) + \boldsymbol{\varepsilon}_i$



z_i and $h_\theta^{-1}(z_i)$ can have very different interpretations... which to choose?

- Data: $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N \times G}$
- Latent factors: $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^T \in \mathbb{R}^{N \times K}$
- Neural network: $f_\theta : \mathbb{R}^K \to \mathbb{R}^G$

How can sparsity help?

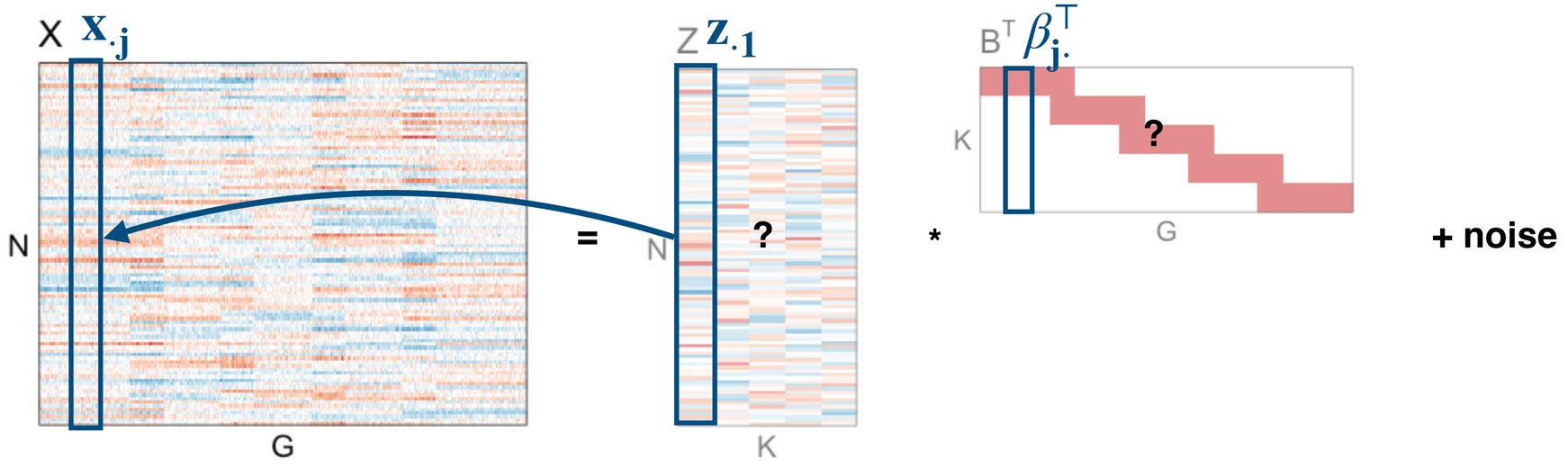


Linear factor analysis

<u>Theorem</u> [*Bing et al. 2020*] Factors z are identifiable if, for every $k = 1, \ldots, K$, there are at least two **anchor features** (+ other assumptions)

How can sparsity help?

$$\mathbf{x}_{\cdot \mathbf{j}} = \mathbf{z}_{\cdot \mathbf{1}} \times \beta_{j1} + \varepsilon_{\cdot j}$$



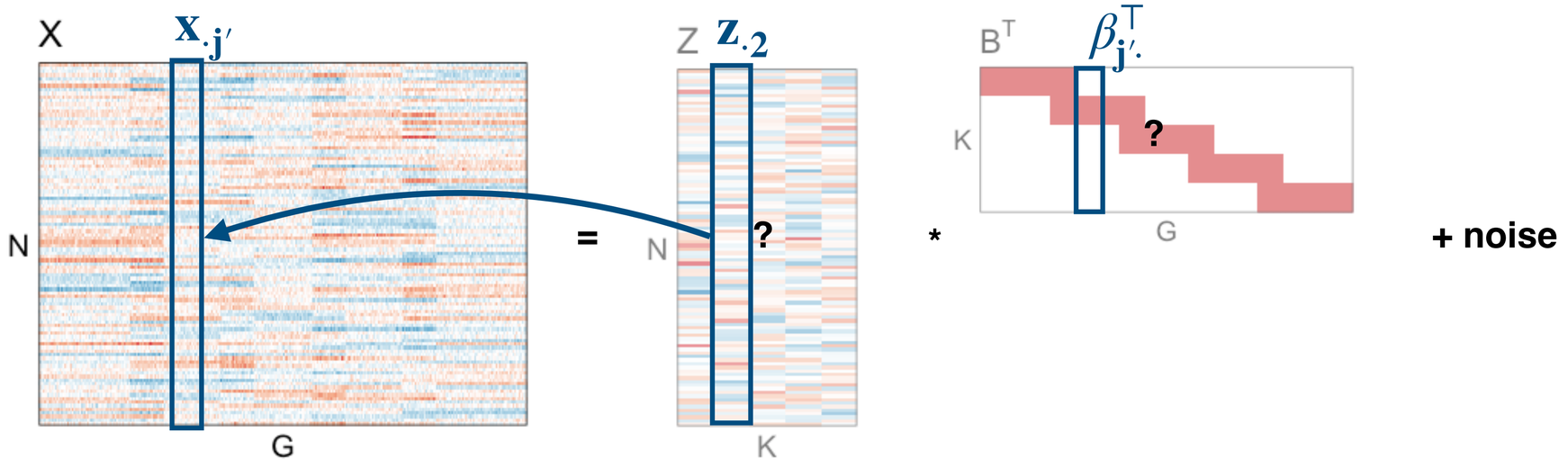
Linear factor analysis

<u>Theorem</u> [*Bing et al. 2020*] Factors z are identifiable if, for every $k = 1, \ldots, K$,

there are at least two **anchor features** (+ other assumptions)

How can sparsity help?

$$\mathbf{x}_{\cdot \mathbf{j}'} = \mathbf{z}_{\cdot \mathbf{2}} \times \beta_{j'2} + \varepsilon_{\cdot j'}$$



Linear factor analysis

<u>Theorem</u> [*Bing et al. 2020*] Factors z are identifiable if, for every $k = 1, \ldots, K$,

there are at least two **anchor features** (+ other assumptions)

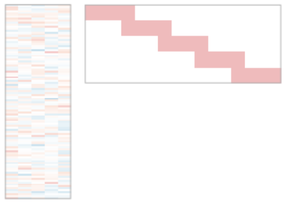
How can sparsity help?

Our work: how do we generalize this to the nonlinear case?

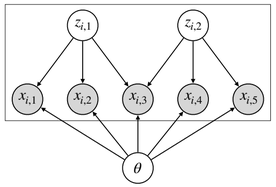
Linear factor analysis

Theorem [*Bing et al. 2020*] Factors z are identifiable if, for every $k = 1, \ldots, K$, there are at least two **anchor features** (+ other assumptions)

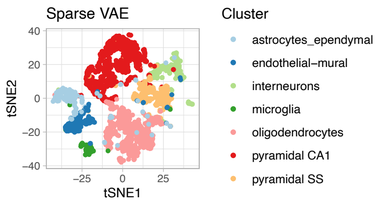
Roadmap



Identifiability and the role of sparsity



The Sparse VAE

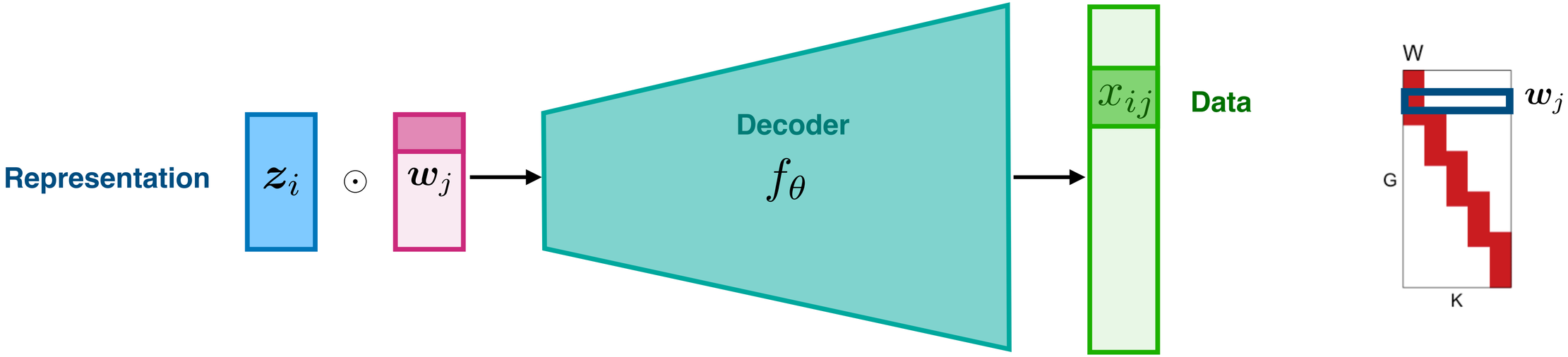


Experimental results

Setup

- Sparse deep generative model: $x_{ij} = f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i)_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$

