MANCHESTER 1824
The University of Manchester

Gaussian Process Variational Autoencoder

Xinxing Shi

Sept. 16th, 2026

Contents

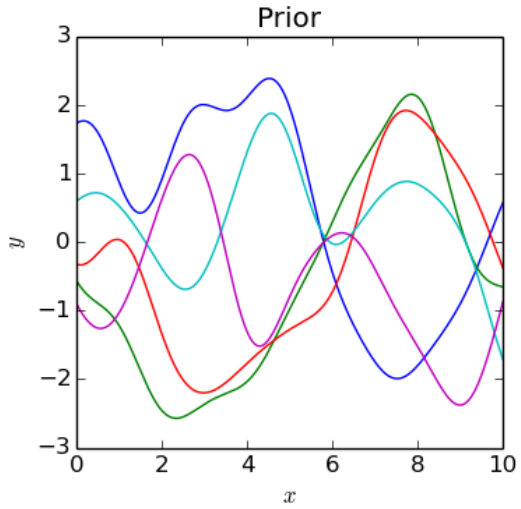
Contents

- Background
  - GP
  - VAE
- GPVAE Models
  - GPVAE with Inducing Points
  - Nearest-Neighbour GPVAE

Gaussian Process (GP)

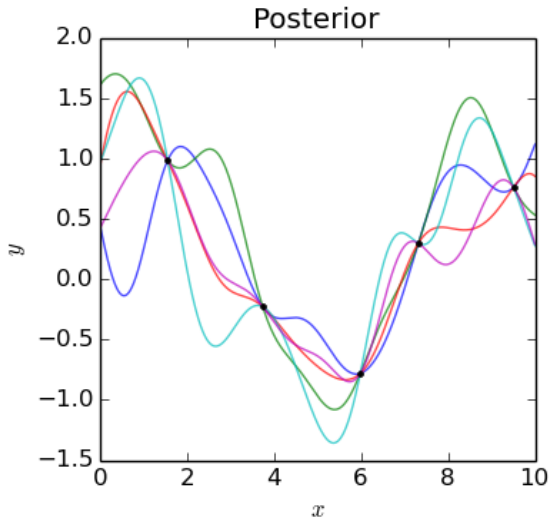
- GP builds correlations and calibrates uncertainty for variables
  - Prior: Any finite collection of random variables is Gaussian

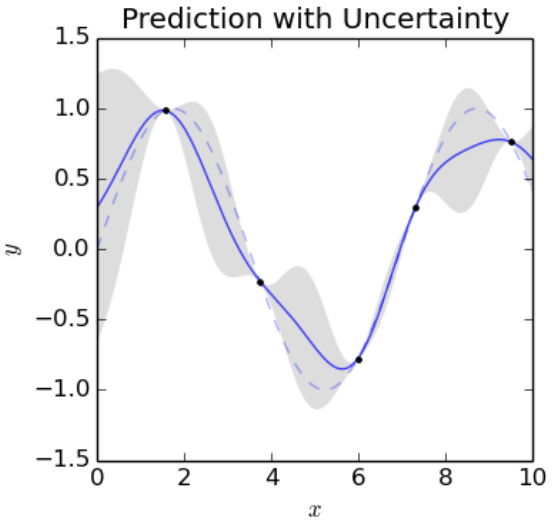
$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_i) \\ \vdots \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{x}_1) \\ \mu(\mathbf{x}_2) \\ \vdots \\ \mu(\mathbf{x}_i) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_i) & \cdots \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_i) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k(\mathbf{x}_i, \mathbf{x}_1) & k(\mathbf{x}_i, \mathbf{x}_2) & \cdots & k(\mathbf{x}_i, \mathbf{x}_i) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \end{bmatrix} \right)$$



  - Posterior: Conditioned on observations through Bayes' theorem

$$p(\mathbf{f} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f})}{p(\mathbf{y})}$$

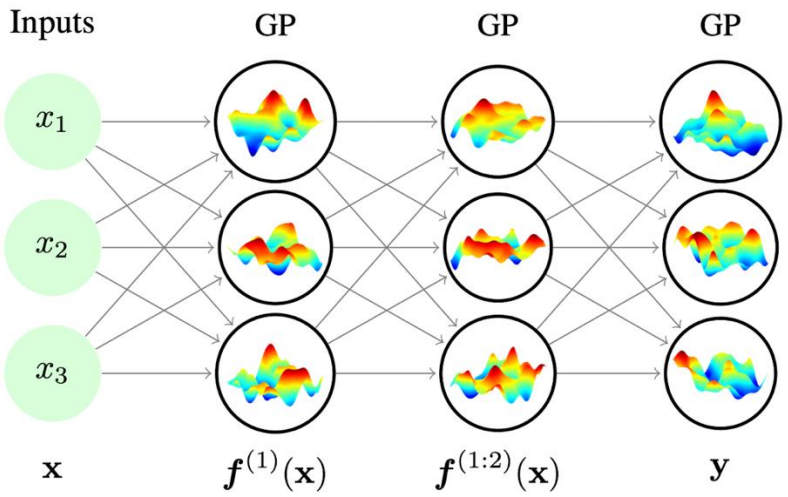




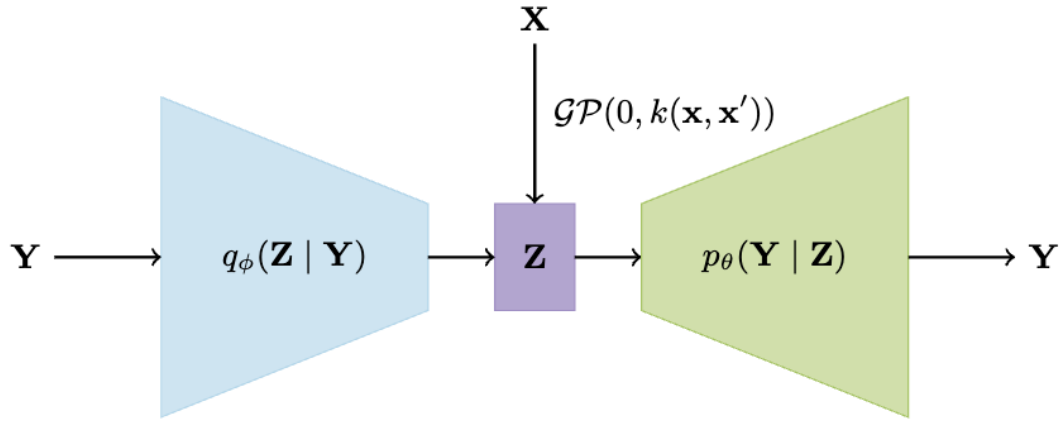
Gaussian processes for machine learning. Rasmussen, C. E. and Williams, C. K. I. 2006.

GP as Building Block

- More complex GP models:
  - (supervised) Deep GP



  - (unsupervised) GP Variational AutoEncoder (GPVAE)



- Problems of GPs are inherited:
  - Intractability of posteriors
  - Scalability issues due to *N*×*N* kernel matrix

Salimbeni, H. and Deisenroth, M. P. Doubly stochastic variational inference for deep Gaussian processes. NeurIPS 2017.
Casale, F. P., et al. Gaussian process prior variational autoencoders. NeurIPS 2018.

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f})\, \mathrm{d}\mathbf{f}$$

- Intractability

- Scalability issues due to *the N×N* covariance matrix

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{K_{ff}} + \sigma^2 \mathbf{I})$$

$$p(\mathbf{f}_* \mid \mathbf{y}) = \mathcal{N}\big(\mathbf{f}_* \mid \mathbf{K_{f_*f}}(\mathbf{K_{ff}} + \sigma^2 \mathbf{I})^{-1}\mathbf{y}, \mathbf{K_{f_*f_*}} - \mathbf{K_{f_*f}}(\mathbf{K_{ff}} + \sigma^2 \mathbf{I})^{-1}\mathbf{K_{ff_*}}\big)$$

