Medical data wrangling with sequential variational autoencoders

Daniel Barrejón, Pablo M. Olmos and Antonio Artés-Rodríguez

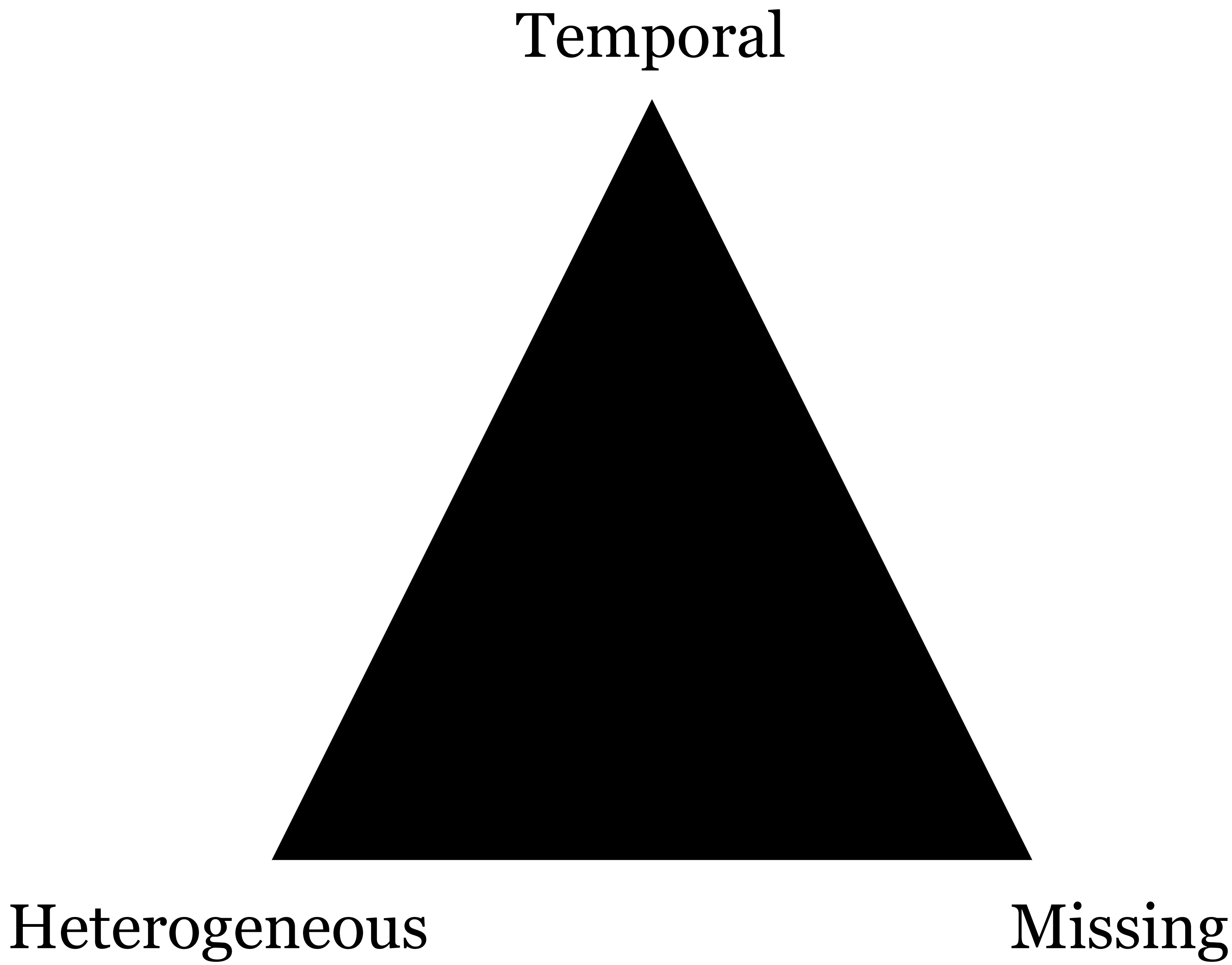
uc3m | Universidad Carlos III de Madrid

4th October 2021

Outline

1. Main problem
2. Proposed Model: Shi-VAE
   - Generative Model: Heterogeneous Decoder
   - Variational Inference: Discrete and continuous latent spaces
   - SOTA Comparison: GP-VAE
3. Evaluation Metrics
   - Error Metrics
   - Cross Correlation
4. Datasets
5. Results
6. Discussion