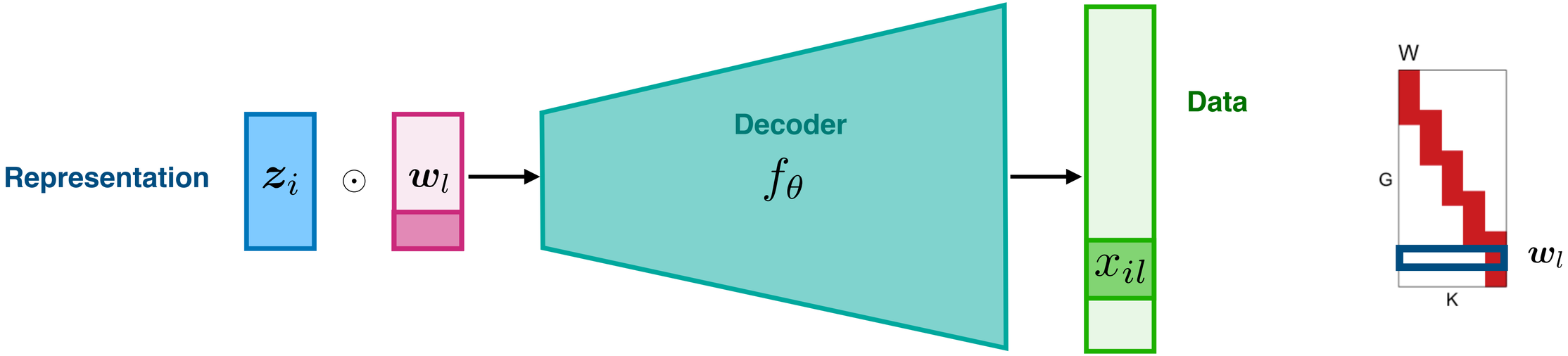
- Introduces masking matrix: $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_G]^T \in \mathbb{R}^{G \times K}$



Setup

- Sparse deep generative model: $x_{ij} = f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i)_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$

- Introduces masking matrix: $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_G]^T \in \mathbb{R}^{G \times K}$

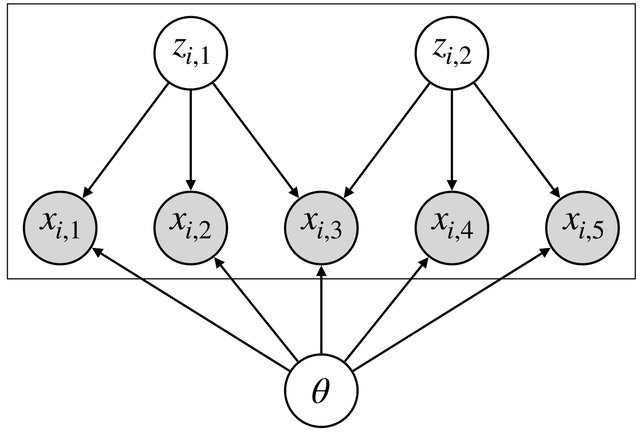


Setup

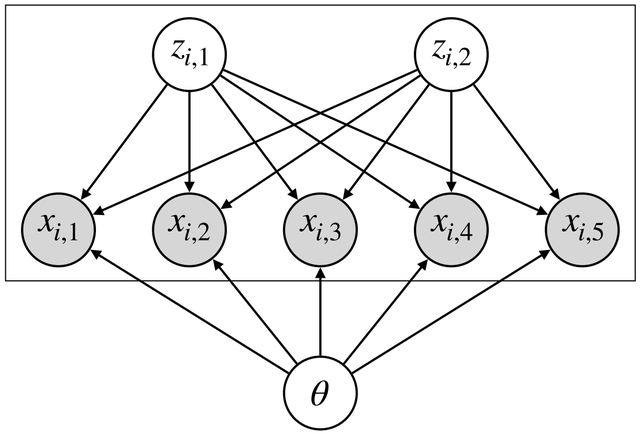
- Sparse deep generative model: $x_{ij} = f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i)_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$

- Introduces masking matrix: $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_G]^T \in \mathbb{R}^{G \times K}$

Sparse deep generative model



Deep generative model

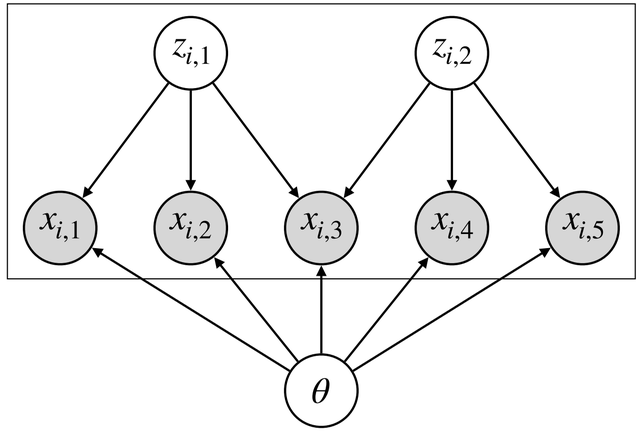


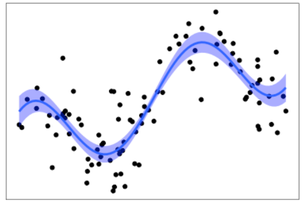
Setup

- Sparse deep generative model: $x_{ij} = f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i)_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$

- Introduces masking matrix: $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_G]^T \in \mathbb{R}^{G \times K}$

Sparse deep generative model





Flexible: uses neural networks for encoder and decoder

Interpretable: can inspect which features depend on which factors

Identifiability

Identifiable: a single optimal representation

Formal definition:

- if two solutions $\{\theta, Z\}$ and $\{\widetilde{\theta}, \widetilde{Z}\}$ have equal likelihood $p_\theta(\boldsymbol{X}|\boldsymbol{Z}) = p_{\widetilde{\theta}}(\boldsymbol{X}|\widetilde{\boldsymbol{Z}})$, then we must have $\{\theta, Z\} = \{\widetilde{\theta}, \widetilde{Z}\}$.

Identifiability

Identifiable: a single optimal representation

Formal definition:

- if two solutions $\{\theta, Z\}$ and $\{\widetilde{\theta}, \widetilde{Z}\}$ have equal likelihood $p_\theta(\boldsymbol{X}|\boldsymbol{Z}) = p_{\widetilde{\theta}}(\boldsymbol{X}|\widetilde{\boldsymbol{Z}})$, then we must have $\{\theta, Z\} = \{\widetilde{\theta}, \widetilde{Z}\}$.

Our goal:

- each dimension of $Z = (z_{\cdot 1}, \ldots, z_{\cdot K})$ is identified up to coordinate-wise transform
- i.e. if two solutions $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}}$ have equal likelihood, then $z_{\cdot k} = g_k(\widetilde{z}_{\cdot k})$

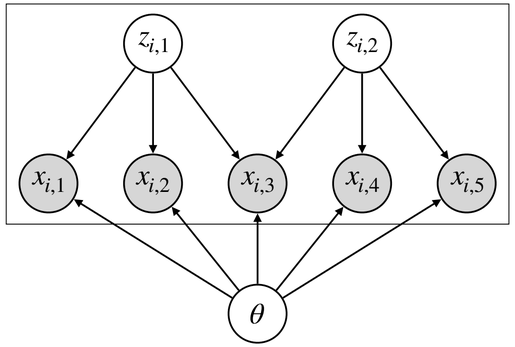
Identifiability

- <u>Theorem</u> *[Moran et al. 2021]* Factors Z are identifiable (up to coordinate-wise transform) under the following assumptions:

For each factor k, there are two features with

$$\mathbb{E}[x_{ij}|\boldsymbol{z}_i] = \mathbb{E}[x_{ij'}|\boldsymbol{z}_i] = f(z_{ik})$$

1. Anchor features



Suppose j is an anchor feature for factor k. For a non anchor feature l, we have

$$|w_{jk}| \sum_{i=1}^{N} |B_{ijk}|^2 \geq \sum_{i=1}^{N} \sum_{k'=1}^{K} |w_{lk'}||B_{ijk}||B_{ilk'}|$$

where $B_{ijk} = \dfrac{\partial f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i))_j}{\partial z_{ik}}$.