Contents

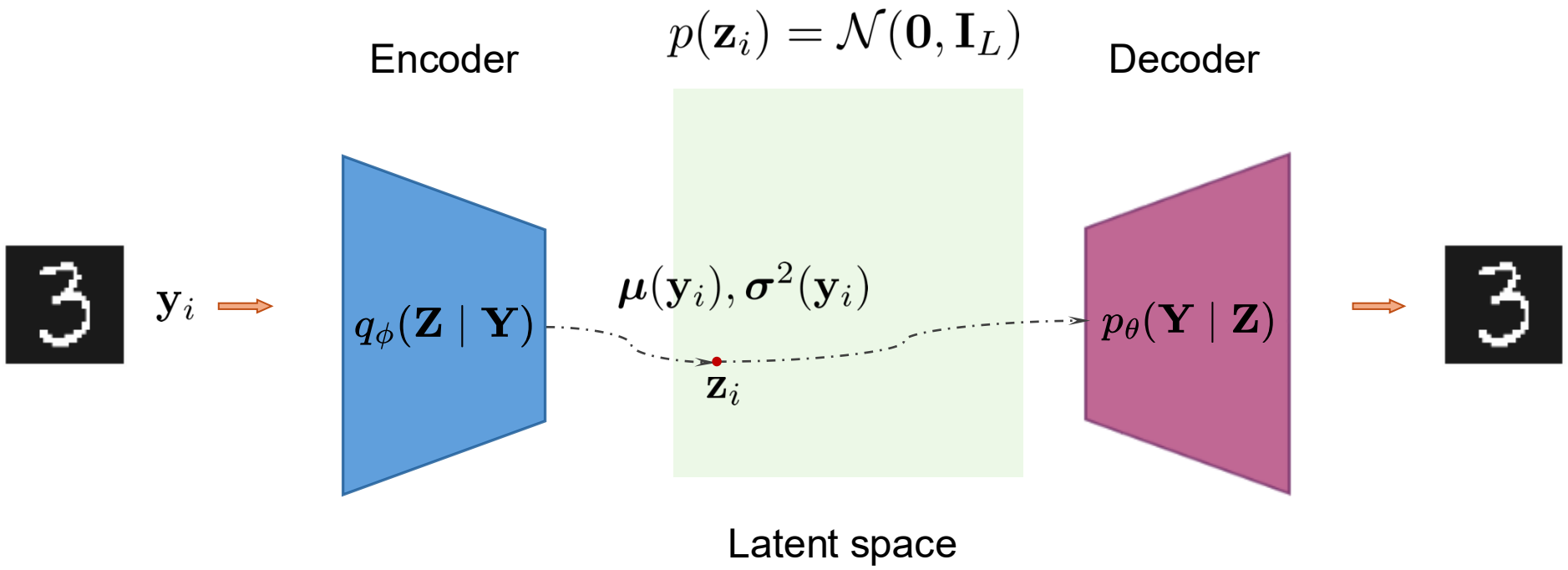
- Background
  - GP
  - VAE
- GPVAE Models
  - GPVAE with Inducing Points
  - Nearest-Neighbour GPVAE

Variational Autoencoder (VAE)

- Represent data points *Y* with Gaussian latent variables *Z*

$$\mathbf{Y} = [\mathbf{y}_i]_{i=1}^{N} \in \mathbb{R}^{N \times K} \qquad \mathbf{Z} = [\mathbf{z}_i]_{i=1}^{N} \in \mathbb{R}^{N \times L}$$

- Assumes a fully factorised Gaussian prior over the latent space

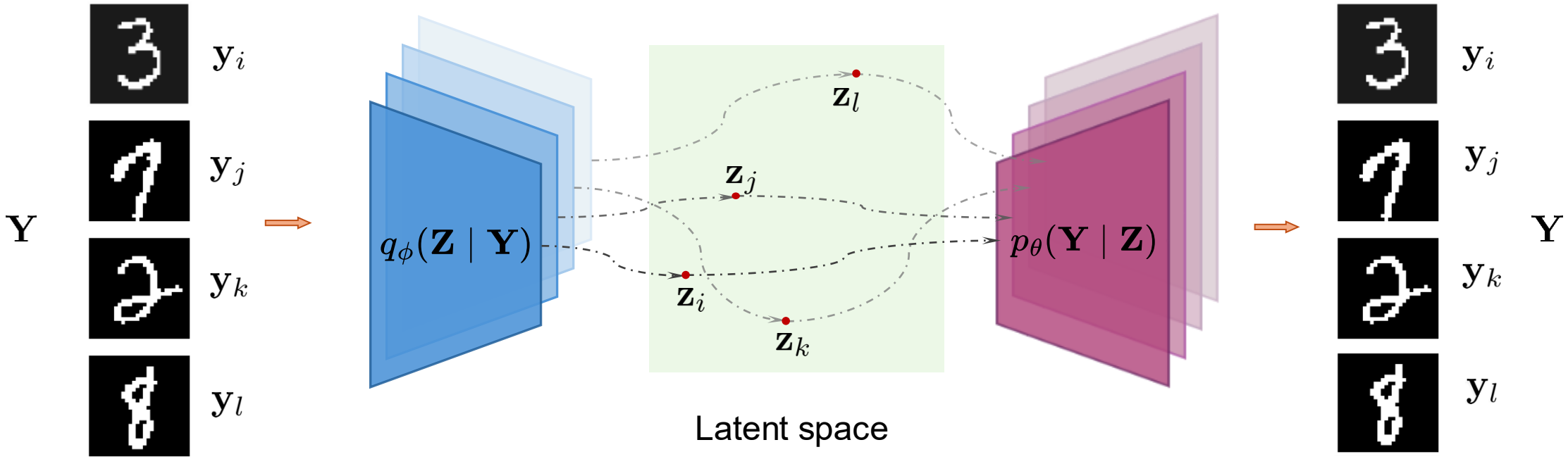


Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. ICLR 2014.

VAE: Factorised Prior

- Assumes a fully factorised Gaussian prior over the latent space

$$\mathbf{Z} \sim \prod_{i=1}^{N} p(\mathbf{z}_i) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{z}_i \mid \mathbf{0}, \mathbf{I}_L)$$

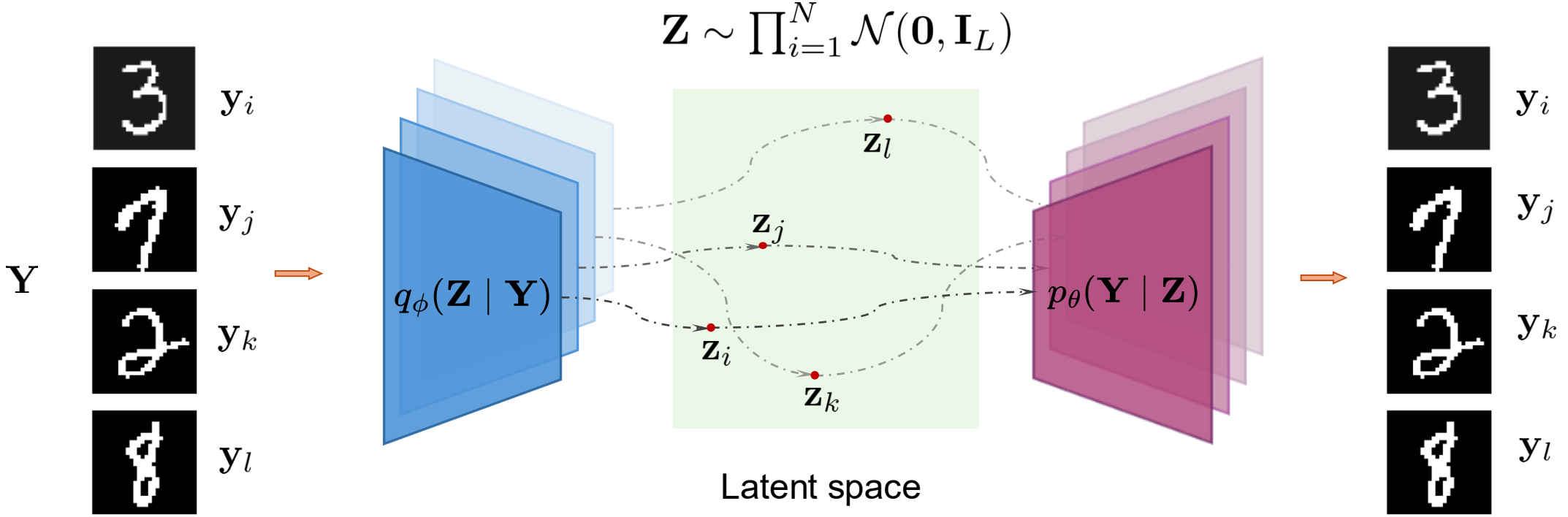


Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. ICLR 2014.

VAE: Training

- A standard VAE assumes a fully factorised Gaussian prior over the latent space for amortised variational inference

$$\mathcal{L}_{\text{VAE}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left\{ \mathbb{E}_{q(\mathbf{z}_i|\mathbf{y}_i)}[\log p(\mathbf{y}_i \mid \mathbf{z}_i)] - \text{KL}[q(\mathbf{z}_i \mid \mathbf{y}_i) \| p(\mathbf{z}_i)] \right\}$$



Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. ICLR 2014.

VAE: Independent Latent Variables

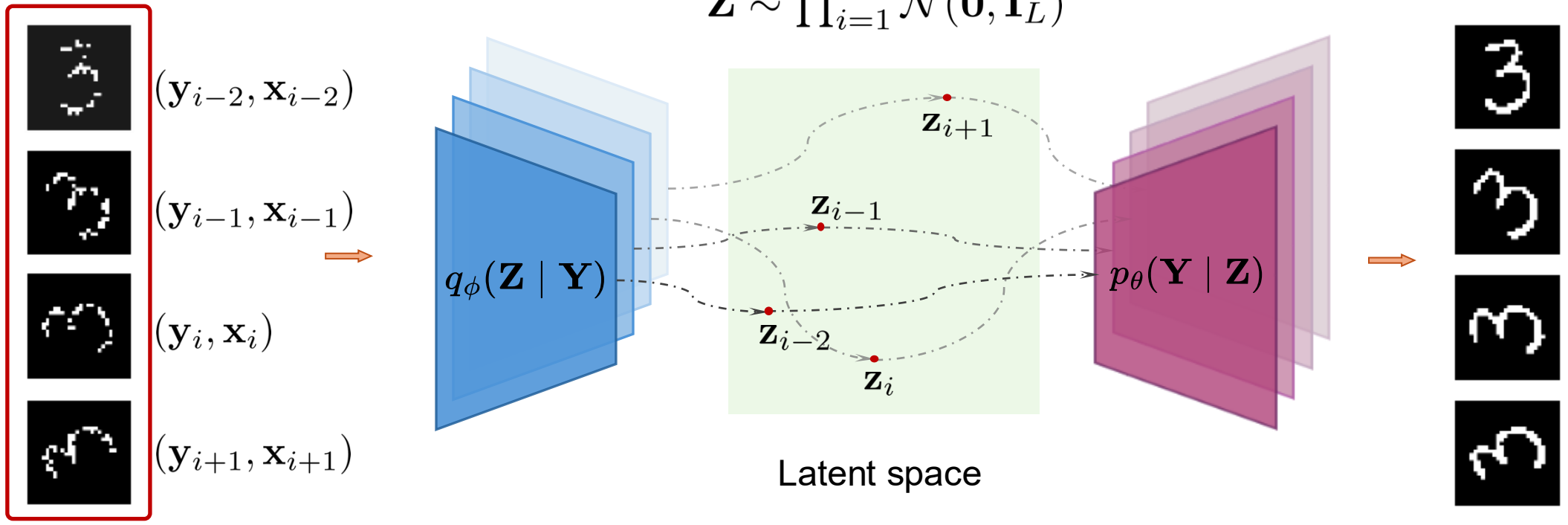
- A fully factorised Gaussian prior in a standard VAE ignores correlations of structured data *Y* with "timestamps" *X*