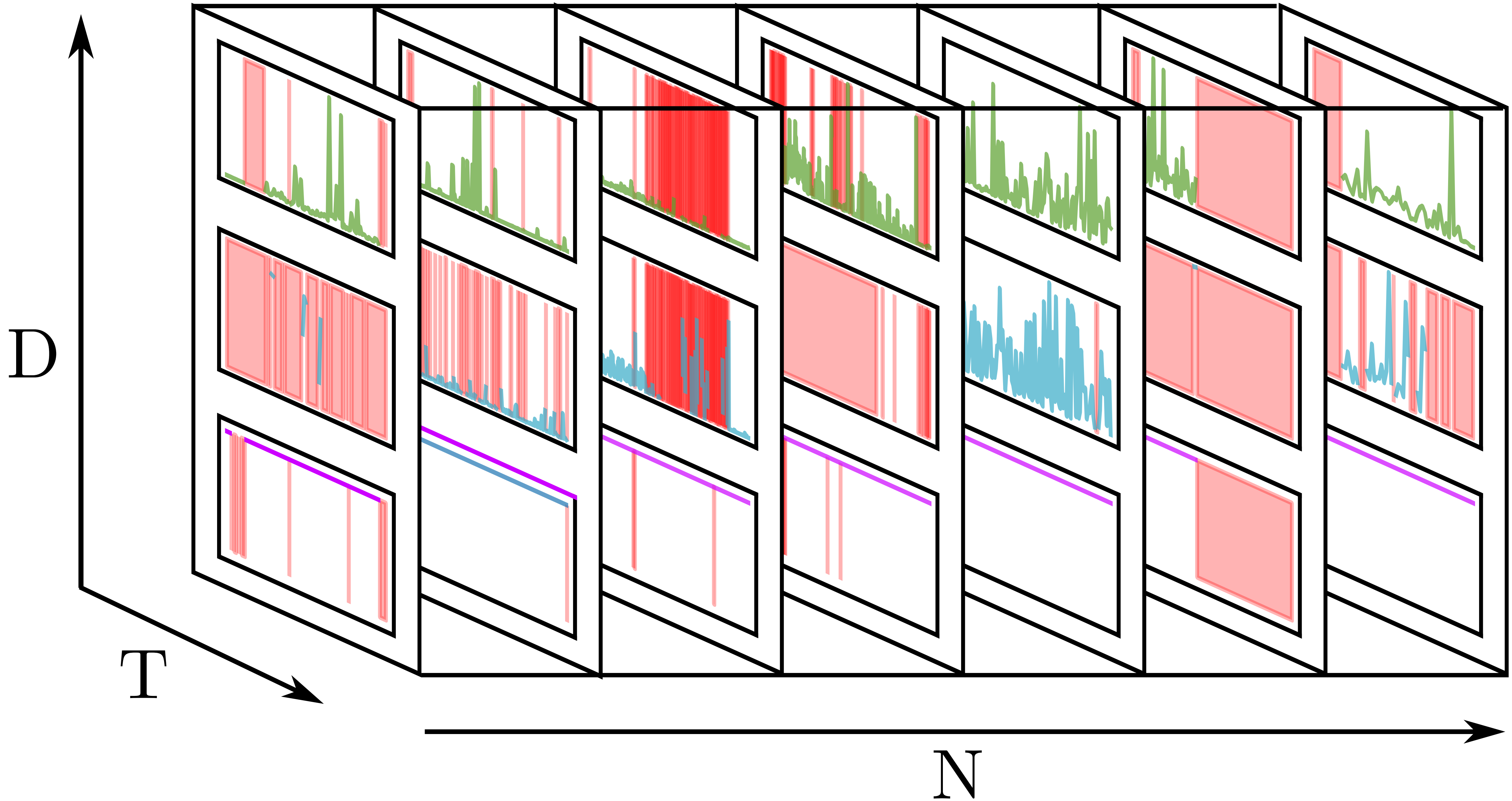
1 Main problem



1

Main problem

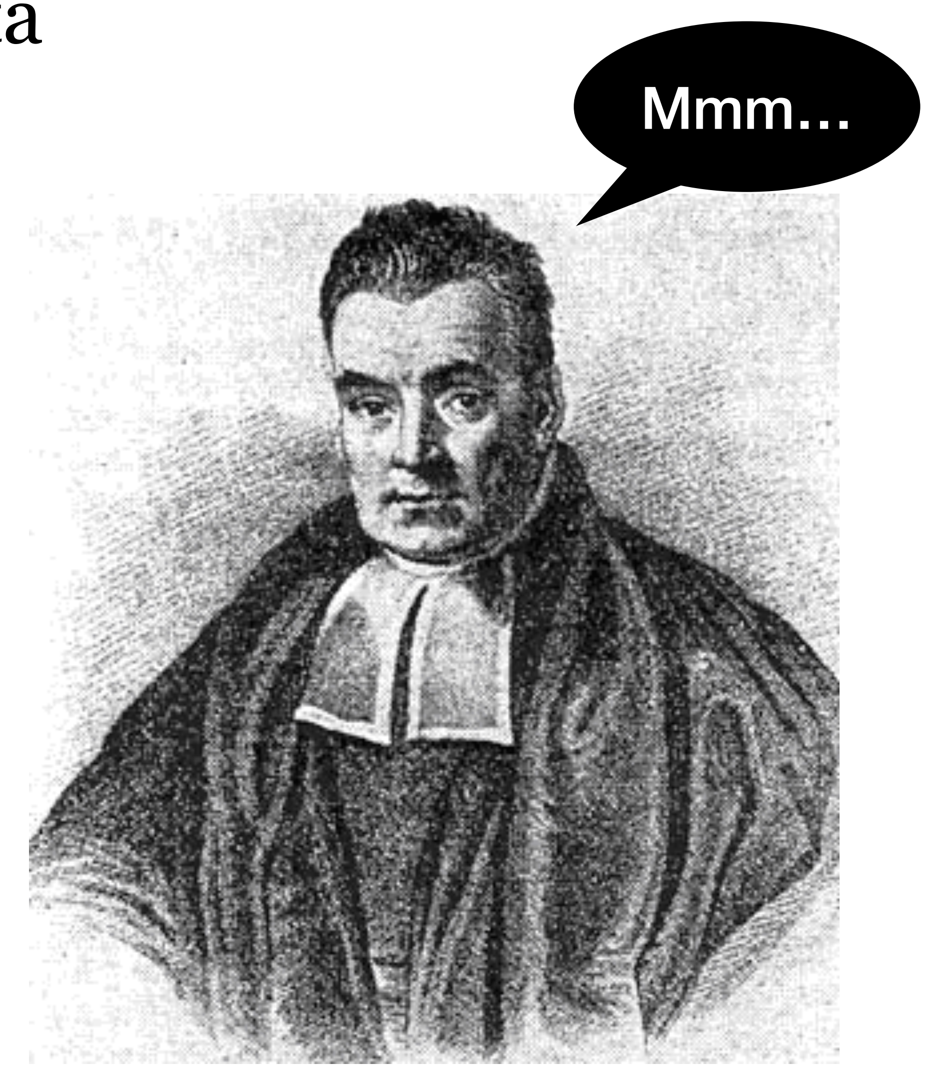
1 # Main problem



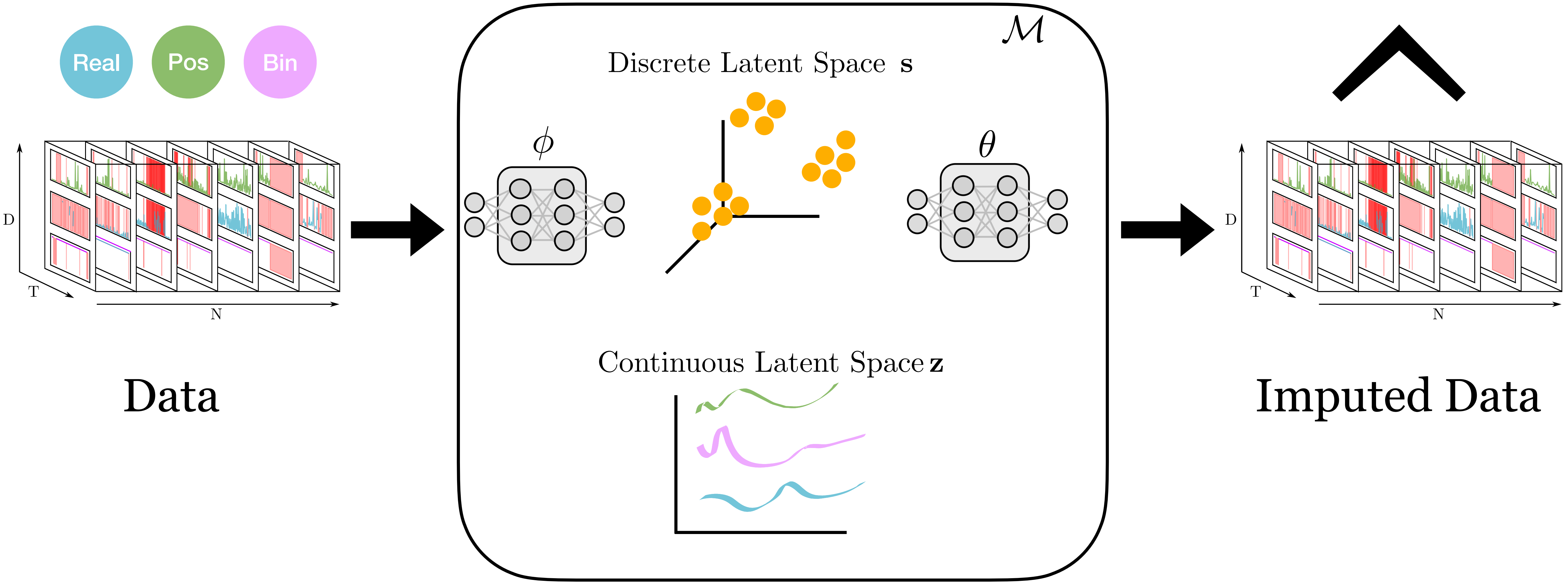
❶ Main problem

Problems in this setup

- Heterogeneous Likelihoods: Penalization
- Sequential Data
- Missing Data
- Noisy Data