2. Scale of neural network

$$\min\{\mathrm{Var}(z_{ik}), \mathrm{Var}(z_{ik'})\} > |\mathrm{Cov}(z_{ik}, z_{ik'})|$$

3. Factor covariance

Identifiability

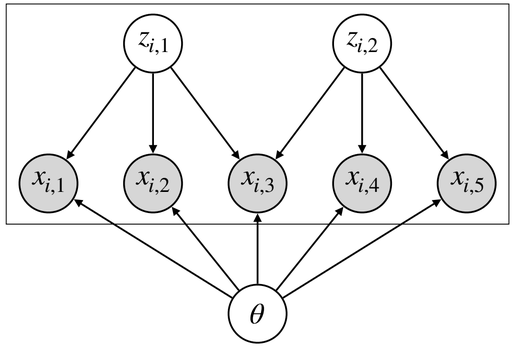
- <u>Theorem</u> *[Moran et al. 2021]* Factors Z are identifiable (up to coordinate-wise transform) under the following assumptions:

For each factor k, there are two features with

$$\mathbb{E}[x_{ij}|\boldsymbol{z}_i] = \mathbb{E}[x_{ij'}|\boldsymbol{z}_i] = f(z_{ik})$$

don't need to be known in advance

1. Anchor features



Suppose j is an anchor feature for factor k. For a non anchor feature l, we have

$$|w_{jk}| \sum_{i=1}^{N} |B_{ijk}|^2 \geq \sum_{i=1}^{N} \sum_{k'=1}^{K} |w_{lk'}||B_{ijk}||B_{ilk'}|$$

where $B_{ijk} = \dfrac{\partial f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i))_j}{\partial z_{ik}}$.

2. Scale of neural network

$$\min\{\mathrm{Var}(z_{ik}), \mathrm{Var}(z_{ik'})\} > |\mathrm{Cov}(z_{ik}, z_{ik'})|$$

3. Factor covariance

Identifiability

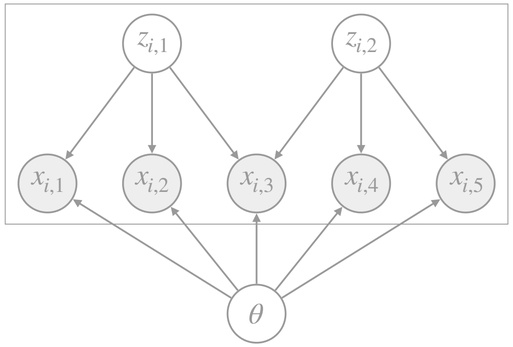
- <u>Theorem</u> *[Moran et al. 2021]* Factors Z are identifiable (up to coordinate-wise transform) under the following assumptions:

For each factor k, there are two features with

$$\mathbb{E}[x_{ij}|\boldsymbol{z}_i] = \mathbb{E}[x_{ij'}|\boldsymbol{z}_i] = f(z_{ik})$$

don't need to be known in advance

1. Anchor features

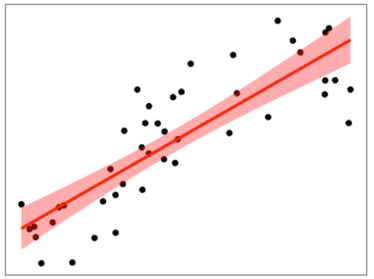


Suppose j is an anchor feature for factor k. For a non anchor feature l, we have

$$|w_{jk}| \sum_{i=1}^{N} |B_{ijk}|^2 \geq \sum_{i=1}^{N} \sum_{k'=1}^{K} |w_{lk'}||B_{ijk}||B_{ilk'}|$$

where $B_{ijk} = \dfrac{\partial f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i))_j}{\partial z_{ik}}$.

2. Scale of neural network



$$\min\{\mathrm{Var}(z_{ik}), \mathrm{Var}(z_{ik'})\} > |\mathrm{Cov}(z_{ik}, z_{ik'})|$$

3. Factor covariance

Identifiability

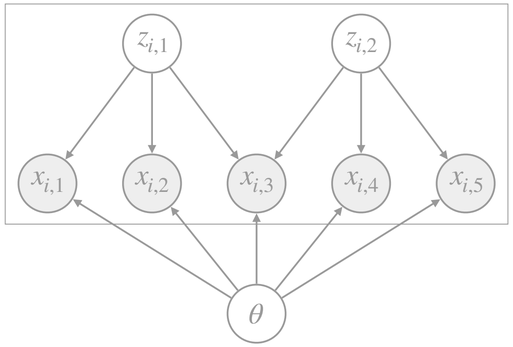
- <u>Theorem</u> *[Moran et al. 2021]* Factors $\boldsymbol{Z}$ are identifiable (up to coordinate-wise transform) under the following assumptions:

For each factor k, there are two features with

$$\mathbb{E}[x_{ij}|\boldsymbol{z}_i] = \mathbb{E}[x_{ij'}|\boldsymbol{z}_i] = f(z_{ik})$$

don't need to be known in advance

1. Anchor features

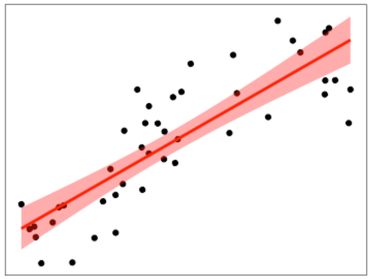


Suppose j is an anchor

Satisfied by neural networks with ReLU activations and independent weights

where $B_{ijk} = \frac{\partial f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i))_j}{\partial z_{ik}}$.

2. Scale of neural network



$$\min\{\mathrm{Var}(z_{ik}), \mathrm{Var}(z_{ik'})\} > |\mathrm{Cov}(z_{ik}, z_{ik'})|$$

3. Factor covariance

Identifiability

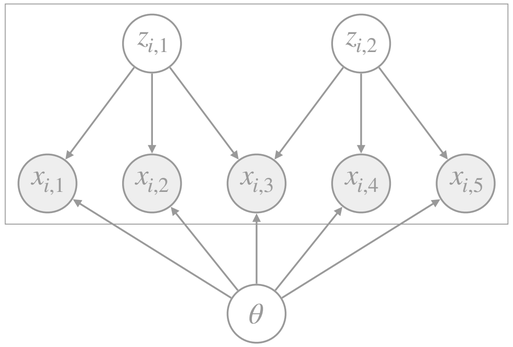
- <u>Theorem</u> *[Moran et al. 2021]* Factors $\boldsymbol{Z}$ are identifiable (up to coordinate-wise transform) under the following assumptions:

For each factor k, there are two features with

$$\mathbb{E}[x_{ij}|\boldsymbol{z}_i] = \mathbb{E}[x_{ij'}|\boldsymbol{z}_i] = f(z_{ik})$$

don't need to be known in advance

1. Anchor features

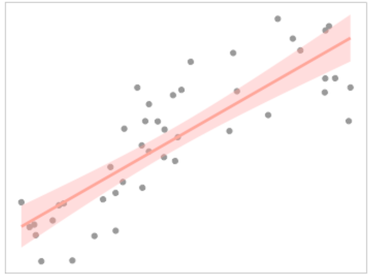


Suppose j is an anchor

Satisfied by neural networks with ReLU activations and independent weights

where $B_{ijk} = \frac{\partial f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i))_j}{\partial z_{ik}}$.

2. Scale of neural network



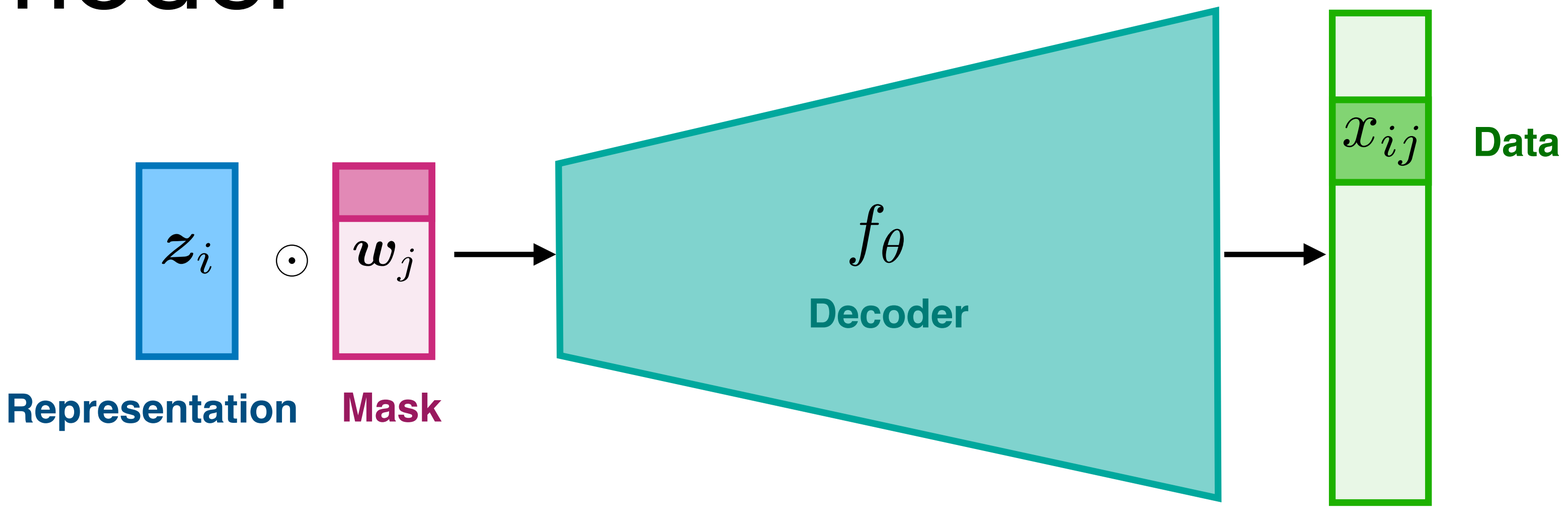
$$\min\{\mathrm{Var}(z_{ik}), \mathrm{Var}(z_{ik'})\} > |\mathrm{Cov}(z_{ik}, z_{ik'})|$$