$$\mathbf{Y} = [\mathbf{y}_i]_{i=1}^N \in \mathbb{R}^{N \times K}, \quad \mathbf{X} = [\mathbf{x}_i]_{i=1}^N \in \mathbb{R}^{N \times D}$$

correlated samples

Introduce GPs for explicit, structured latent modelling

$$\mathbf{Z} \sim \prod_{i=1}^N \mathcal{N}(\mathbf{0}, \mathbf{I}_L)$$



Contents

- Background
  - GP
  - From VAE to GPVAE
- GPVAE Models
  - GPVAE with Inducing Points
  - Nearest-Neighbour GPVAE

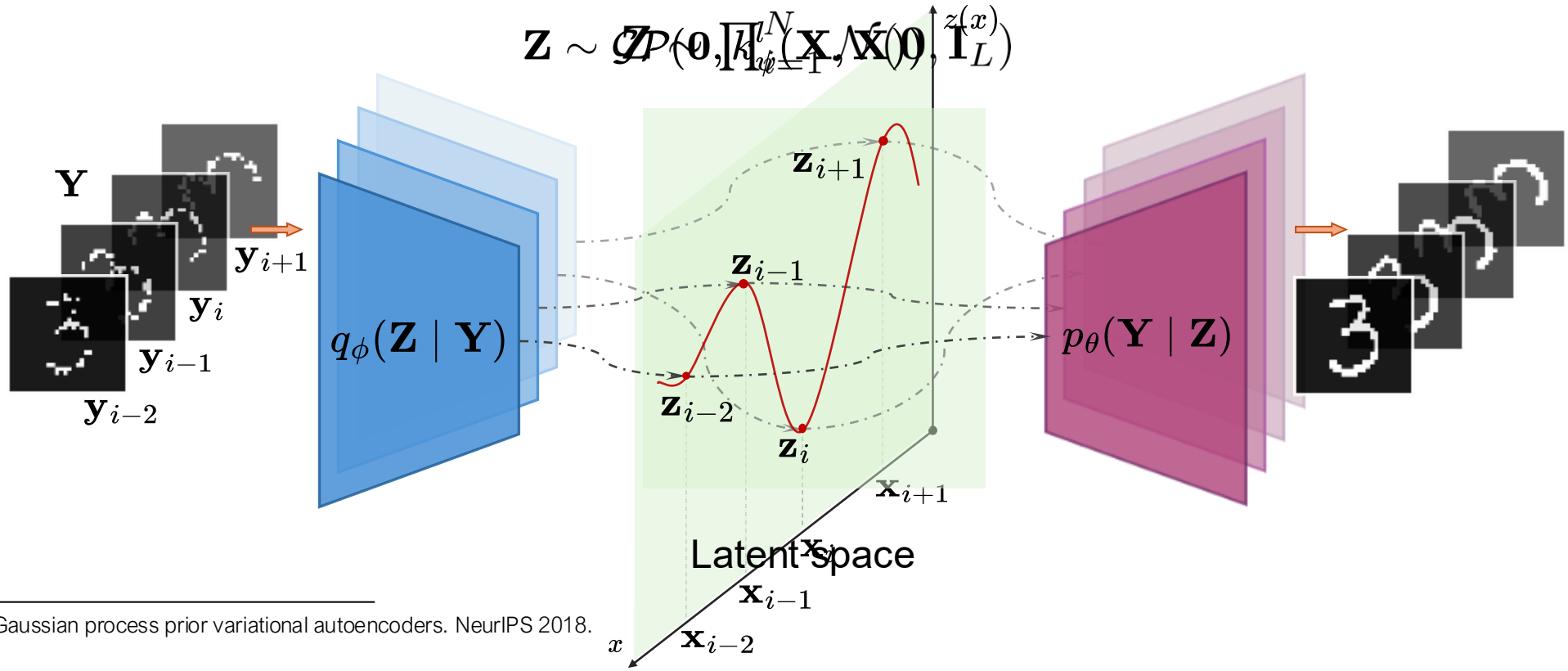
GPVAE: GP Prior in VAE but with Poor Scalability

- Build correlations via GP prior on data pairs *(X, Y)*

$$\mathbf{Y} = [\mathbf{y}_i]_{i=1}^N, \quad \mathbf{X} = [\mathbf{x}_i]_{i=1}^N$$

- Evidence lower bound (ELBO) scales poorly with $\mathcal{O}(N^3)$

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z}|\mathbf{Y})}[\log p(\mathbf{Y} \mid \mathbf{Z})] - \mathrm{KL}[q(\mathbf{Z} \mid \mathbf{Y}) \,\|\, p(\mathbf{Z})]$$



Casale, F. P., et al. Gaussian process prior variational autoencoders. NeurIPS 2018.

GPVAE: ELBO Derivation

- Data pairs *(X, Y)*

$$\mathbf{Y} = [\mathbf{y}_i]_{i=1}^{N} \in \mathbb{R}^{N \times K}, \quad \mathbf{X} = [\mathbf{x}_i]_{i=1}^{N} \in \mathbb{R}^{N \times D}$$

- Latent variables *Z*

$$\text{GP prior: } p(\mathbf{Z}) = \mathcal{N}(\mathbf{Z} \mid \mathbf{0}, \mathbf{K}_{\mathbf{XX}})$$

$$\text{Variational posterior: } q(\mathbf{Z} \mid \mathbf{Y}) = \mathcal{N}(\mathbf{Z} \mid \mu(\mathbf{Y}), \sigma^2(\mathbf{Y}))$$

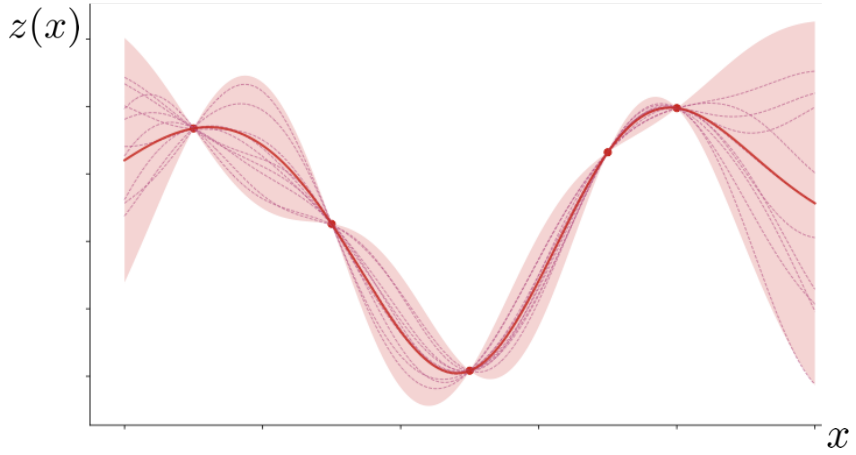
- ELBO

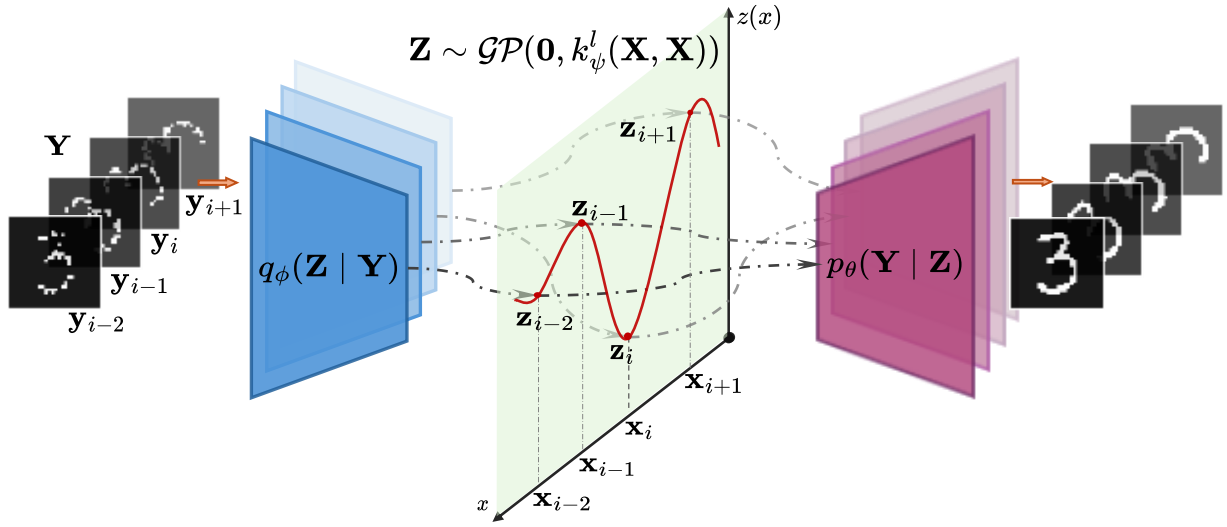
$$\begin{aligned}
\log p(\mathbf{Y}) &= \log \int p(\mathbf{Z}) p(\mathbf{Y} \mid \mathbf{Z}) \, \mathrm{d}\mathbf{Z} \\
&\geq \int q(\mathbf{Z} \mid \mathbf{Y}) \log \frac{p(\mathbf{Z}) p(\mathbf{Y} \mid \mathbf{Z})}{q(\mathbf{Z} \mid \mathbf{Y})} \, \mathrm{d}\mathbf{Z} \\
&= \mathbb{E}_{q(\mathbf{Z}|\mathbf{Y})}[\log p(\mathbf{Y} \mid \mathbf{Z})] - \mathrm{KL}[q(\mathbf{Z} \mid \mathbf{Y}) \,\|\, p(\mathbf{Z})]
\end{aligned}$$