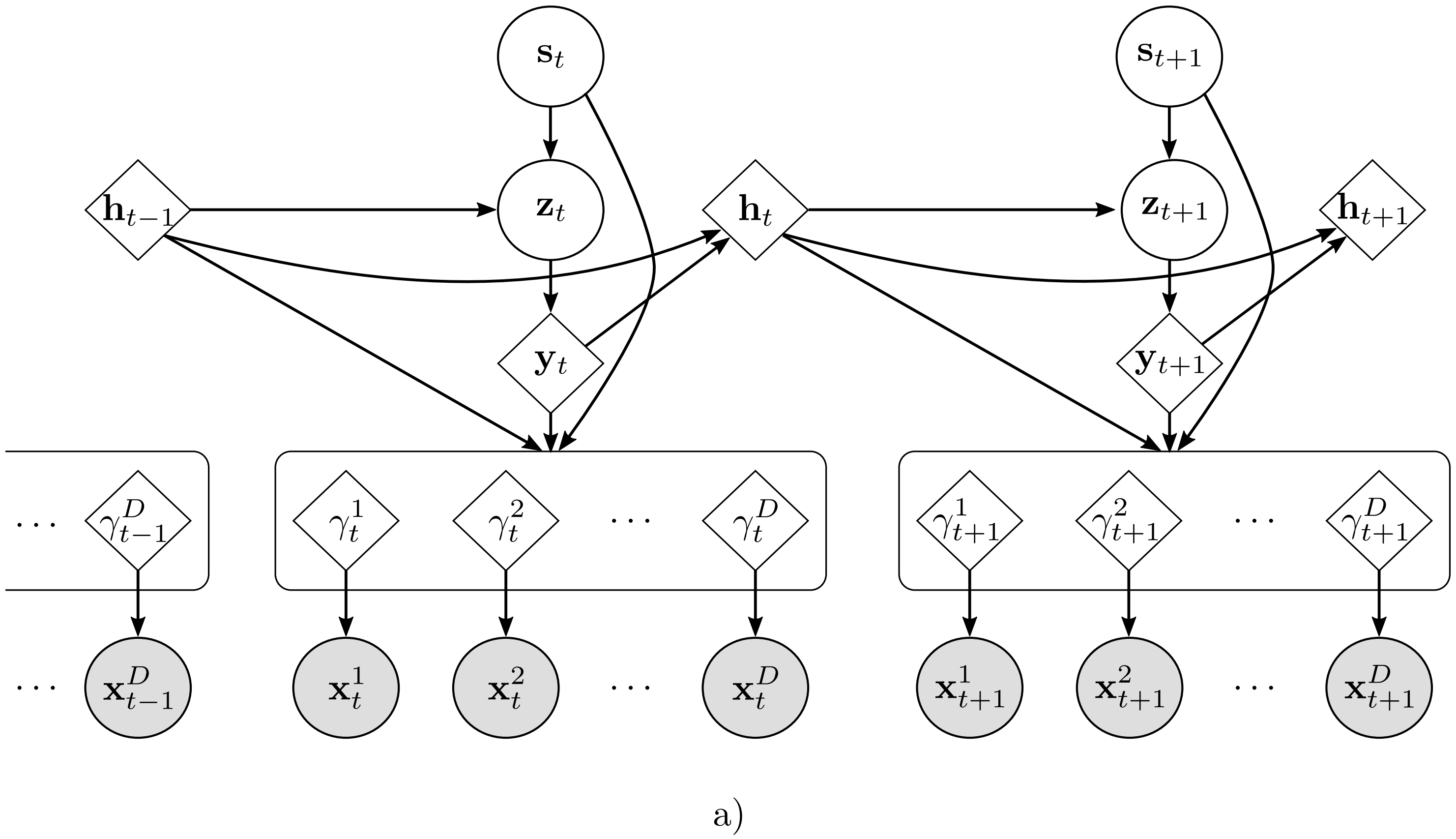
2 Shi-VAE



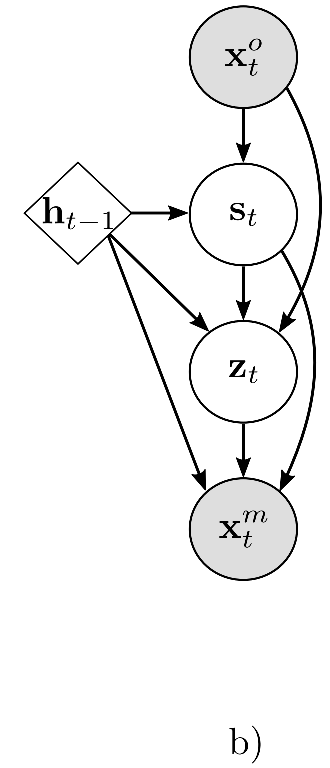
Real
Pos
Bin
D
T
N
Data
$\mathcal{M}$
Discrete Latent Space $\mathbf{s}$
ϕ
θ
Continuous Latent Space $\mathbf{z}$
D
T
N
Imputed Data

Shi-VAE: Generative Model

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{S}) = \prod_{t=1}^{T} p_{\theta_x}(\mathbf{x}_t | \mathbf{z}_{\leq t}, \mathbf{s}_t) p_{\theta_z}(\mathbf{z}_t | \mathbf{z}_{<t}, \mathbf{s}_t) p_{\theta_s}(\mathbf{s}_t)$$



Generative Model



Inference Model

2 Shi-VAE: Generative Model

Heterogeneous Likelihood

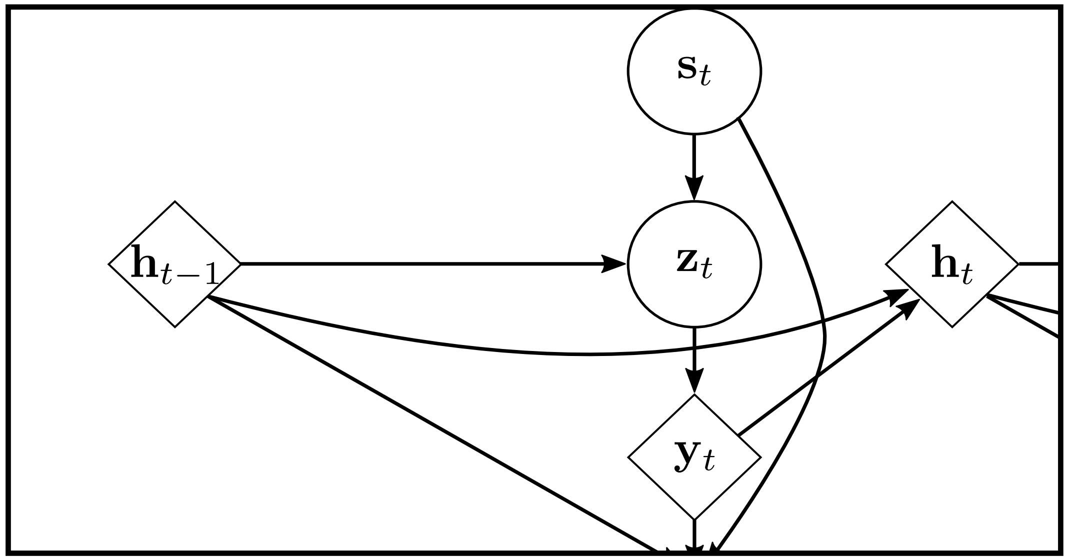
$$p(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{s}_t) = \underbrace{\prod_{d \in \mathcal{O}_t} p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t)}_{\text{Observations}} \underbrace{\prod_{d \in \mathcal{M}_t} p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t)}_{\text{Missing}}$$

$\mathbf{x}_t$ →

- Real: $\mathbf{x}_{td} \in \mathbb{R}$
- Pos: $\mathbf{x}_{td} \in \mathbb{R}^+$
- Bin: $\mathbf{x}_{td} \in [0, 1]$
- Cat: $\mathbf{x}_{td} \in$ e.g. $[-1, 0, 1]$, [negative, neutral, negative]

HIVAE (Alfredo Nazabal et al. 2018)

2 Shi-VAE: Generative Model



- Temporal Dependency encoded in $\mathbf{z}_t$

$$p(\mathbf{z}_t | \mathbf{z}_{<t}, \mathbf{s}_t) = \mathcal{N}(\mathbf{z}_t | \boldsymbol{\mu}_{0,t}, \boldsymbol{\Sigma}_{0,t})$$

Where $[\boldsymbol{\mu}_{0,t}, \boldsymbol{\Sigma}_{0,t}] = \varphi_\omega^{\text{prior}}(\mathbf{h}_{t-1}, \mathbf{s}_t)$

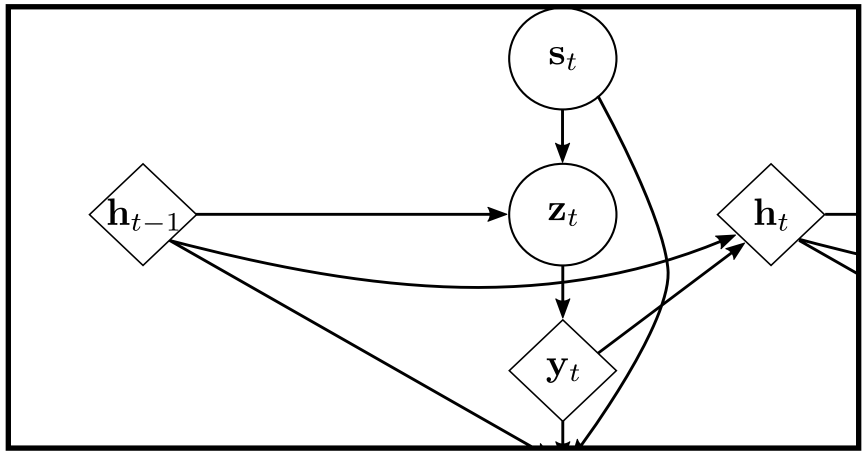
Temporality is encapsulated in $\mathbf{h}_{t-1}$: RNN Hidden State

$$\mathbf{h}_{t-1} = f_{\boldsymbol{\tau}}(\mathbf{y}_{t-1}, \mathbf{h}_{t-2}),$$

Where $\mathbf{y}_{t-1} = \varphi_\omega^{\mathbf{z}}(\mathbf{z}_{t-1})$

VRNN (J Chung et al 2015)

2 Shi-VAE: Generative Model



- Time-independent Discrete Latent $\mathbf{s}_t$

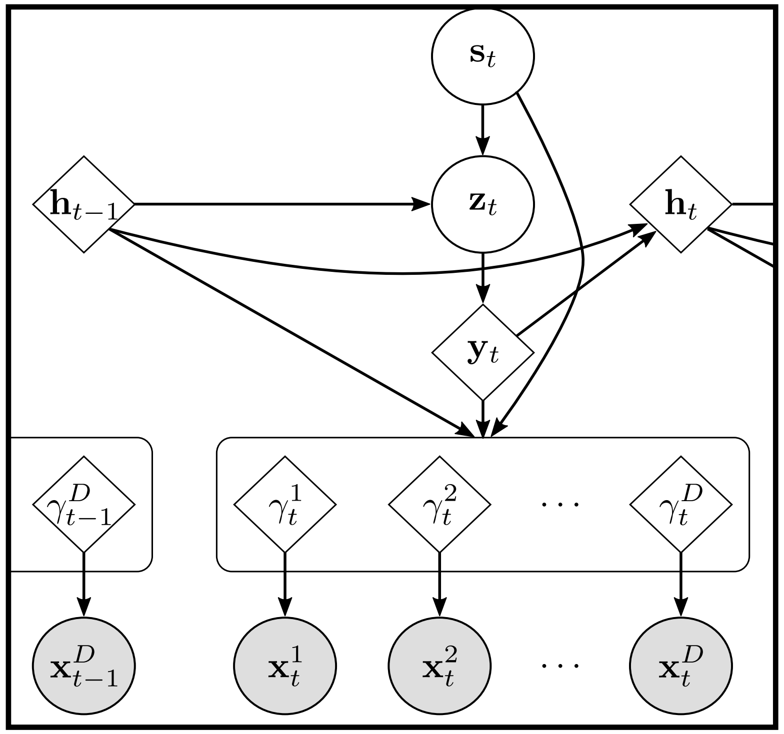
$$p(\mathbf{s}_t) = \text{Categorical}(\mathbf{s}_t|\boldsymbol{\pi}),$$