3. Factor covariance

The Sparse VAE model

Generative model:

$$\theta \sim \mathcal{N}(0, \tau)$$
$$w_{jk} \sim \text{Spike-and-Slab Lasso}(\lambda_0, \lambda_1, a, b)$$
$$\boldsymbol{z}_i \sim \mathcal{N}_K(0, \boldsymbol{C})$$
$$x_{ij} \sim \mathcal{N}((f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i)_j, \sigma_j^2)$$



The Sparse VAE model

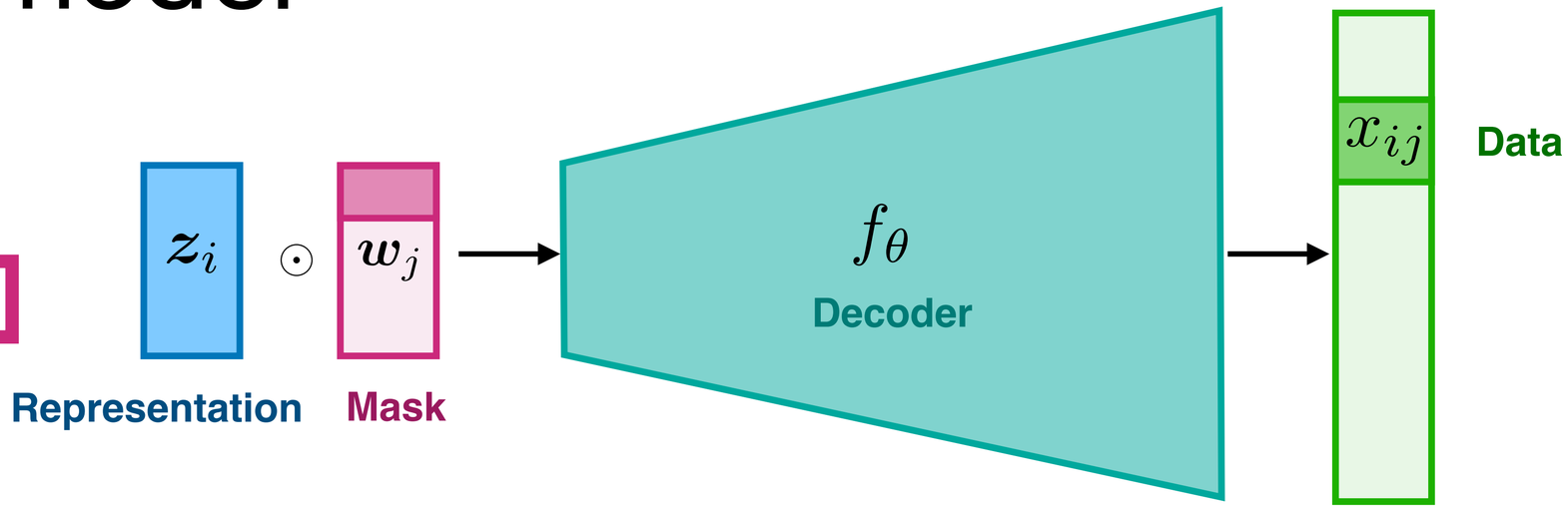
Generative model:

$$\theta \sim \mathcal{N}(0, \tau)$$

$$\boxed{w_{jk} \sim \text{Spike-and-Slab Lasso}(\lambda_0, \lambda_1, a, b)}$$

$$\boldsymbol{z}_i \sim \mathcal{N}_K(0, \boldsymbol{C})$$

$$x_{ij} \sim \mathcal{N}((f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i)_j, \sigma_j^2)$$

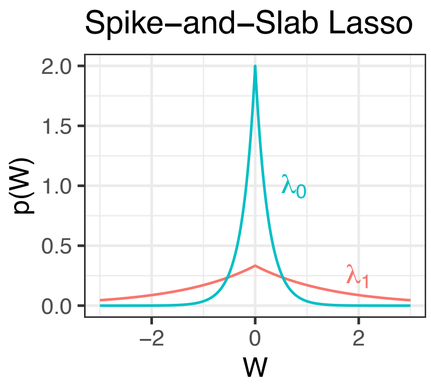


Induce sparsity in W via Spike-and-Slab Lasso prior
[Rockova and George, 2018]

$$\eta_k \sim \text{Beta}(a, b)$$

$$\gamma_{jk} \sim \text{Bernoulli}(\eta_k)$$

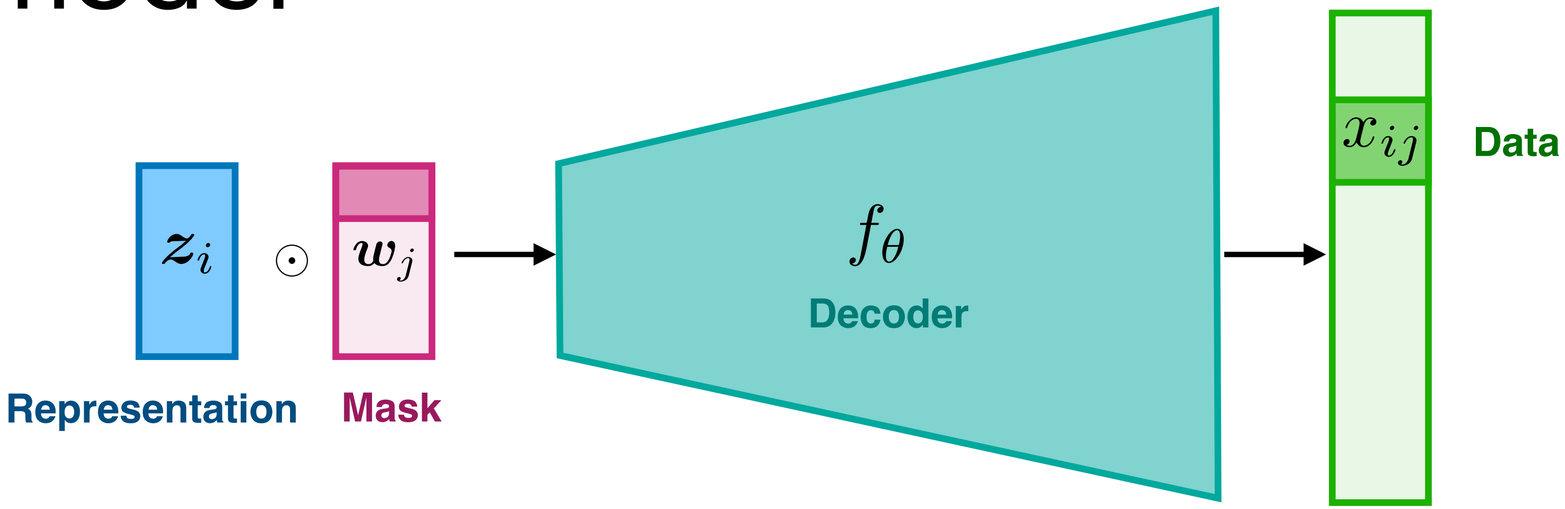
$$w_{jk} = \gamma_{jk} \frac{\lambda_1}{2} e^{-\lambda_1 |w_{jk}|} + (1 - \gamma_{jk}) \frac{\lambda_0}{2} e^{-\lambda_0 |w_{jk}|}$$