O(N^3) complexity!

Mindset: Find Solutions from Scalable GP

- Desiderata:
  - Maintain the VAE architecture
  - Enable mini-batch training
  - Support various kernel choices
  - ...
- Draw inspirations from scalable GPs
  - Sparse GP with inducing point methods
  - Nearest-neighbour GP
  - Markovian GP with state-space representation
  - Sampling-based GP
  - …





Jazbec, M., et al. Scalable Gaussian process variational autoencoders. AISTATS 2021.
Shi, X., et al. Neighbour-driven Gaussian process variational autoencoders for scalable structured latent modelling. ICML 2025.
Zhu, H., et al. Markovian Gaussian process variational autoencoders. ICML 2023.
Tran, B.-H., et al. Fully Bayesian autoencoders with latent sparse Gaussian processes. ICML 2023.

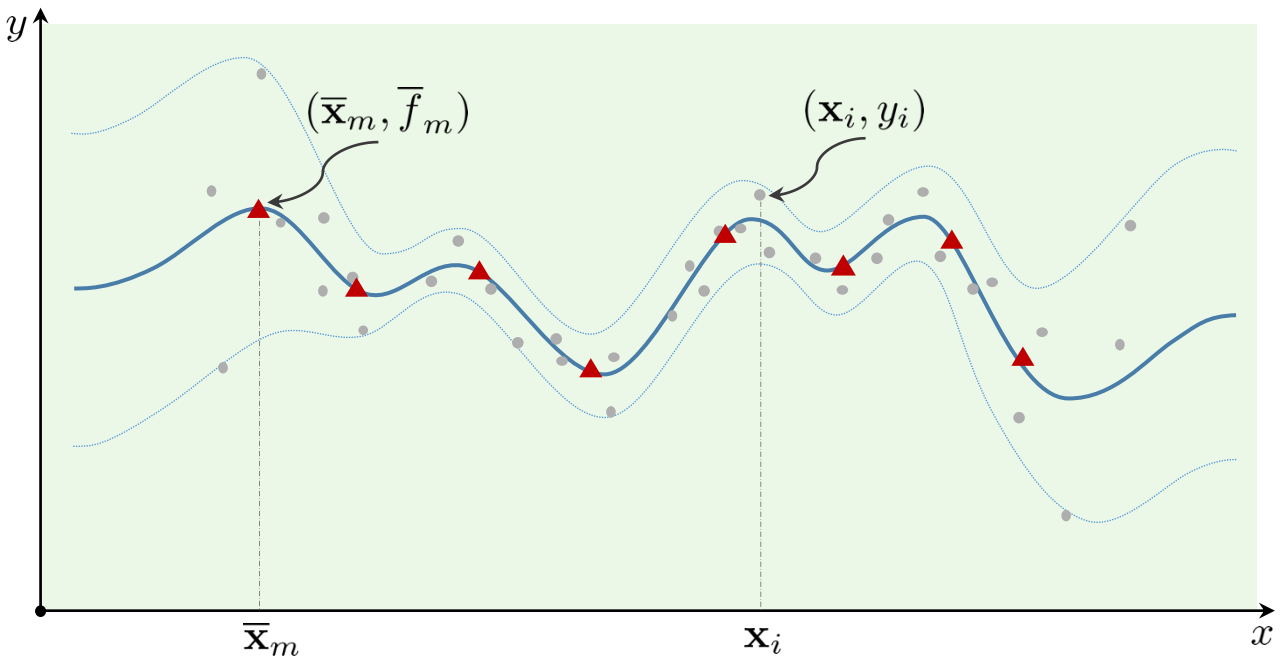
Contents

- Background
  - GP
  - From VAE to GPVAE

- GPVAE Models
  - GPVAE with Inducing Points
  - Nearest-Neighbour GPVAE

Sparse Variational GP (SVGP)

- Use a small set of M (<< N) representative points to summarise data



- Observations $\mathbf{y} = [y_i]_{i=1}^N,\ \mathbf{X} = [\mathbf{x}_i]_{i=1}^N$

  GP prior: $p(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mathbf{0}, \mathbf{K_{ff}})$

  Likelihood: $p(\mathbf{y} \mid \mathbf{f}) = \prod_i p(y_i \mid f_i)$

- Inducing points $\overline{\mathbf{f}} = [\overline{f}_m]_{m=1}^M,\ \overline{\mathbf{X}} = [\overline{\mathbf{x}}_m]_{m=1}^M$

  Augmented GP prior $p(\mathbf{f}, \overline{\mathbf{f}})$:

$$\begin{bmatrix} \mathbf{f} \\ \overline{\mathbf{f}} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K_{ff}} & \mathbf{K_{f\overline{f}}} \\ \mathbf{K_{\overline{f}f}} & \mathbf{K_{\overline{f}\overline{f}}} \end{bmatrix}\right)$$

  Variational posterior:

$$q(\mathbf{f}, \overline{\mathbf{f}}) = p(\mathbf{f} \mid \overline{\mathbf{f}}) q(\overline{\mathbf{f}}),\ q(\overline{\mathbf{f}}) = \mathcal{N}(\overline{\mathbf{f}} \mid \mathbf{m}, \mathbf{S})$$

Maximise:

$$\log p(\mathbf{y}) \geq \mathbb{E}_{p(\mathbf{f}|\overline{\mathbf{f}})q(\overline{\mathbf{f}})}[\log p(\mathbf{y} \mid \mathbf{f})] - \mathrm{KL}[q(\overline{\mathbf{f}}) \| p(\overline{\mathbf{f}})]$$

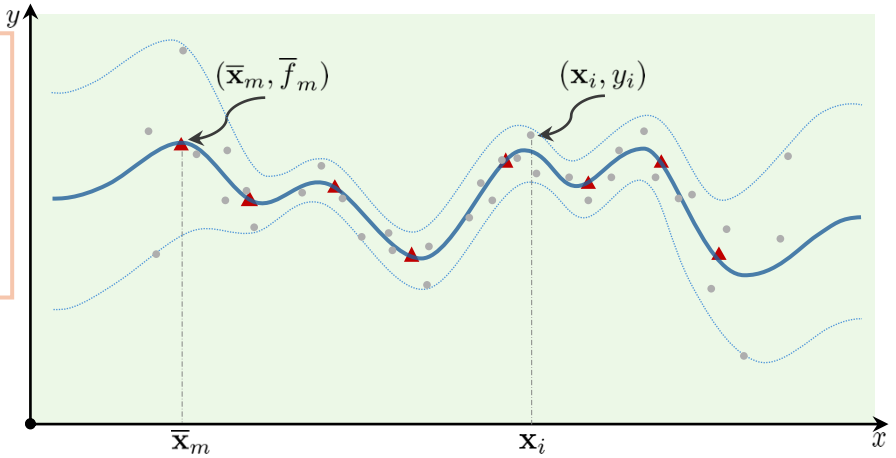
Evidence Lower Bound (ELBO)

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. AISTATS 2009.
Hensman, J., et al. Scalable variational Gaussian process classification. AISTATS 2015.