Where $\pi_k = \frac{1}{L}$, with L being the number of components

GMVAE (Dilokthanakul et al. 2016)

Shi-VAE: Generative Model

Heterogeneous Decoder



$$\boxed{p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t) = p(x_{td}|\gamma_t^d)}$$

Real

$$p(x_{td}|\gamma_t^d) = \mathcal{N}(\mu_{x,t}^d, \sigma_{x,t}^{2,d}),$$
$$\text{where } [\mu_{x,t}^d, \sigma_{x,t}^{2,d}] = \varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1})$$

Pos

$$p(x_{td}|\gamma_t^d) = \log \mathcal{N}(\mu_{x,t}^d, \sigma_{x,t}^{2,d}),$$
$$\text{where } [\mu_{x,t}^d, \sigma_{x,t}^{2,d}] = \varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1})$$

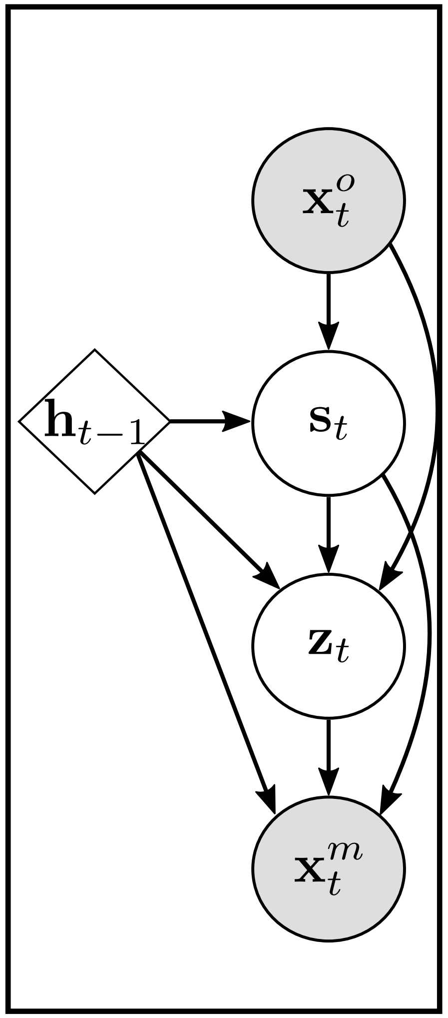
Bin

$$p(x_{td}|\gamma_t^d) = Be(p_{x,t}^d),$$
$$\text{where } p_{x,t}^d = \sigma\big(\varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1})\big),$$

Cat

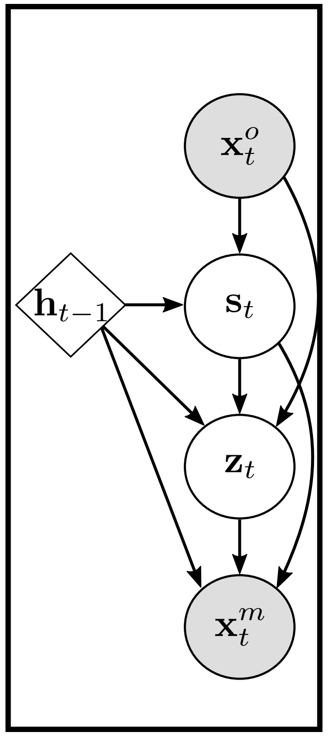
$$\log p(x_{td} = c|\gamma_t^d) = \varphi_{\omega,d}^{\text{dec}}(\mathbf{y}_t, \mathbf{s}_t, \mathbf{h}_{t-1})|_c$$

Variational Inference



$$q_\phi(\mathbf{x}^m_{\leq T}, \mathbf{z}_{\leq T}, \mathbf{s}_{\leq T} | \mathbf{x}^o_{\leq T}) = \prod_{t=1}^{T} q_{\phi_z}(\mathbf{z}_t | \mathbf{z}_{<t}, \mathbf{s}_t, \mathbf{x}^o_t)$$
$$q_{\phi_s}(\mathbf{s}_t | \mathbf{x}^o_t, \mathbf{z}_{<t})$$
$$p(\mathbf{x}^m_t | \mathbf{z}_{\leq t}, \mathbf{s}_t).$$

2 Shi-VAE: Inference Model



- Variational Family on $\mathbf{z}_T$

$$q_{\phi_z}(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t, \mathbf{x}_t^o) = \mathcal{N}(\boldsymbol{\mu}_{z,t}, \boldsymbol{\Sigma}_{z,t}),$$
$$\text{where } [\boldsymbol{\mu}_{z,t}, \boldsymbol{\Sigma}_{z,t}] = \varphi_\omega^{\text{enc}}(\varphi_\omega^{\mathbf{x}}(\tilde{\mathbf{x}}_t), \mathbf{h}_{t-1}, \mathbf{s}_t).$$

$\tilde{\mathbf{x}}_t$ denotes a D-dimensional vector with missing entries filled with zeros.

- Variational Family on $\mathbf{s}_t$

$$q_{\phi_s}(\mathbf{s}_t|\mathbf{x}_t^o, \mathbf{z}_{<t}) = \text{Categorical}(\boldsymbol{\pi}(\varphi_\omega^{\mathbf{s}}(\tilde{\mathbf{x}}_t, \mathbf{h}_{t-1}))),$$

Shi-VAE: Inference Model

ELBO

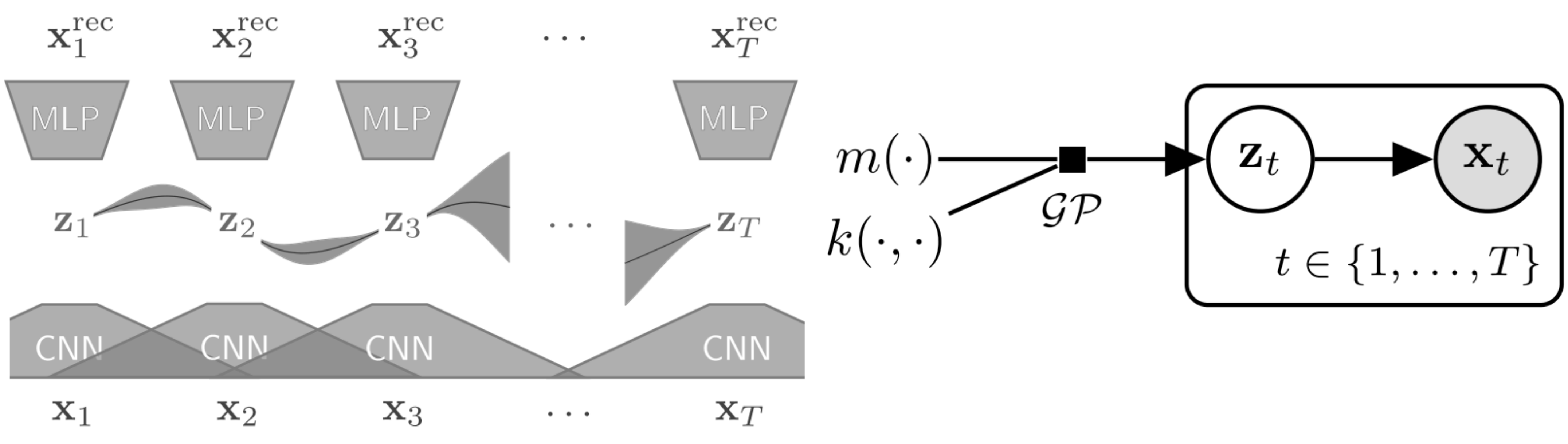
$$\log p(\mathbf{X}^o) \geq \int q(\mathbf{X}^o, \mathbf{X}^m, \mathbf{Z}, \mathbf{S}) \log \frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{S})}{q(\mathbf{X}^o, \mathbf{X}^m, \mathbf{Z}, \mathbf{S})} d\mathbf{Z} d\mathbf{S} d\mathbf{X}^m$$

$$\log p(\mathbf{X}^o) \geq \sum_{t=1}^{T} \Bigg[\underbrace{\mathop{\mathbb{E}}_{\substack{q(\mathbf{s}_t|\mathbf{x}_t^o, \mathbf{z}_{<t},) \\ q(\mathbf{z}_t|\mathbf{x}_t^o, \mathbf{z}_{<t}, \mathbf{s}_t)}} \big[\log p(\mathbf{x}_t^o|\mathbf{z}_{\leq t}, \mathbf{s}_t)\big]}_{\text{Reconstruction}}$$

$$\underbrace{- \mathop{\mathbb{E}}_{q(\mathbf{s}_t|\mathbf{x}_t^o, \mathbf{z}_{<t})} \big[\beta \mathrm{KL}\big(q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{x}_t^o, \mathbf{s}_t) || p(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t)\big)\big] - \beta \mathrm{KL}\big(q(\mathbf{s}_t|\mathbf{x}_t^o, \mathbf{z}_{<t},) || p(\mathbf{s}_t)\big)}_{\text{Regularization}} \Bigg]$$