The Sparse VAE model

Generative model:

$$\theta \sim \mathcal{N}(0, \tau)$$
$$w_{jk} \sim \text{Spike-and-Slab Lasso}(\lambda_0, \lambda_1, a, b)$$
$$\boldsymbol{z}_i \sim \mathcal{N}_K(0, \boldsymbol{C})$$
$$x_{ij} \sim \mathcal{N}((f_\theta(\boldsymbol{w}_j \odot \boldsymbol{z}_i)_j, \sigma_j^2)$$

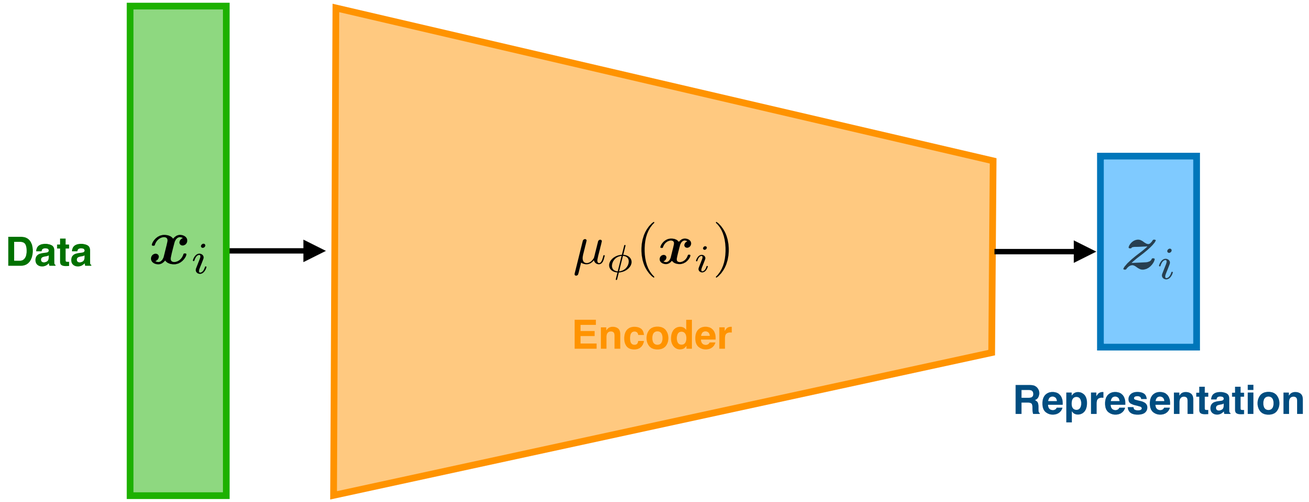


Inference:

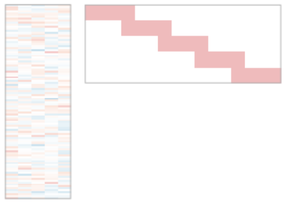
- Fit with a variational EM algorithm
- Variational approximation for z_i:

$$q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i) \sim \mathcal{N}_K(\mu_\phi(\boldsymbol{x}_i), \sigma_\phi^2(\boldsymbol{x}_i))$$

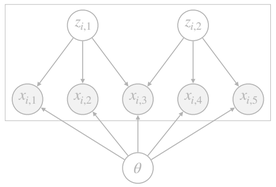
- Stochastic gradient ascent for W, θ, ϕ
- EM steps for Spike-and-Slab Lasso parameters



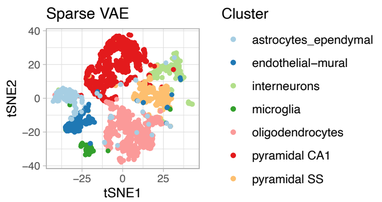
Roadmap



Identifiability and the role of sparsity



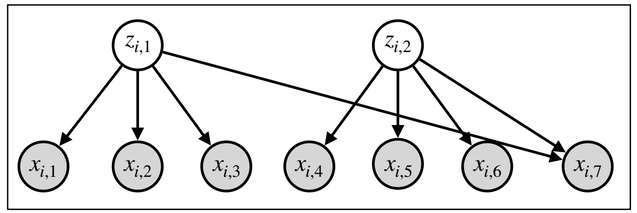
The Sparse VAE



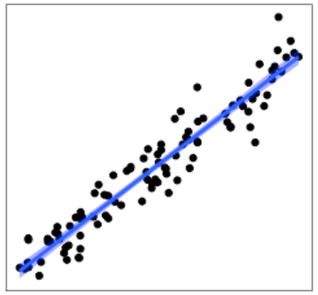
Experimental results

Experiments

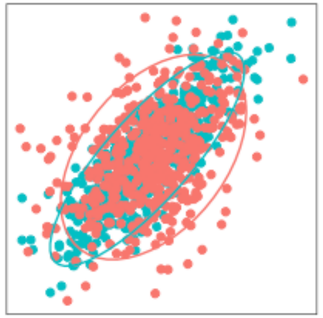
1. Recovering true factors



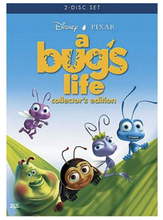
2. Prediction



3. Domain adaptation (train/test data have different distributions)



4. Interpretability



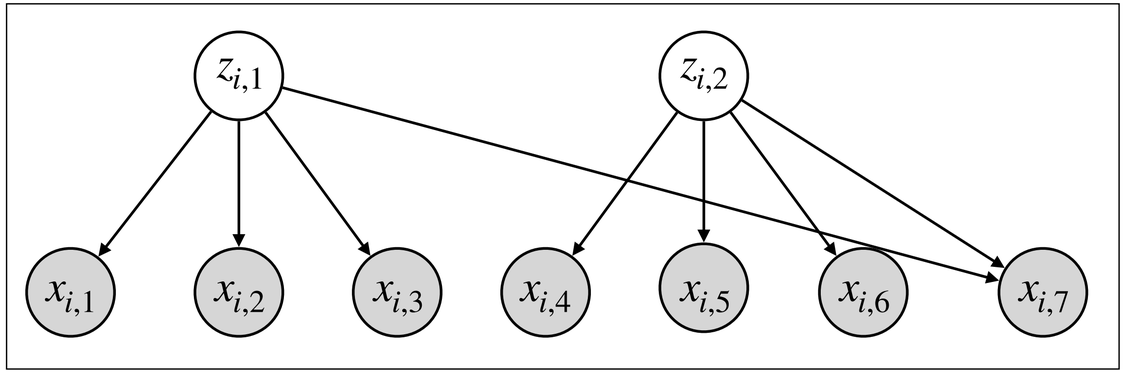
Experiments

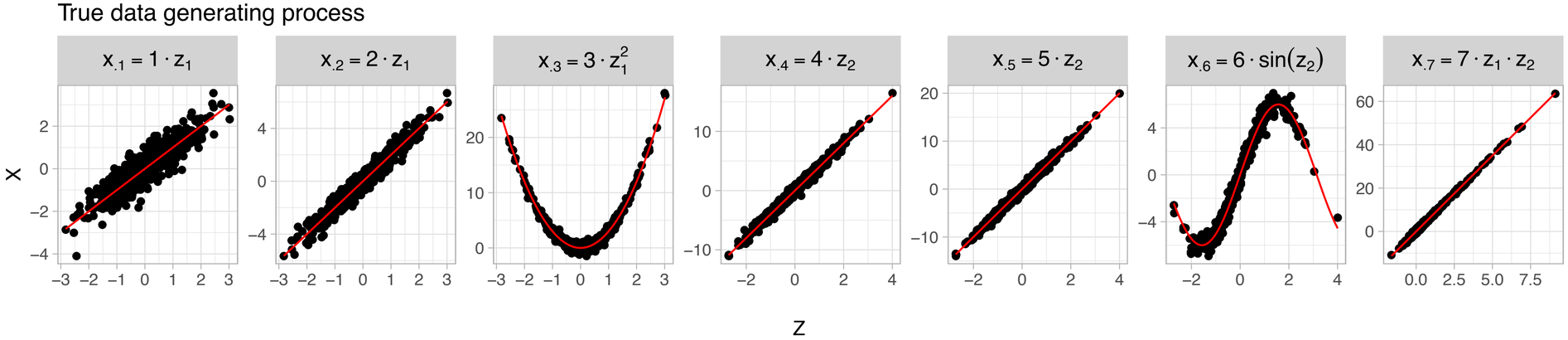
Methods compared	Identifiable?	Sparse?
Sparse VAE *[Moran et al. 2021]*	✓	✓ Decoder
VAE [*Kingma and Welling, 2014; Rezende et al. 2014]*	✗	✗
beta-VAE *[Higgins et al. 2017]*	✗	✗
Variational Sparse Coding *[Tonolini et al. 2020]*	✗	✓ Factors z_i