SVGP: ELBO

- A generic likelihood $p(y_i \mid f_i)$ may be non-Gaussian

$$\mathbb{E}_{p(\mathbf{f}|\bar{\mathbf{f}})q(\bar{\mathbf{f}})}[\log p(\mathbf{y} \mid \mathbf{f})] - \mathrm{KL}[q(\bar{\mathbf{f}}) \| p(\bar{\mathbf{f}})]$$



- Closed-form KL

$$p(\bar{\mathbf{f}}) = \mathcal{N}(\bar{\mathbf{f}} \mid \mathbf{0}, \mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}) \qquad q(\bar{\mathbf{f}}) = \mathcal{N}(\bar{\mathbf{f}} \mid \mathbf{m}, \mathbf{S})$$

$$\mathrm{KL}[q\|p] = \tfrac{1}{2}\left[\mathbf{m}^\top \mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{m} + \mathrm{tr}\{\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{S}\} - \log \frac{|\mathbf{S}|}{|\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}|} - M\right]$$

- Decomposable expected log-likelihood

$$\mathbb{E}_{p(\mathbf{f}|\bar{\mathbf{f}})q(\bar{\mathbf{f}})}[\log p(\mathbf{y} \mid \mathbf{f})] = \sum_{i=1}^{N} \mathbb{E}_{q(f_i)}[\log p(y_i \mid f_i)] \approx \frac{N}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathbb{E}_{q(f_i)}[\log p(y_i \mid f_i)]$$

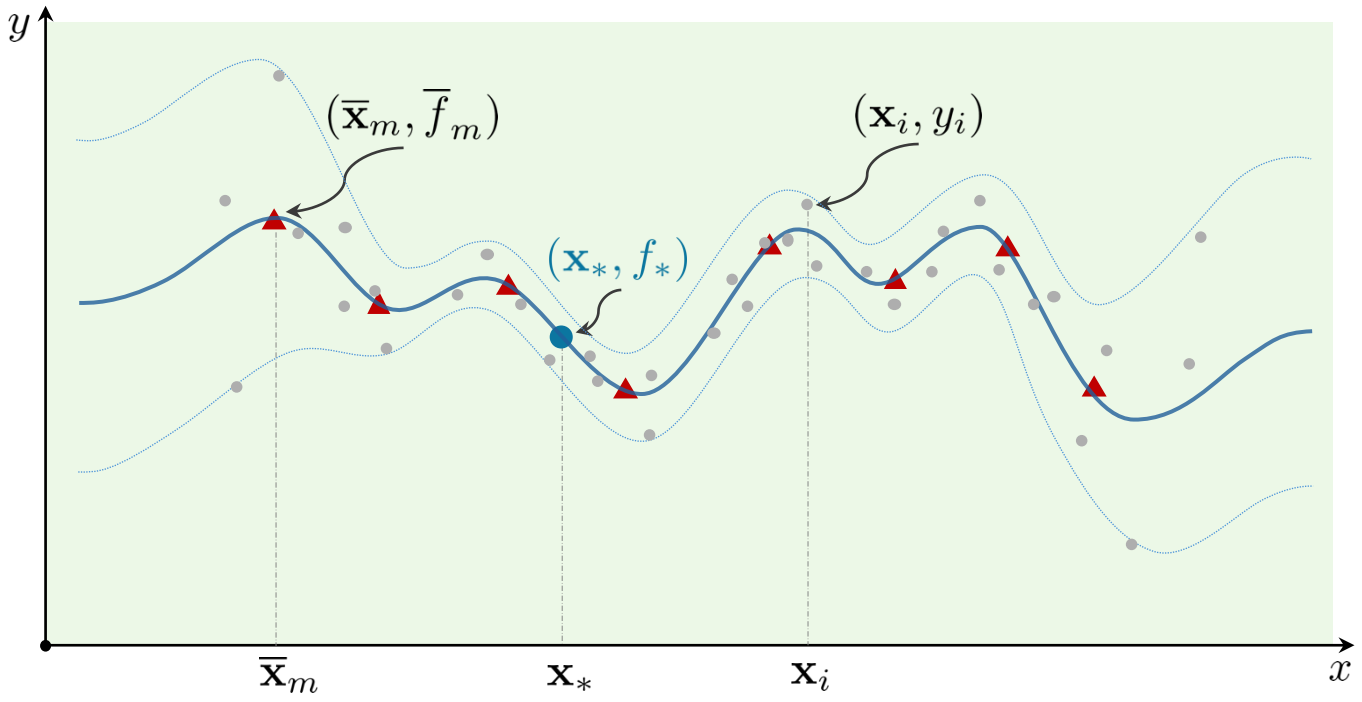
Marginalising + Mini-batching

$$q(f_i) = \int p(f_i \mid \bar{\mathbf{f}}) q(\bar{\mathbf{f}})\, \mathrm{d}\bar{\mathbf{f}} = \mathcal{N}\left(f_i \mid \mathbf{K}_{f_i\bar{\mathbf{f}}}\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{m},\ \underbrace{K_{f_i f_i} + \mathbf{K}_{f_i\bar{\mathbf{f}}}\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}(\mathbf{S} - \mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}})\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{K}_{\bar{\mathbf{f}}f_i}}_{\mathbf{K}_{q(\mathbf{f})}}\right)$$

Hensman, J., et al. Scalable variational Gaussian process classification. AISTATS 2015.

SVGP: Prediction

- Compute posterior predictive distribution via $q(f_*, \overline{\mathbf{f}})$



- GP Prediction $f_*(\mathbf{x}_*)$

$$q(f_*, \overline{\mathbf{f}}) = p(f_* \mid \overline{\mathbf{f}}) q(\overline{\mathbf{f}})$$

$$q(f_*) = \int p(f_* \mid \overline{\mathbf{f}}) q(\overline{\mathbf{f}}) \, \mathrm{d}\overline{\mathbf{f}}$$

$$= \mathcal{N}\big(f_* \mid \mathbf{K}_{f_*\overline{\mathbf{f}}} \mathbf{K}_{\overline{\mathbf{f}}\overline{\mathbf{f}}}^{-1} \mathbf{m},$$

$$K_{f_* f_*} + \mathbf{K}_{f_*\overline{\mathbf{f}}} \mathbf{K}_{\overline{\mathbf{f}}\overline{\mathbf{f}}}^{-1} (\mathbf{S} - \mathbf{K}_{\overline{\mathbf{f}}\overline{\mathbf{f}}}) \mathbf{K}_{\overline{\mathbf{f}}\overline{\mathbf{f}}}^{-1} \mathbf{K}_{\overline{\mathbf{f}} f_*}\big)$$

- Prediction $y_*(\mathbf{x}_*)$

$$q(y_* \mid \mathbf{y}) = \int p(y_* \mid f_*) q(f_*) \, \mathrm{d}f_*$$

Hensman, J., et al. Scalable variational Gaussian process classification. AISTATS 2015.

SVGP with Gaussian Likelihood

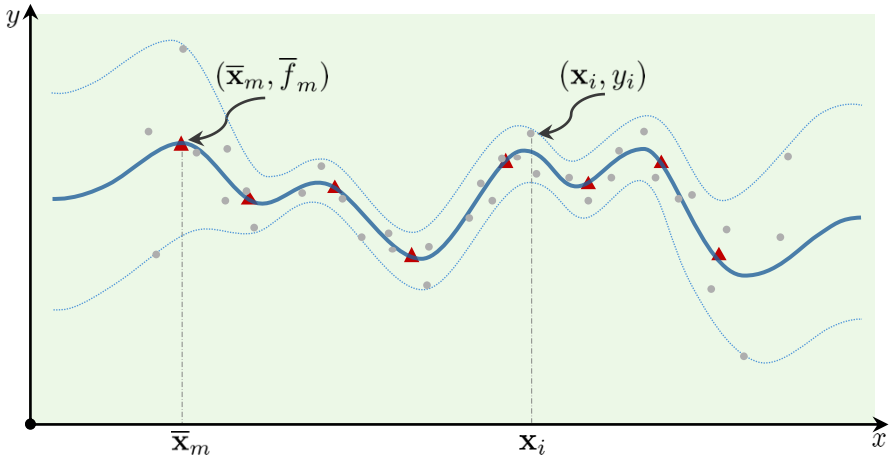
- When the likelihood is Gaussian $\mathcal{N}(y_i \mid f_i, \sigma^2)$

$$\mathbb{E}_{p(\mathbf{f}|\bar{\mathbf{f}})q(\bar{\mathbf{f}})}[\log p(\mathbf{y} \mid \mathbf{f})] - \mathrm{KL}[q(\bar{\mathbf{f}}) \| p(\bar{\mathbf{f}})]$$

  - Uncollapsed bound with mini-batching

$$\log \mathcal{N}\left(\mathbf{y} \mid \mathbf{K}_{\mathbf{f}\bar{\mathbf{f}}}\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{m}, \sigma^2\mathbf{I}\right) - \frac{\mathrm{tr}\left\{\mathbf{K}_{q(\mathbf{f})}\right\}}{2\sigma^2}$$

$$\approx \frac{N}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\{\log \mathcal{N}\left(y_i \mid \mathbf{K}_{f_i\bar{\mathbf{f}}}\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{m}, \sigma^2\right) - \frac{\mathrm{tr}\left\{\mathbf{K}_{q(f_i)}\right\}}{2\sigma^2}\right\}$$



  - Collapsed bound with analytical $q(\bar{\mathbf{f}})$

$$q^*(\bar{\mathbf{f}}) = \mathcal{N}\left(\bar{\mathbf{f}} \mid \sigma^{-2}\mathbf{K}_{\bar{\mathbf{f}}\mathbf{f}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{K}_{\bar{\mathbf{f}}\mathbf{f}}\mathbf{y}, \mathbf{K}_{\bar{\mathbf{f}}\mathbf{f}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{K}_{\bar{\mathbf{f}}\mathbf{f}}\right) \qquad \hat{\mathbf{\Sigma}} = \mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}} + \sigma^{-2}\mathbf{K}_{\bar{\mathbf{f}}\mathbf{f}}\mathbf{K}_{\mathbf{f}\bar{\mathbf{f}}}$$

No mini-batching

$$\mathrm{ELBO} = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \mathbf{K}_{\mathbf{f}\bar{\mathbf{f}}}\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{K}_{\bar{\mathbf{f}}\mathbf{f}} + \sigma^2\mathbf{I}\right) - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathbf{K}_{\mathbf{f}\mathbf{f}} - \mathbf{K}_{\mathbf{f}\bar{\mathbf{f}}}\mathbf{K}_{\bar{\mathbf{f}}\bar{\mathbf{f}}}^{-1}\mathbf{K}_{\bar{\mathbf{f}}\mathbf{f}}\right\}$$

Hensman, J., et al. Gaussian process for big data. UAI 2013.
Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. AISTATS 2009.