Shi-VAE: GP-VAE

SOTA Model: GP-VAE



(a) Architecture sketch (b) Graphical model

Complexity! $\mathcal{O}(T^3)$

GP-VAE (Fortuin et al. 2016)

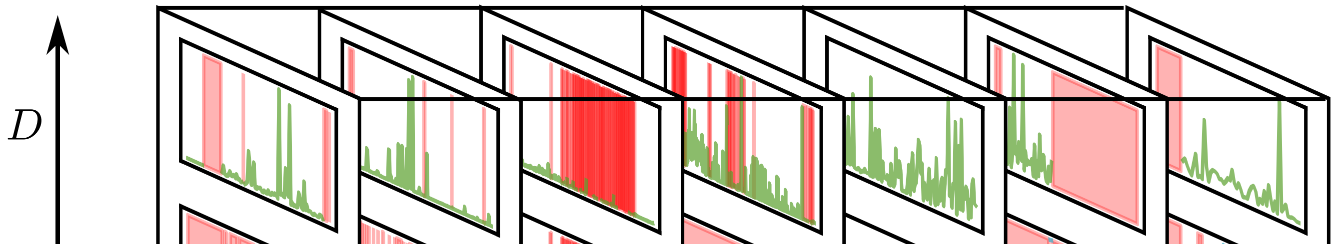
Error Metrics

Continuous Data $$err(d) = \frac{\sqrt{1/N_d \sum_n \sum_t (x_{td}^n - \hat{x}_{td}^n)^2}}{\max(\mathbf{X}_d) - \min(\mathbf{X}_d)}$$

Discrete Data $$err(d) = \frac{1}{N_d} \sum_n \sum_t I(x_{td}^n \neq \hat{x}_{td}^n)$$

$$\text{Error} = 1/D \sum_d err(d)$$

3 Evaluation Metrics



- Cross Correlation

$\mathbf{X}_d$ true d-portion from dataset
$\hat{\mathbf{X}}_d$ imputed d-portion from dataset
N_d # missing entries at D

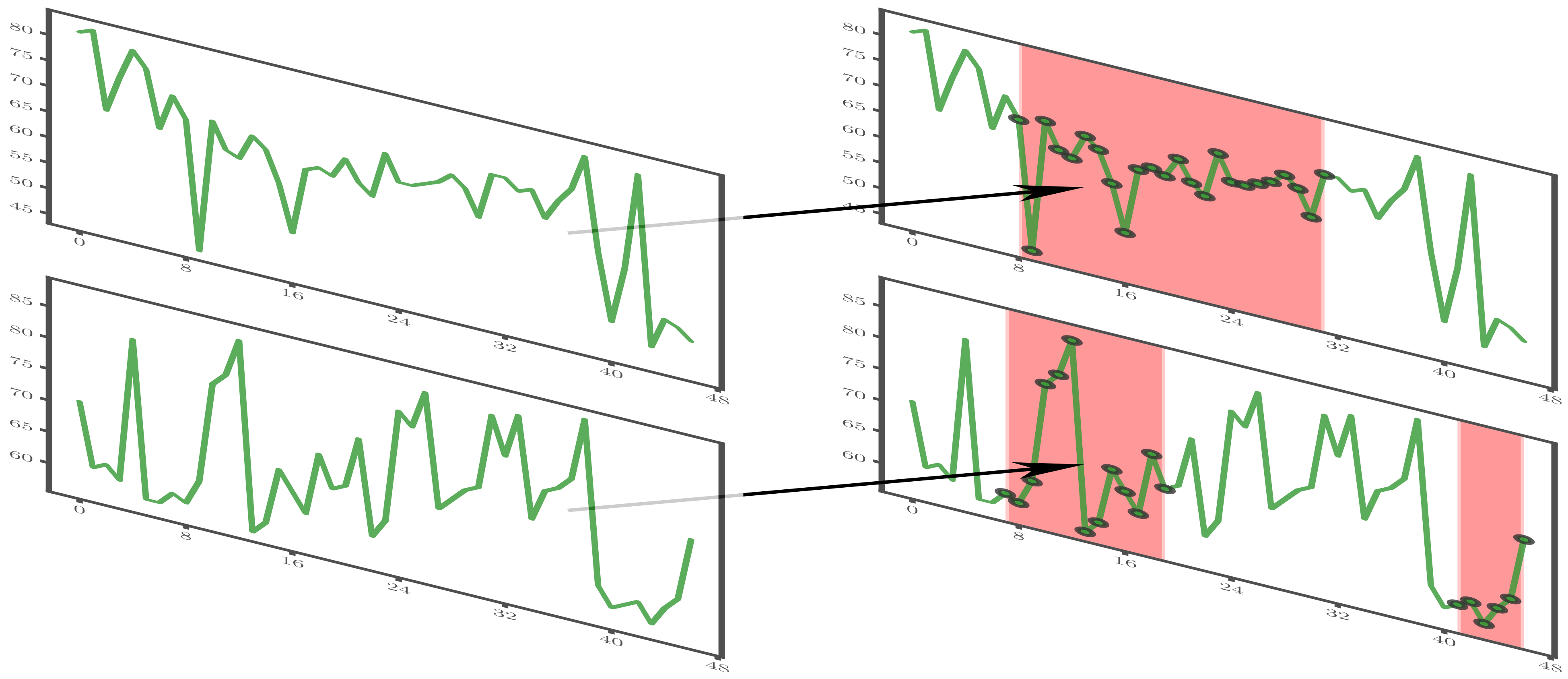
$$c(\mathbf{w}, \hat{\mathbf{w}}) = \max[(\mathbf{w} - \mu_{\mathbf{w}}) \star (\hat{\mathbf{w}} - \mu_{\hat{\mathbf{w}}})]$$

$\mathbf{w}$ true value of missing burst

$\hat{\mathbf{w}}$ imputed value of missing burst

$$\phi(d) = \frac{\sum_{\mathbf{w}, \hat{\mathbf{w}} \in \mathbf{X}_d} c(\mathbf{w}, \hat{\mathbf{w}})}{N_d} \qquad \text{Cross. Corr} = 1/D \sum_d \phi(d)$$

- Missing bursts generation



4 Datasets

Dataset	Dimension Dataset **D**	Dimension **z**	Dimension **s**
Synthetic	4	2	3
Physionet	35	35	10
Human Monitoring	7	5	3