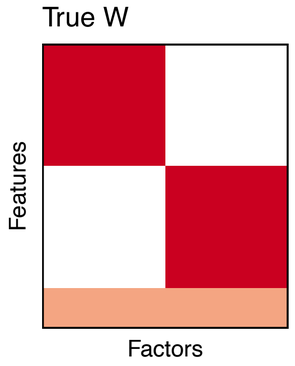
Recovering true factors

Synthetic Gaussian data

$N = 1000$ samples, $G = 7$ features, $K = 2$ factors



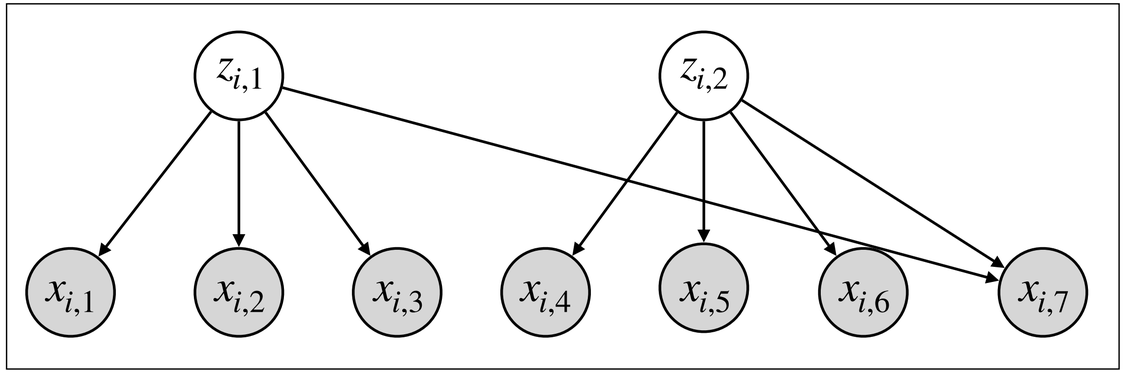


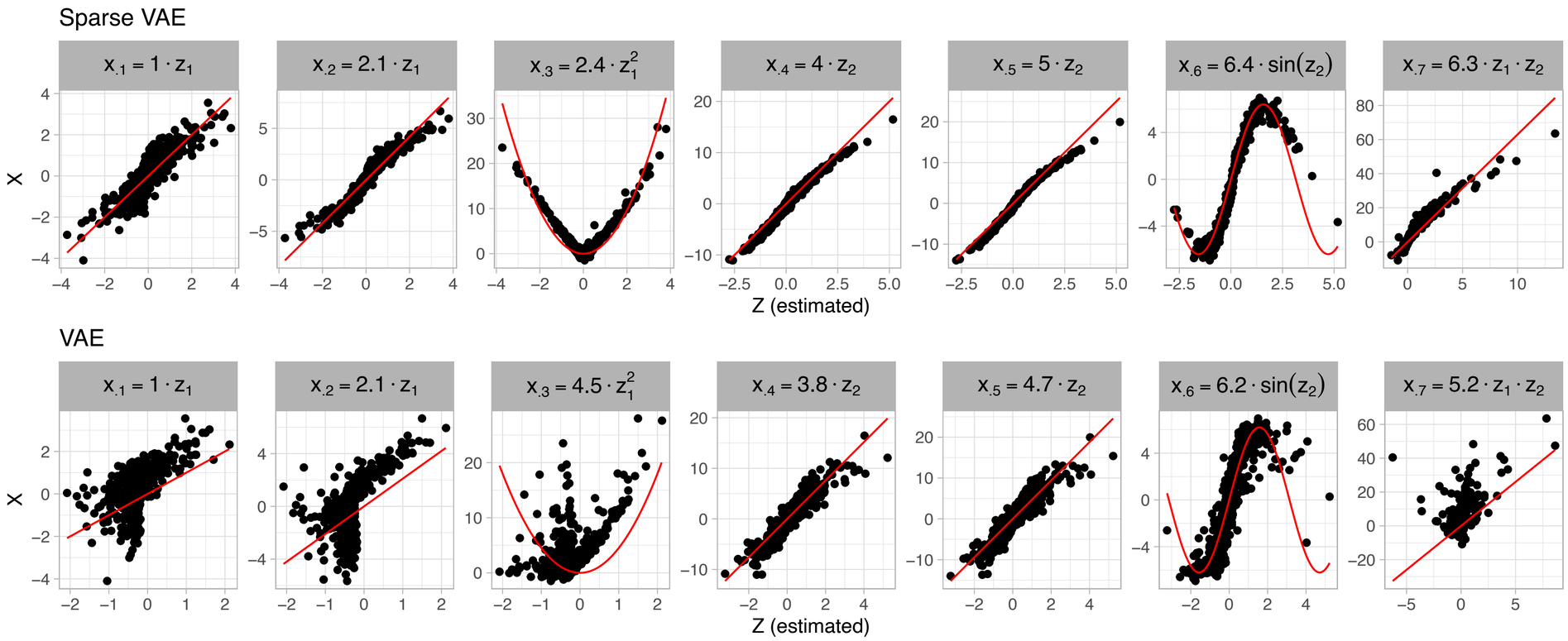


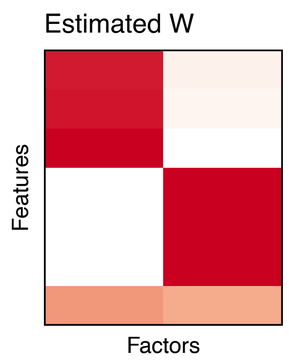
Recovering true factors

Synthetic Gaussian data

$N = 1000$ samples, $G = 7$ features, $K = 2$ factors

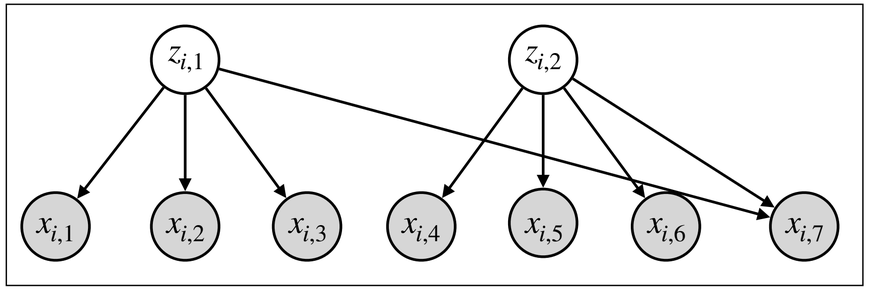






Sparse VAE successfully recovers true factors, VAE does not

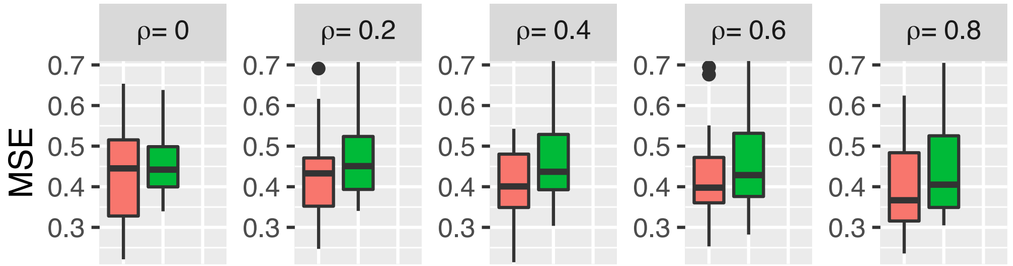
Recovering true factors



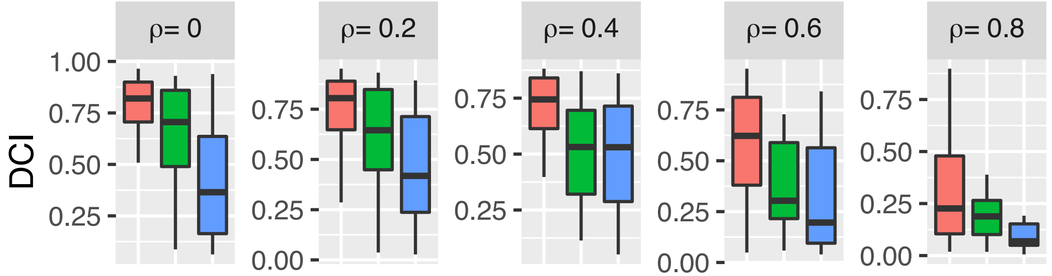
Synthetic Gaussian data

$N = 1000$ samples,
$G = 7$ features,
$K = 2$ factors

Increasing correlation of true factors



Heldout Mean Square Error (MSE) (lower is better)