SVGPVAE: Amortised Inducing Points

- Start from the inducing-point distribution

$$p(\overline{\mathbf{f}}) = \mathcal{N}(\overline{\mathbf{f}} \mid \mathbf{0}, \mathbf{K}_{\overline{\mathbf{ff}}})$$

$$q^*(\overline{\mathbf{f}}) = \mathcal{N}\left(\overline{\mathbf{f}} \mid \sigma^{-2}\mathbf{K}_{\overline{\mathbf{ff}}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{K}_{\overline{\mathbf{ff}}}\mathbf{y}, \mathbf{K}_{\overline{\mathbf{ff}}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{K}_{\overline{\mathbf{ff}}}\right) \qquad \hat{\mathbf{\Sigma}} = \mathbf{K}_{\overline{\mathbf{ff}}} + \sigma^{-2}\mathbf{K}_{\overline{\mathbf{ff}}}\mathbf{K}_{\overline{\mathbf{ff}}}$$

- Introduce inducing points in latent space

$$p(\overline{\mathbf{Z}}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{\overline{\mathbf{ZZ}}}), \quad k(\overline{\mathbf{X}}, \overline{\mathbf{X}}) = \mathbf{K}_{\overline{\mathbf{ZZ}}}$$

- Design variational posterior
  - Encoder outputs pseudo-observations / variances for a mini-batch $(\mathbf{X}_\mathcal{B}, \mathbf{Y}_\mathcal{B}, \mathbf{Z}_\mathcal{B})$

$$\mathbf{y} \longrightarrow \boldsymbol{\mu}_\mathcal{B}, \qquad \sigma^2 \longrightarrow \boldsymbol{\sigma}^2_\mathcal{B}$$

  - Mini-batch version of "$q^*(\overline{\mathbf{f}})$": $q(\overline{\mathbf{Z}} \mid \mathbf{Y}_\mathcal{B}) = \mathcal{N}(\overline{\mathbf{Z}} \mid \mathbf{m}_\mathcal{B}, \mathbf{S}_\mathcal{B})$

$$\mathbf{m}_\mathcal{B} = \tfrac{N}{|\mathcal{B}|}\mathbf{K}_{\overline{\mathbf{zz}}}\mathbf{\Sigma}_\mathcal{B}^{-1}\mathbf{K}_{\overline{\mathbf{z}}\mathbf{z}_\mathcal{B}}\boldsymbol{\sigma}_\mathcal{B}^{-2}\boldsymbol{\mu}_\mathcal{B} \quad \mathbf{S}_\mathcal{B} = \mathbf{K}_{\overline{\mathbf{zz}}}\mathbf{\Sigma}_\mathcal{B}^{-1}\mathbf{K}_{\overline{\mathbf{zz}}} \qquad \mathbf{\Sigma}_\mathcal{B} = \mathbf{K}_{\overline{\mathbf{zz}}} + \tfrac{N}{|\mathcal{B}|}\mathbf{K}_{\overline{\mathbf{z}}\mathbf{z}_\mathcal{B}}\boldsymbol{\sigma}_\mathcal{B}^{-2}\mathbf{K}_{\mathbf{z}_\mathcal{B}\overline{\mathbf{z}}}$$

  - Variational posterior: $q(\mathbf{z}_n \mid \mathbf{Y}_\mathcal{B}) = \int p(\mathbf{z}_n \mid \overline{\mathbf{Z}}_\mathcal{B}) q(\overline{\mathbf{Z}}_\mathcal{B} \mid \mathbf{Y}_\mathcal{B}) \, \mathrm{d}\overline{\mathbf{Z}}_\mathcal{B}$

Jazbec, M. et al. Scalable Gaussian process variational autoencoders. AISTATS 2021.

GPVAE with Inducing Points

- Training objective based on VAE and SVGP formulation

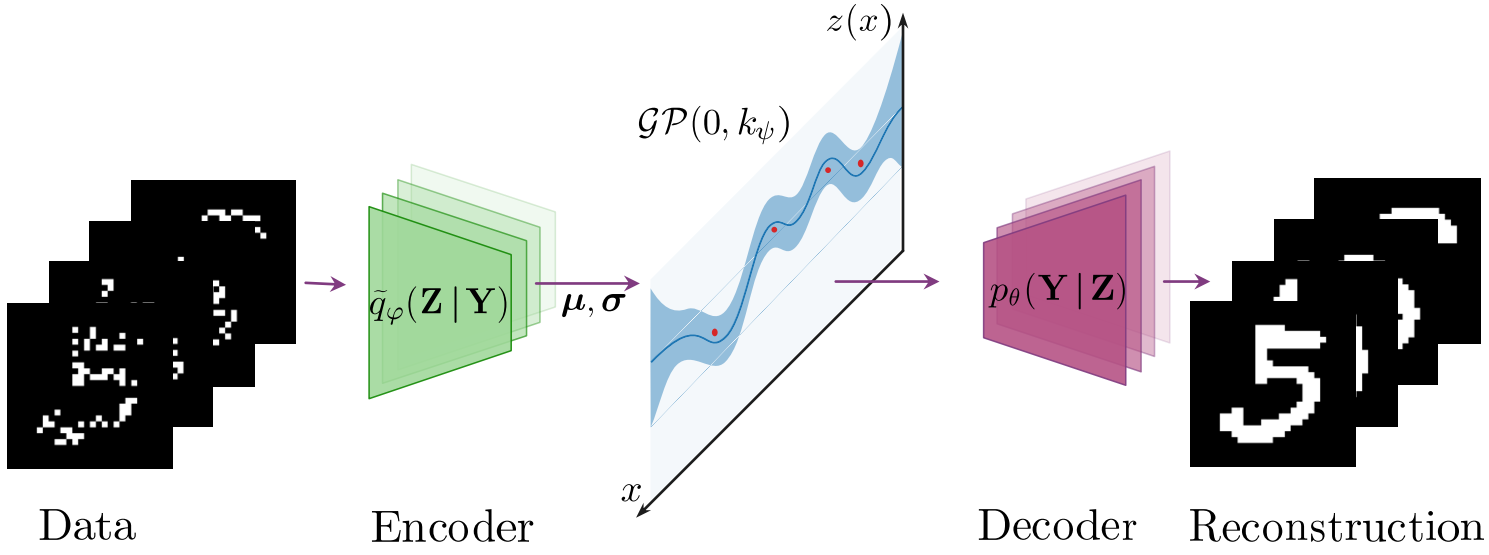
$$\mathcal{L}_{\text{SVGPVAE}} = \frac{N}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \mathbb{E}_{q(\mathbf{z}_n \mid \mathbf{Y}_\mathcal{B})} [\log p(\mathbf{y}_n \mid \mathbf{z}_n)] - \mathrm{KL}[q(\overline{\mathbf{Z}} \mid \mathbf{Y}_\mathcal{B}) \,\|\, p(\overline{\mathbf{Z}})]$$

Simplified reconstruction-minus-KL training objective

- Prediction through Gaussian conditional

$$q(\mathbf{z}_* \mid \mathbf{Y}) = \int p(\mathbf{z}_* \mid \overline{\mathbf{Z}})\, q(\overline{\mathbf{Z}} \mid \mathbf{Y})\, \mathrm{d}\overline{\mathbf{Z}}$$

Use full-batch or mini-batch



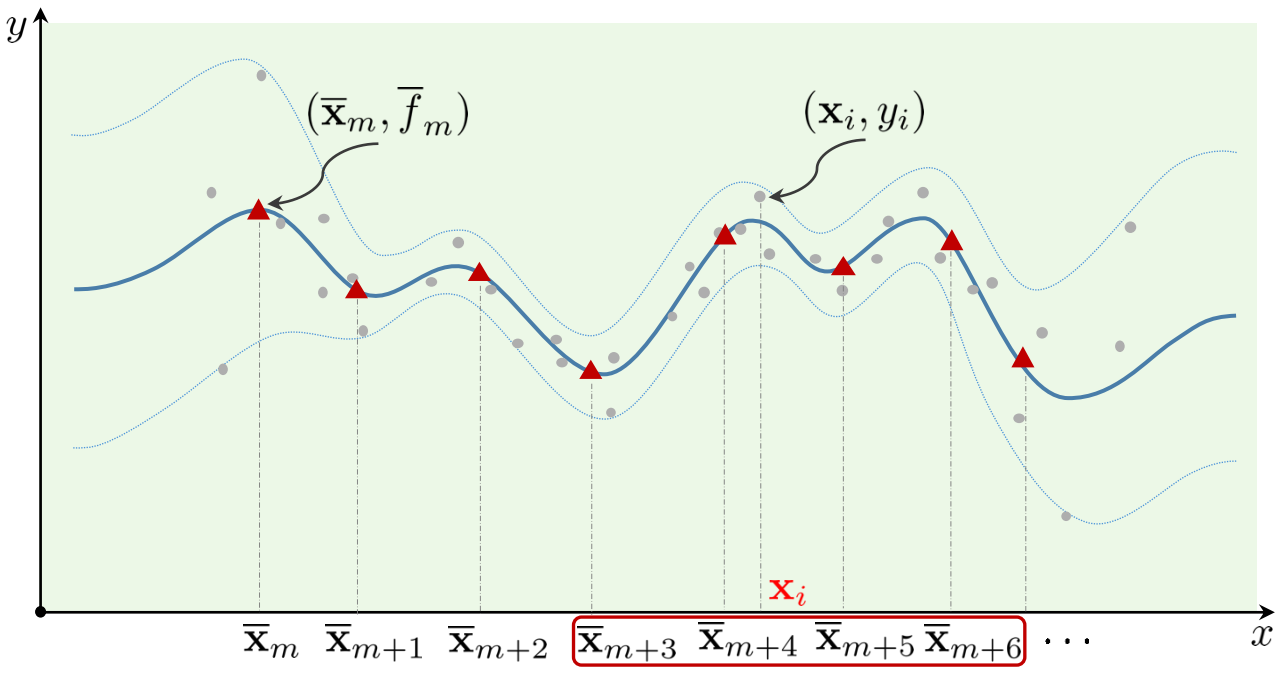
Jazbec, M. et al. Scalable Gaussian process variational autoencoders. AISTATS 2021.