Synthetic Dataset

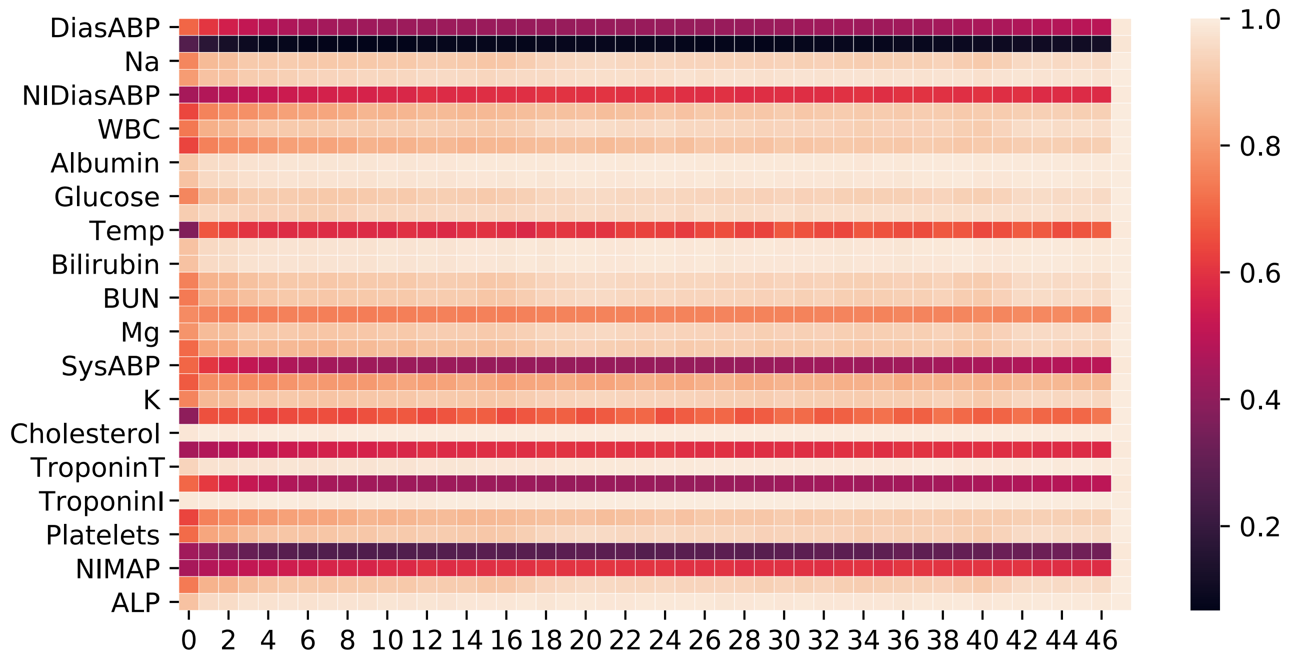
1000 samples from HMM

Real Pos Bin Cat

10 artificial masks: 10%, 30%, 50% missing rates
T=100

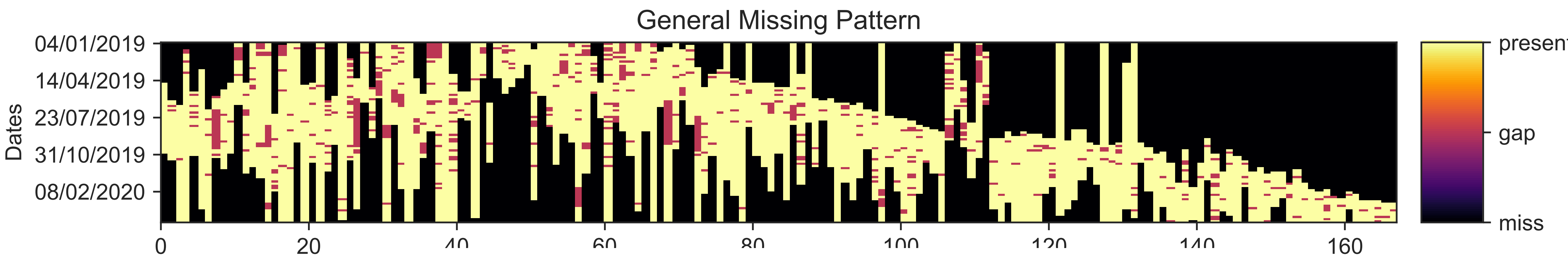
Dataset	Dimension Dataset **D**	Dimension **z**	Dimension **s**
Synthetic	4	2	3
Physionet	35	35	10
Human Monitoring	7	5	3

Physionet



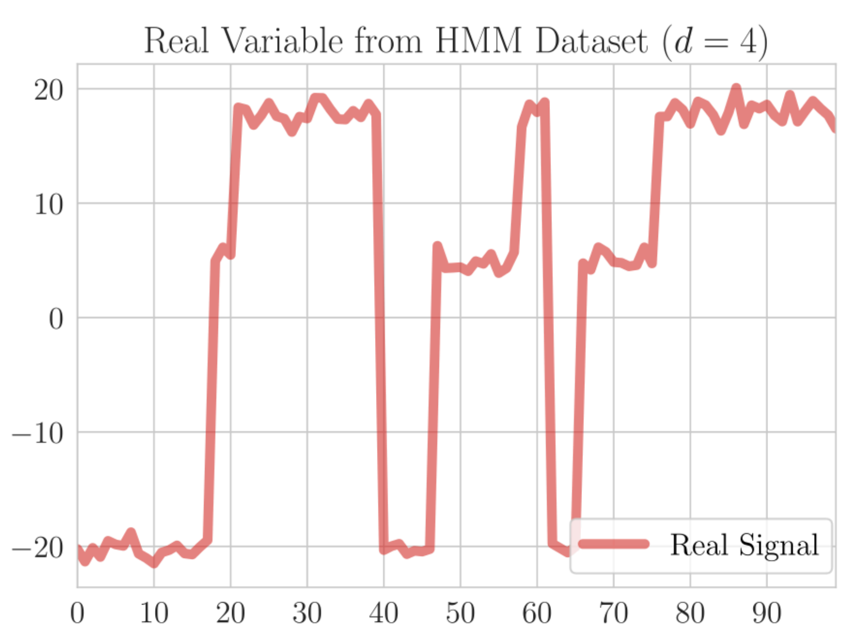
D=35
T=48 (2 days at ICU)
10% of artificial missing

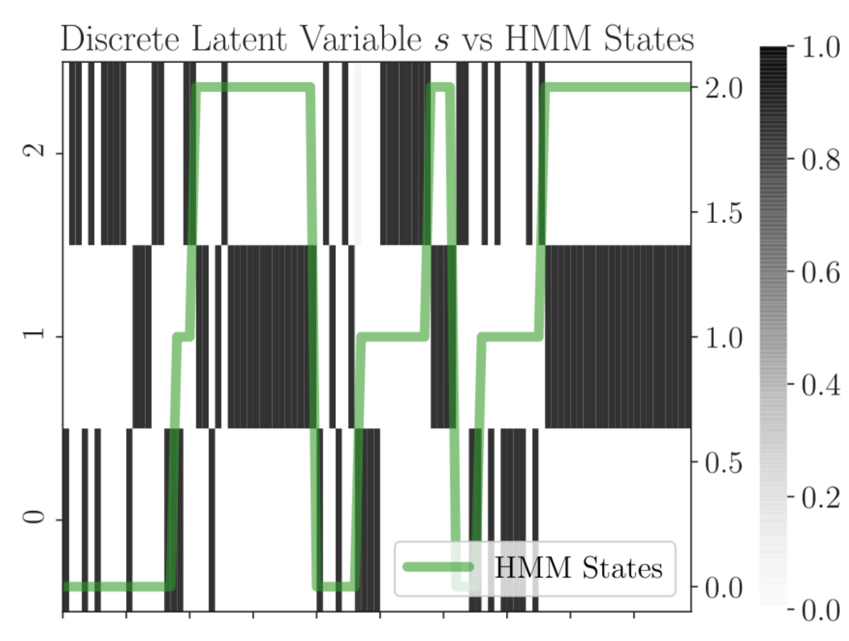
Human Monitoring Database (eb2)

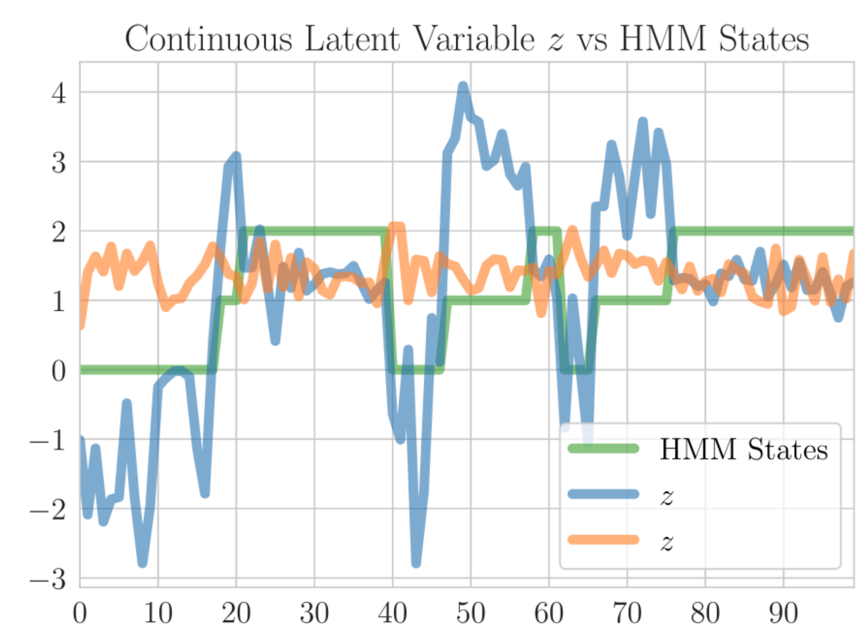


Variable	Type	Missing Percentage [%]
Distance	Positive	42
Steps Home	Binary	66
Steps Total	Positive	22
App Usage	Positive	38
Sport	Binary	62
Sleep	Positive	31
Vehicle	Positive	44