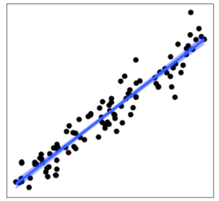


Method

- SparseVAE
- VAE
- VSC

Ground truth factor recovery (DCI disentanglement score) (higher is better)

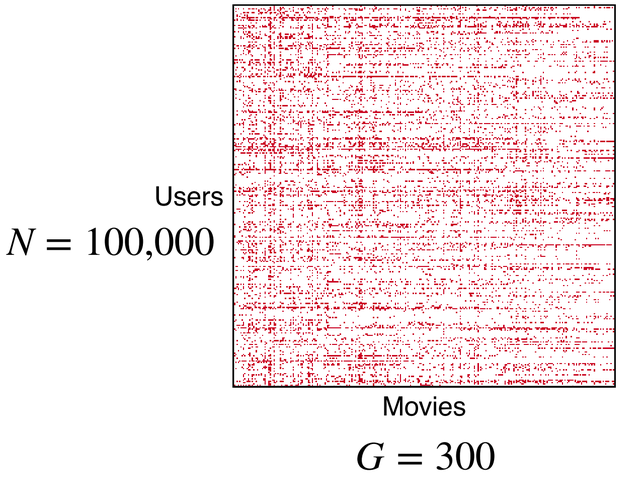
Prediction



MovieLens: ratings of movies by different users *[Harper and Konstan, 2015]*

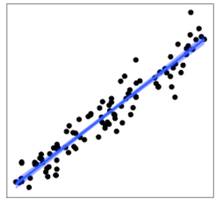
MovieLens

Method	Log loss	Recall@5	NDCG@10
Sparse VAE	**170.9 (2.1)**	**0.98 (0.002)**	**0.98 (0.003)**
VAE	175.9 (2.4)	0.97 (0.001)	0.96 (0.001)
β-VAE ($\beta = 2$)	178.2 (2.4)	0.95 (0.002)	0.93 (0.002)
VSC	192.2 (2.3)	0.79 (0.008)	0.77 (0.009)



NDCG: normalized discounted cumulative gain

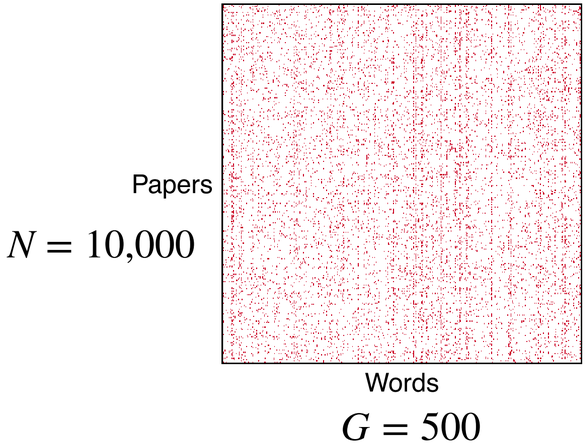
Prediction



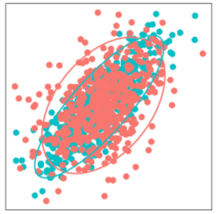
- **PeerRead:** word counts in paper abstracts *[Kang et al. 2018]*

PeerRead

Method	Log Loss
Sparse VAE	**245.0 (2.0)**
VAE	252.6 (1.4)
β-VAE ($\beta = 2$)	254.5 (3.0)
VSC	252.9 (2.0)



Domain adaptation

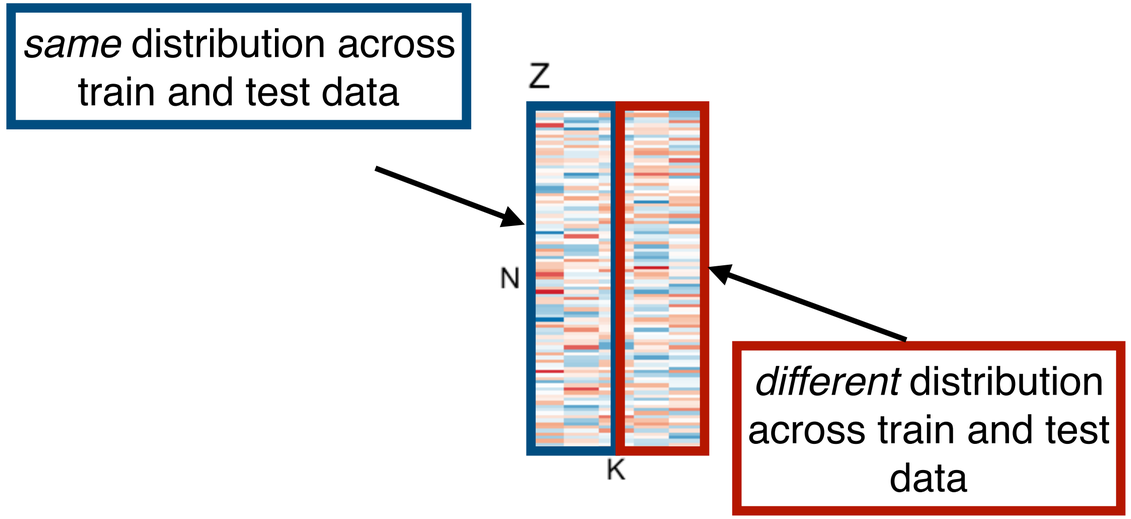


Data: semi-synthetic PeerRead ($N = 10{,}000,\ G = 500$)

- Generate factors with different distributions across train and test data
- Difficulty: correlation level of factors in training data

Average log loss on test data

	Difficulty		
Method	High	Medium	Low
Sparse VAE	**52.4 (0.4)**	**49.2 (0.3)**	**48.6 (0.1)**
VAE	54.6 (0.5)	52.3 (0.2)	50.8 (0.2)
β-VAE ($\beta = 2$)	54.7 (0.3)	52.1 (0.2)	51 (0.4)
VSC	58.7 (0.6)	56.1 (0.3)	55.4 (0.2)



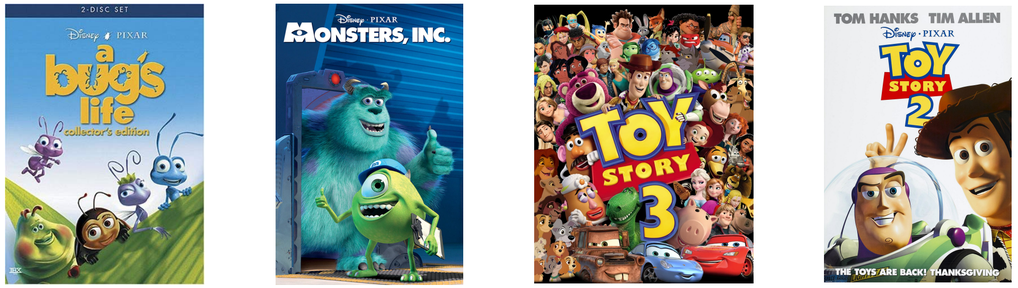
Interpretability

MovieLens: Sparse VAE finds meaningful factors via the matrix W

Sci-Fi Factor

The Fifth Element; Alien; Gattaca; Aliens

Pixar Factor



A Bug's Life; Monsters, Inc.; Toy Story 3 & 2

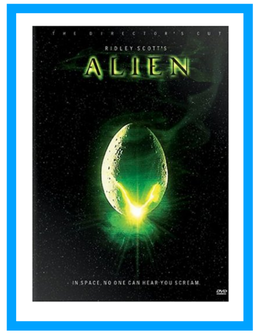
Action/Adventure Factor

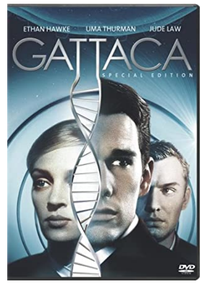
Star Wars V, IV & VI; Indiana Jones I

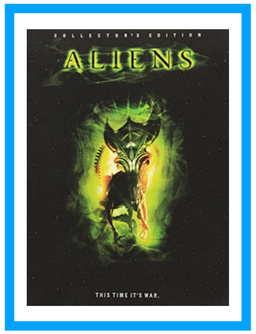
Interpretability

MovieLens: Sparse VAE finds meaningful factors via the matrix W

Sci-Fi Factor

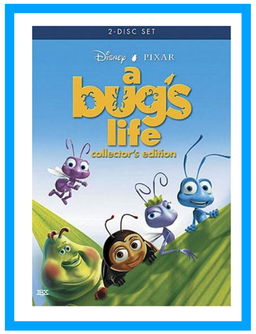






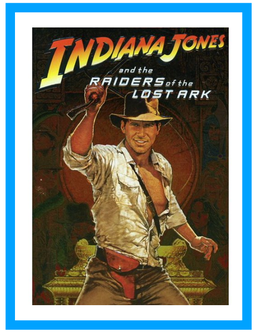
The Fifth Element; Alien; Gattaca; Aliens