Contents

- Background
  - GP
  - From VAE to GPVAE
- GPVAE Models
  - GPVAE with Inducing Points
  - Nearest-Neighbour GPVAE

Nearest-Neighbour GP (NNGP)

- Consider the *H*-nearest inducing points of SVGP



neighbours of $\mathbf{x}_i$:

$$n(i) = \{m+3, m+4, m+5, m+6\}$$

- NNGP: "sparse within sparse"

$$p(f_i \mid \overline{\mathbf{f}}) \longrightarrow p(f_i \mid \overline{\mathbf{f}}_{n(i)})$$

$$\mathcal{N}\left(f_i \mid \mathbf{K}_{f_i\overline{\mathbf{f}}}\mathbf{K}_{\overline{\mathbf{f}}\overline{\mathbf{f}}}^{-1}\overline{\mathbf{f}}, K_{f_if_i} - \mathbf{K}_{f_i\overline{\mathbf{f}}}\mathbf{K}_{\overline{\mathbf{f}}\overline{\mathbf{f}}}^{-1}\mathbf{K}_{\overline{\mathbf{f}}f_i}\right)$$

$$q(\overline{\mathbf{f}}) \longrightarrow q(\overline{\mathbf{f}}_{n(i)})$$

$$p(\overline{\mathbf{f}}) \longrightarrow p(\overline{\mathbf{f}}_{n(i)})$$

- ELBO:

$$\frac{N}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}_{p(f_i|\overline{\mathbf{f}})q(\overline{\mathbf{f}})}[\log p(y_i \mid f_i)] - \mathrm{KL}[q(\overline{\mathbf{f}})\|p(\overline{\mathbf{f}})]$$

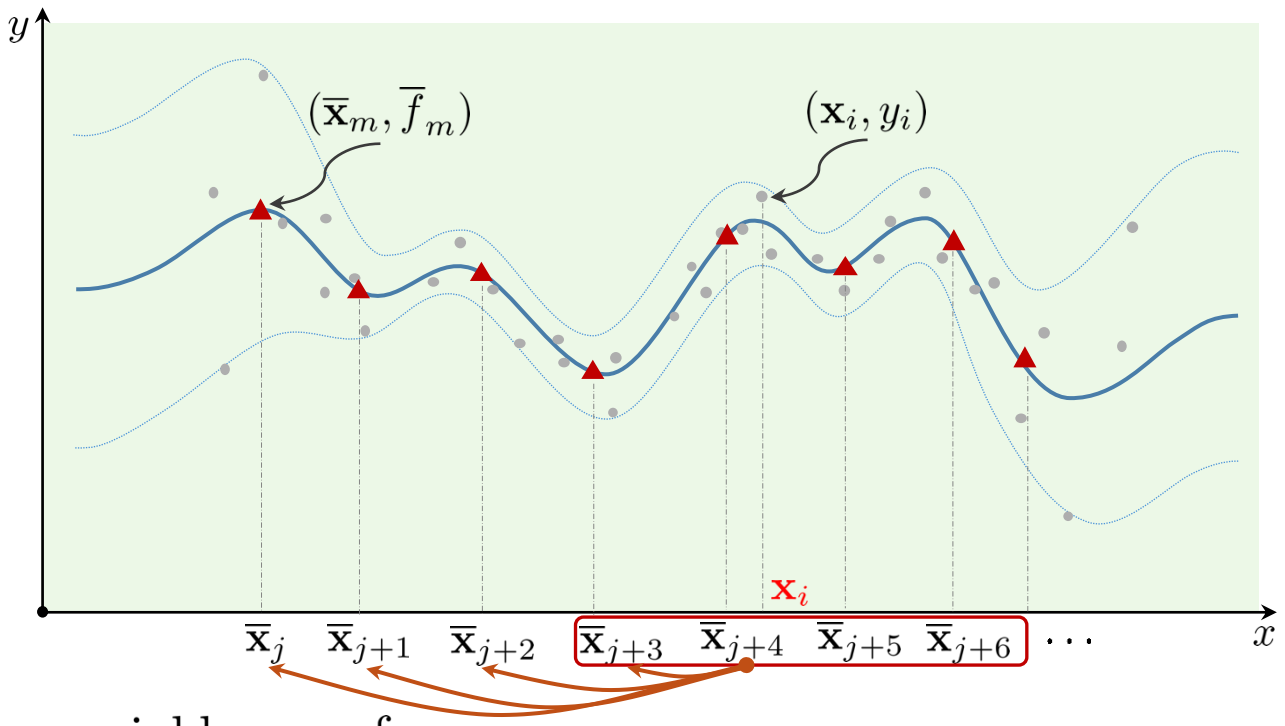
$$\downarrow$$

$$\frac{N}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}_{p(f_i|\overline{\mathbf{f}}_{n(i)})q(\overline{\mathbf{f}}_{n(i)})}[\log p(y_i \mid f_i)] - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathrm{KL}[q(\overline{\mathbf{f}}_{n(i)})\|p(\overline{\mathbf{f}}_{n(i)})]$$

Tran, G.-L., et al. Sparse within sparse Gaussian processes using neighbor information. ICML 2021.

NNGP: Vecchia Approximation

- Apply the Vecchia approximation to prior $p(\overline{\mathbf{f}})$



neighbours of $\mathbf{x}_i$:

$$n(i) = \{j+3, j+4, j+5, j+6\}$$

neighbours of $\overline{\mathbf{x}}_{j+4}$ (selected by "looking backward"):

$$n(j+4) = \{j, j+1, j+2, j+3\}$$

- NNGP based on Vecchia approximation

$$p(f_i \mid \overline{\mathbf{f}}) \longrightarrow p(f_i \mid \overline{\mathbf{f}}_{n(i)})$$

$$q(\overline{\mathbf{f}}) \longrightarrow \text{mean-field } q(\overline{\mathbf{f}})$$

$$p(\overline{\mathbf{f}}) \longrightarrow p(\overline{f}_1) \prod_{j=2}^{M} p(\overline{f}_m \mid \overline{\mathbf{f}}_{n(j)})$$

- ELBO:

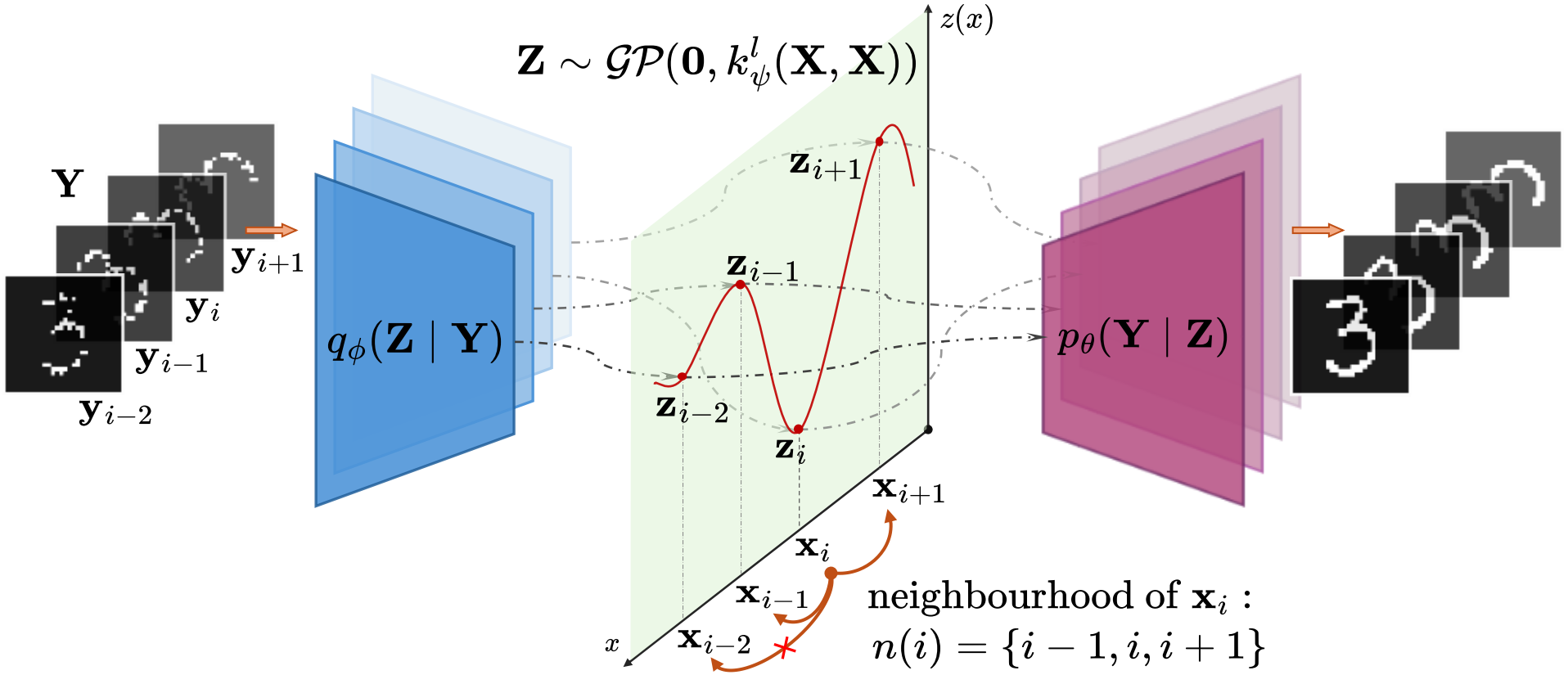
$$\frac{N}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\mathbb{E}_{q(f_i)} \log p(y_i \mid f_i) \right]$$