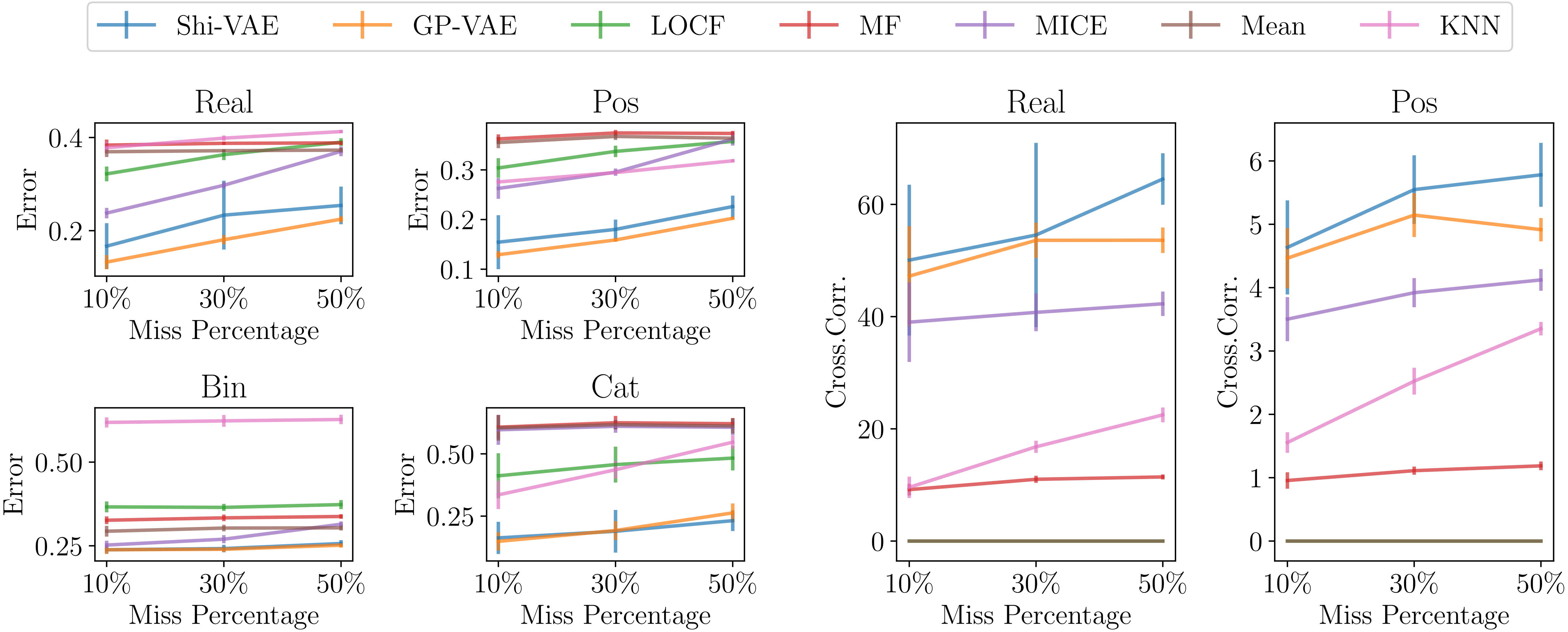
Synthetic Dataset







Synthetic Dataset



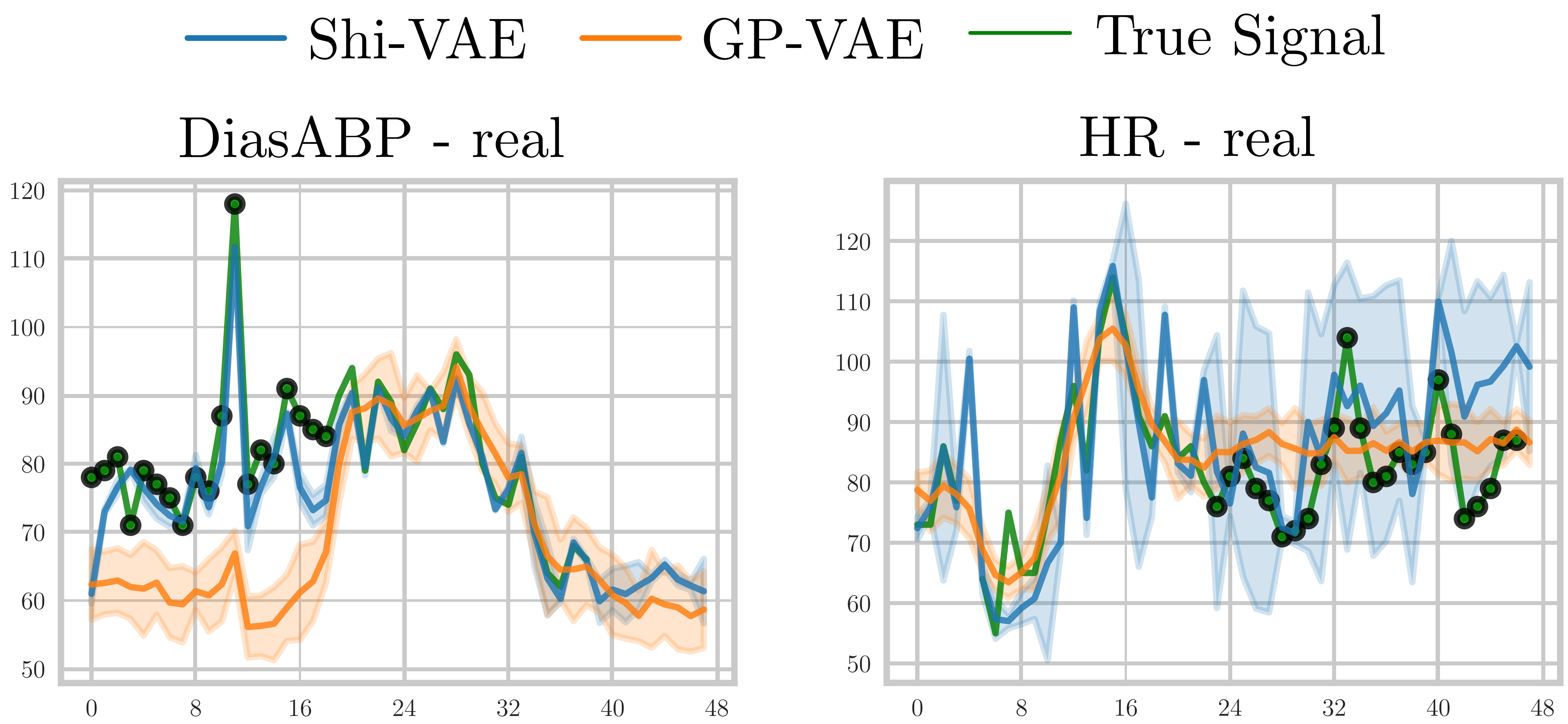
Physionet

Model	Avg. Error	Cross. Corr
Shi-VAE	**0.064 ± 0.003**	**38.061 ± 5.000**
GP-VAE	**0.060 ± 0.002**	31.414 ± 1.016

Human Monitoring Database

Variable	Model	Error	Cross Correlation
Average	Shi-VAE	**0.200 ± 0.038**	**0.369± 0.140**
	GP-VAE	**0.184 ± 0.022**	0.157 ± 0.031
Distance	Shi-VAE	**0.201 ± 0.012**	**0.783 ± 0.249**
	GP-VAE	**0.205 ± 0.014**	0.389 ± 0.092
Steps home	Shi-VAE	**0.170 ± 0.054**	**0.010 ± 0.009**
	GP-VAE	**0.151 ± 0.016**	**0.011 ± 0.009**
Steps total	Shi-VAE	**0.269 ± 0.046**	**0.444 ± 0.181**
	GP-VAE	**0.268 ± 0.044**	0.205 ± 0.038
App usage	Shi-VAE	**0.113 ± 0.014**	**0.088 ± 0.045**
	GP-VAE	**0.115 ± 0.013**	**0.039 ± 0.008**
Sport	Shi-VAE	**0.216 ± 0.086**	**0.013 ± 0.005**
	GP-VAE	**0.121 ± 0.030**	**0.009 ± 0.004**
Sleep	Shi-VAE	**0.063 ± 0.010**	**0.034 ± 0.016**
	GP-VAE	**0.059 ± 0.010**	0.013 ± 0.003
Vehicle	Shi-VAE	**0.372 ± 0.043**	**1.215 ± 0.477**
	GP-VAE	**0.370 ± 0.028**	0.436 ± 0.064

Physionet

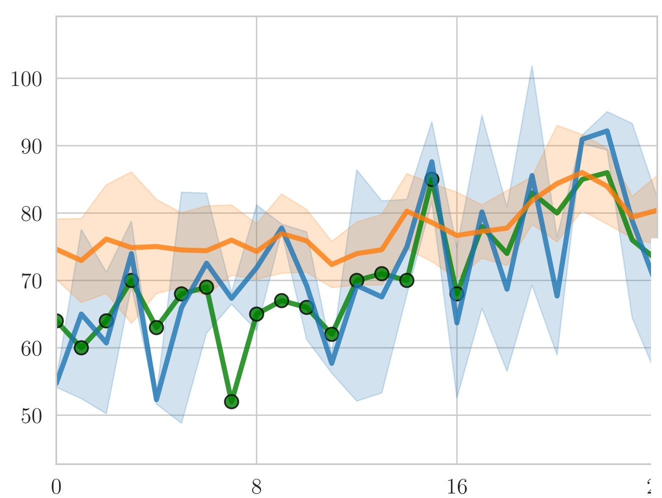

Shi-VAE
GP-VAE
True Signal
DiasABP - real
HR - real
120
110
100
90
80
70
60
50
0
8
16
24
32
40
48