Pixar Factor



A Bug's Life; Monsters, Inc.; Toy Story 3 & 2

Action/Adventure Factor



Star Wars V, IV & VI; Indiana Jones I

Anchor features

Downstream tasks

Single-cell RNA molecule counts in mouse cortex cells *[Zeisel et al. 2015]*

- $N = 3005$ cells, $G = 558$ genes
- Cells are from different regions of mouse cortex

Downstream tasks

Single-cell RNA molecule counts in mouse cortex cells *[Zeisel et al. 2015]*

- $N = 3005$ cells, $G = 558$ genes
- Cells are from different regions of mouse cortex

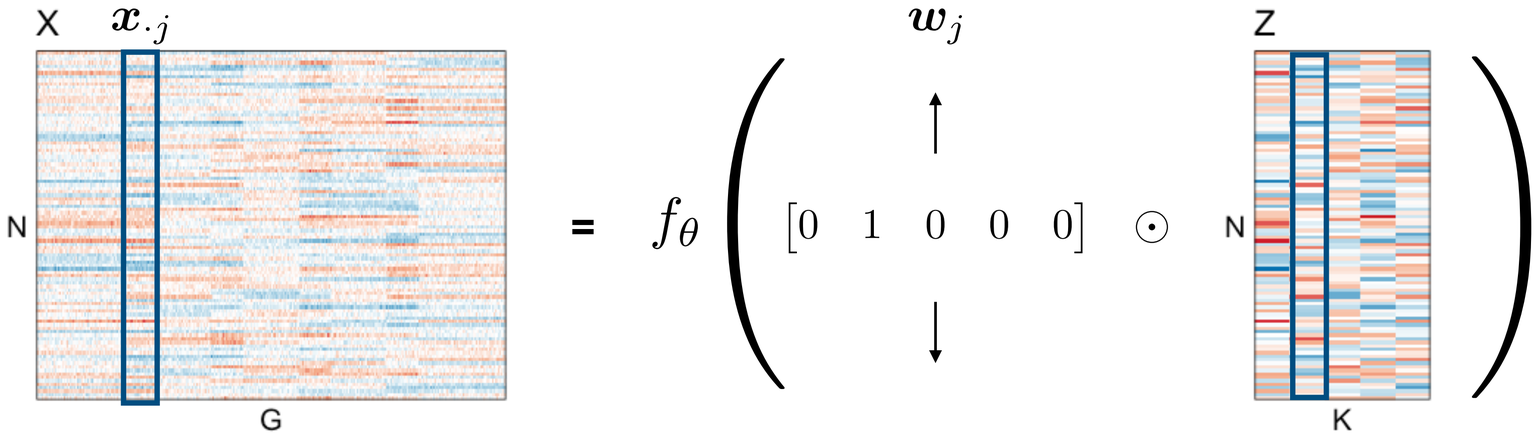
- Sparse VAE factors better predict cell label compared to other methods

Method	Sparse VAE	VAE	NMF	VSC
Accuracy	**0.95 (0.003)**	0.94 (0.016)	0.91 (0.007)	0.89 (0.062)

- Genes found by Sparse VAE were enriched for 39 biological processes

Future work

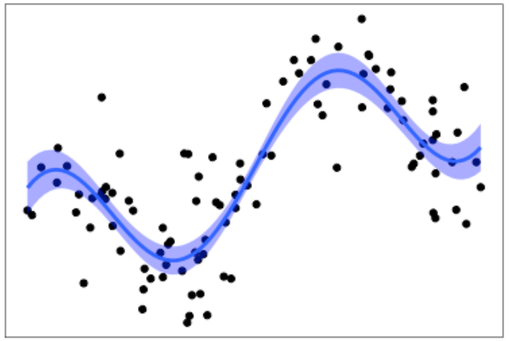
- Extend to image data
  - Sparse VAE requires each feature have a consistent meaning (pixels do not)



  - How to learn image "features" and retain identifiability?

Conclusion

We developed the **Sparse VAE** for representation learning:



Flexible

Interpretable

Identifiable

Moran, Sridhar, Wang and Blei

Identifiable Deep Generative Models via Sparse Decoding

[arXiv:2110.10804]

GitHub

gemoran/sparse-vae-code

References

- Ainsworth, S. K., Foti, N. J., Lee, A. K., and Fox, E. B. (2018). *oi-VAE: Output interpretable VAEs for nonlinear group factor analysis.* ICML
- Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013). *A practical algorithm for topic modeling with provable guarantees.* ICML
- Bing, X., Bunea, F., Ning, Y., Wegkamp, M., et al. (2020). *Adaptive estimation in structured factor models with applications to overlapping clustering.* Annals of Statistics, 48(4):2055–2081.
- Donoho, D. L. and Stodden, V. (2003). *When does non-negative matrix factorization give a correct decomposition into parts?* NeurIPS
- Gresele, L., Von Kügelgen, J., Stimper, V., Schölkopf, B. and Besserve, M. (2021). *Independent mechanism analysis, a new concept?*. NeurIPS
- Hälvä, H., Corff, S. L., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvarinen, A. (2021). *Disentangling identifiable features from noisy data with structured nonlinear ICA*. arXiv preprint arXiv:2106.09620.
- Harper, F. M. and Konstan, J. A. (2015). *The MovieLens Datasets: History and Context.* ACM TiiS, 5(4):1–19.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). *beta-vae: Learning basic visual concepts with a constrained variational framework.* ICLR
- Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., and Schwartz, R. (2018). *A dataset of peer reviews (PeerRead): Collection, insights and NLP applications.* arXiv preprint arXiv:1804.09635.

References (continued)

- Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). *Variational autoencoders and nonlinear ICA: A unifying framework.* AISTATS.
- Kingma, D. P. and Welling, M. (2014). *Auto-encoding variational Bayes.* ICLR
- Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., & Gavves, S. (2022). *CITRIS: Causal Identifiability from Temporal Intervened Sequences*. ICML
- Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). *Weakly supervised disentanglement without compromises*. ICML
- Ročková, V. and George, E. I. (2018). *The spike-and-slab lasso.* Journal of the American Statistical Association, 113(521):431–444.
- Rohe, K. and Zeng, M. (2020). *Vintage factor analysis with varimax performs statistical inference.* arXiv preprint arXiv:2004.05387.
- Tonolini, F., Jensen, B. S., and Murray-Smith, R. (2020). *Variational sparse coding.* UAI
- Van der Maaten, L. and Hinton, G. (2008). *Visualizing data using t-SNE.* Journal of Machine Learning Research, 9(11).
- von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). *Self-supervised learning with data augmentations provably isolates content from style.* arXiv preprint arXiv:2106.04619.
- Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). *Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, 347(6226):1138–1142.
- Zheng, Y., Ng, I., & Zhang, K. (2022). *On the Identifiability of Nonlinear ICA: Sparsity and Beyond.* arXiv preprint arXiv:2206.07751.

Sparse VAE algorithm

- MAP objective function:

$$\mathcal{L}(\theta, \phi, \boldsymbol{W}, \boldsymbol{\eta}) = \sum_{i=1}^{N} \left\{ \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)} [\log p_\theta(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W})] - D_{KL}(q_\phi(\boldsymbol{z}_i|\boldsymbol{x}_i)||p(\boldsymbol{z}_i)) \right\} \tag{9}$$
$$+ \mathbb{E}_{\boldsymbol{\Gamma}|\boldsymbol{W}^{(t)}, \boldsymbol{\eta}^{(t)}} \left[\log[p(\boldsymbol{W}|\boldsymbol{\Gamma}) p(\boldsymbol{\Gamma}|\boldsymbol{\eta}) p(\boldsymbol{\eta})] \right],$$

Algorithm 1: The Sparse VAE

input: data $\boldsymbol{X}$, hyperparameters $\lambda_0, \lambda_1, a, b, \boldsymbol{C}$
output: factor distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, selector matrix $\boldsymbol{W}$, parameters θ
while *not converged* **do**
For $j = 1, \ldots, G;\ k = 1, \ldots, K$, update:

$$\mathbb{E}\left[\gamma_{jk}|w_{jk}, \eta_k\right] = [1 + (1 - \eta_k)/\eta_k \psi_0(w_{jk})/\psi_1(w_{jk})]^{-1}.$$

For $k = 1, \ldots, K$, update:

$$\eta_k = \left(\sum_{j=1}^{G} \mathbb{E}\left[\gamma_{jk}|w_{jk}, \eta_k\right] + a - 1 \right) \Big/ (a + b + G - 2).$$

Update $\theta, \phi, \boldsymbol{W}$ with stochastic gradient ascent according to Eq. 9.
end