5 Results

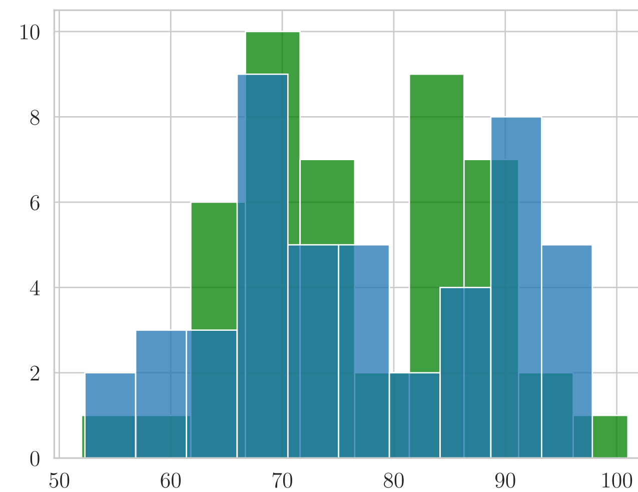
Physionet

True Signal
Shi-VAE
GP-VAE

NIMAP - real

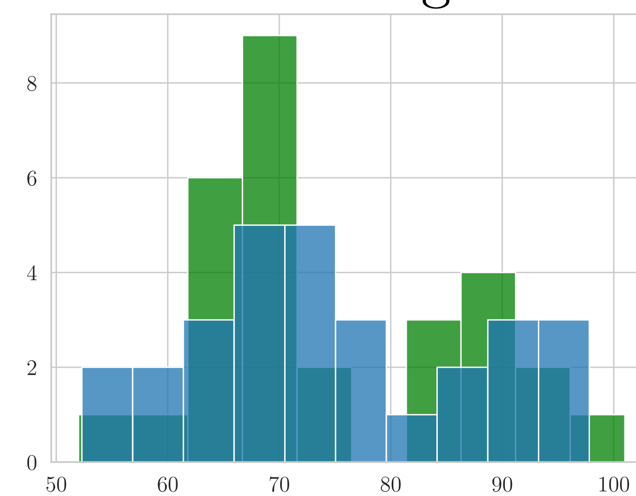


Observations

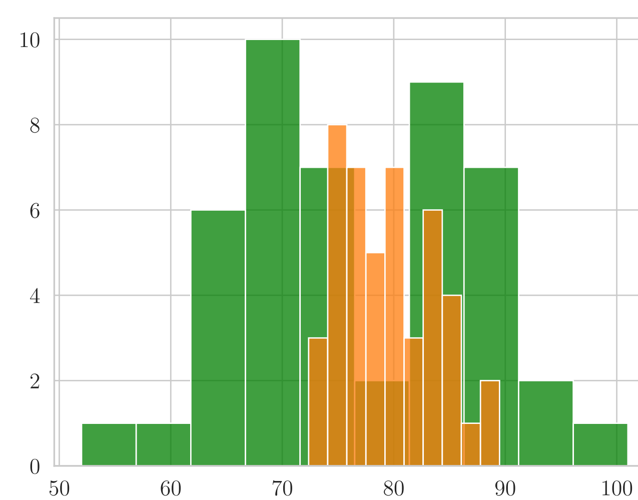


RMSE: 4.24
Cross.Corr: 113.18

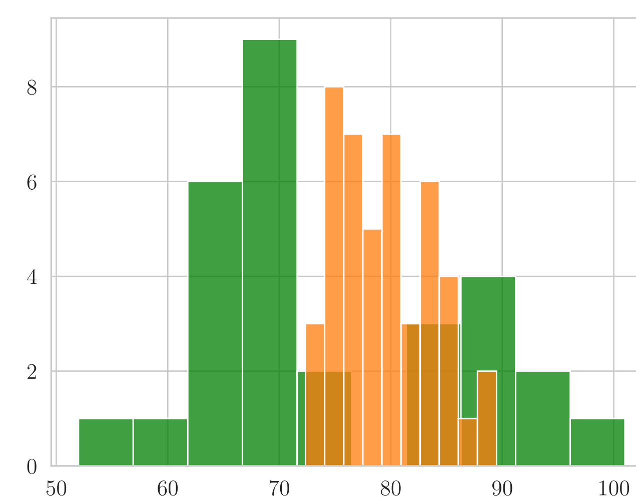
Missing



RMSE: 3.68
Cross.Corr: 127.52

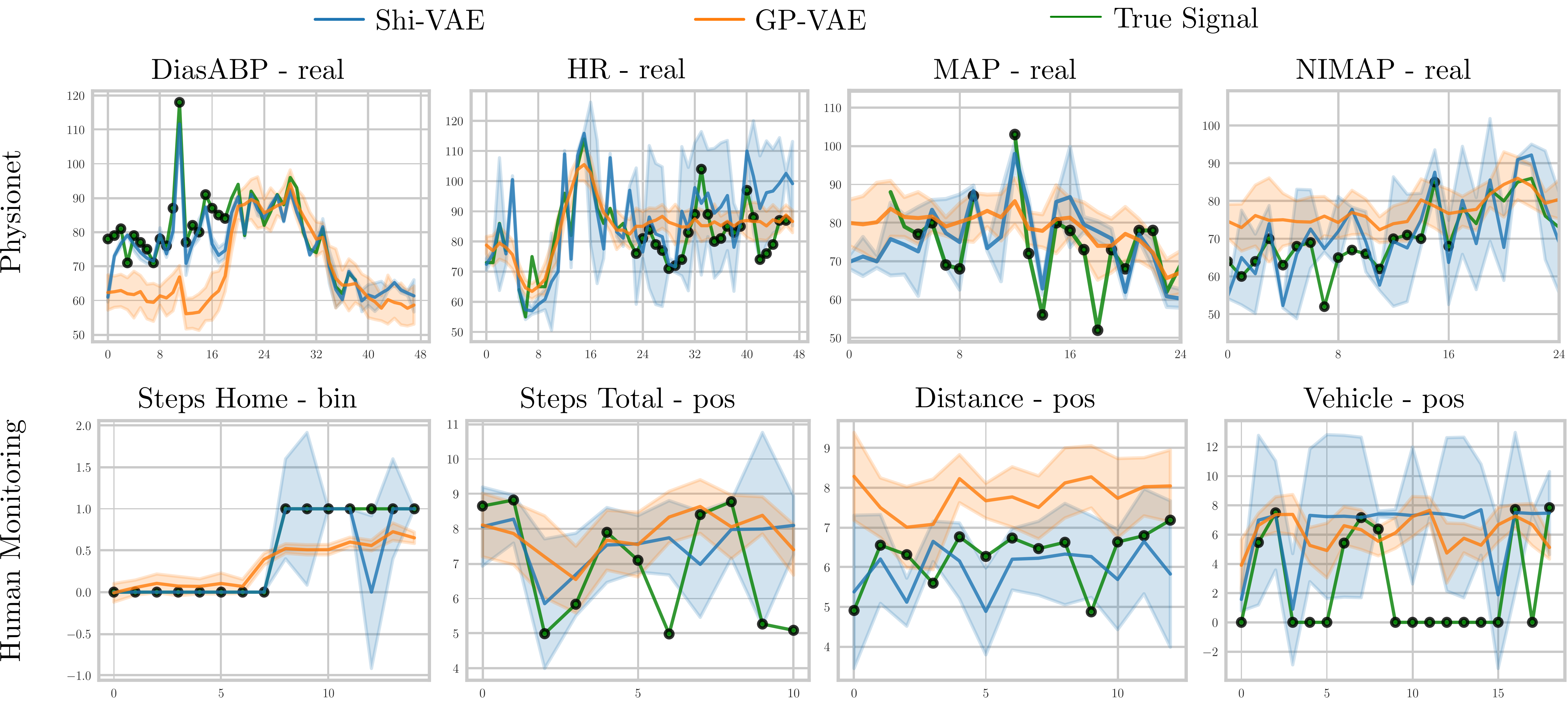


RMSE: 7.52
Cross.Corr: 38.86



RMSE: 5.87
Cross.Corr: 43.50

Physionet and Human Monitoring



Take Home Message

- Correlation in temporal scenarios is important!
- Standard error metrics can be not conclusive enough.
- The Shi-VAE model is able to capture hidden correlation within heterogeneous streams of data.

Bibliography

- Alfredo Nazabal, P. M. Olmos, Z. Gharhamani, and I. Valera “Handling Incomplete Heterogeneous data using VAEs”, *Pattern Recognition 2020*

- N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, “Deep unsupervised clustering with gaussian mixture variational autoencoders.” *CoRR, vol. abs/1611.02648, 2016*.

- J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. "A recurrent latent variable model for sequential data”. *CoRR, abs/1506.02216, 2015*.

- V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, “GP-VAE: Deep probabilistic time series imputation,” in *International Conference on Artificial Intelligence and Statistics*. *PMLR, 2020, pp*. *1651–1661*.

Code

Code: https://github.com/dbarrejon/Shi-VAE

Questions? Thank you!