$$- \frac{M}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \mathbb{E}_{q(\overline{\mathbf{f}}_{n(j)})} \left\{ \mathrm{KL}[q(\overline{f}_j) \| p(\overline{f}_j \mid \overline{\mathbf{f}}_{n(j)})] \right\}$$

Wu, L., et al. Variational nearest neighbor Gaussian processes. ICML 2022.

NNGPVAE: Talk Only to Nearest Neighbours

- Inspired by the “first law of geography” in recent NNGP
- Focusing on a small set of nearest neighbours captures most of the essential correlation structures
- Place each inducing point over each data point



Shi, X., et al. Neighbour-driven Gaussian process variational autoencoders for scalable structured latent modelling. ICML 2025.

Two Neighbour-Driven Prior Approximations

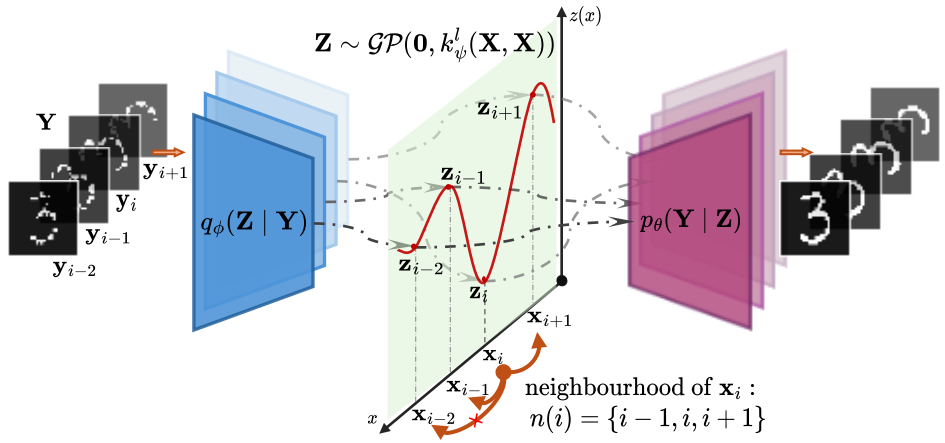
- Hierarchical Prior Approximation (HPA): sparsify covariance via hierarchical masks

$$\mathcal{L}_{\text{HPA}} \approx \frac{N}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left\{ \mathbb{E}_{q(\mathbf{z}_i|\mathbf{y}_i)} \left[\log p(\mathbf{y}_i \mid \mathbf{z}_i) \right] - \frac{1}{N} \text{KL} \left[q(\mathbf{Z}_{n(i)}) \mid\mid p(\mathbf{Z}_{n(i)}) \right] \right\}$$

- Sparse Precision Approximation (SPA): sparsify precision matrix via chained conditionals

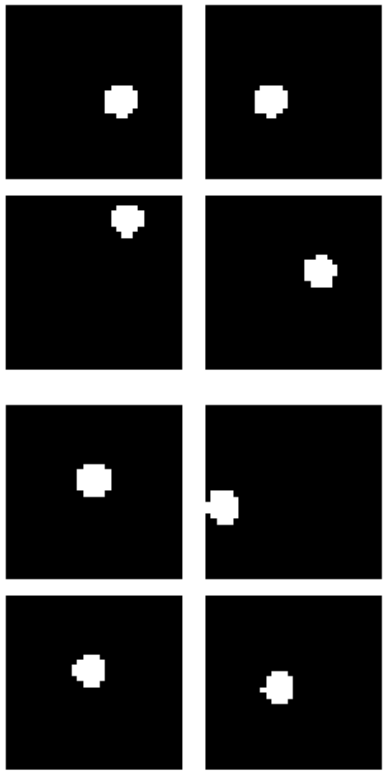
$$\mathcal{L}_{\text{SPA}} \approx \frac{N}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{y}_i)} \left[\log p(\mathbf{y}_i \mid \mathbf{z}_i) \right] - \frac{N}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \mathbb{E}_{q(\mathbf{Z}_{n(j)})} \text{KL} \left[q(\mathbf{z}_j) \mid\mid p(\mathbf{z}_j \mid \mathbf{Z}_{n(j)}) \right]$$

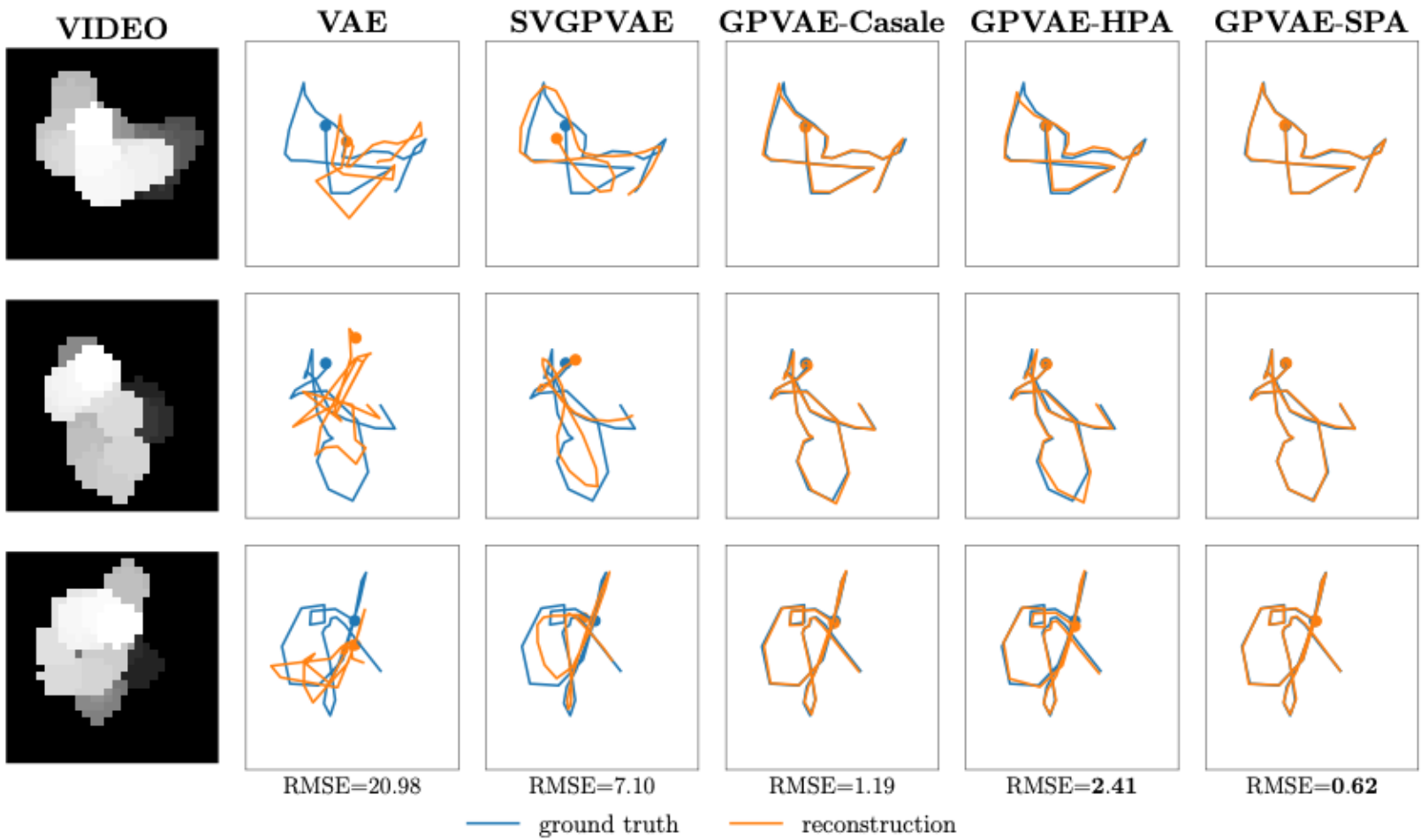
KL only involves H (<< N) nearest neighbouring variables!



Experiments: Latent Representation in Moving Ball

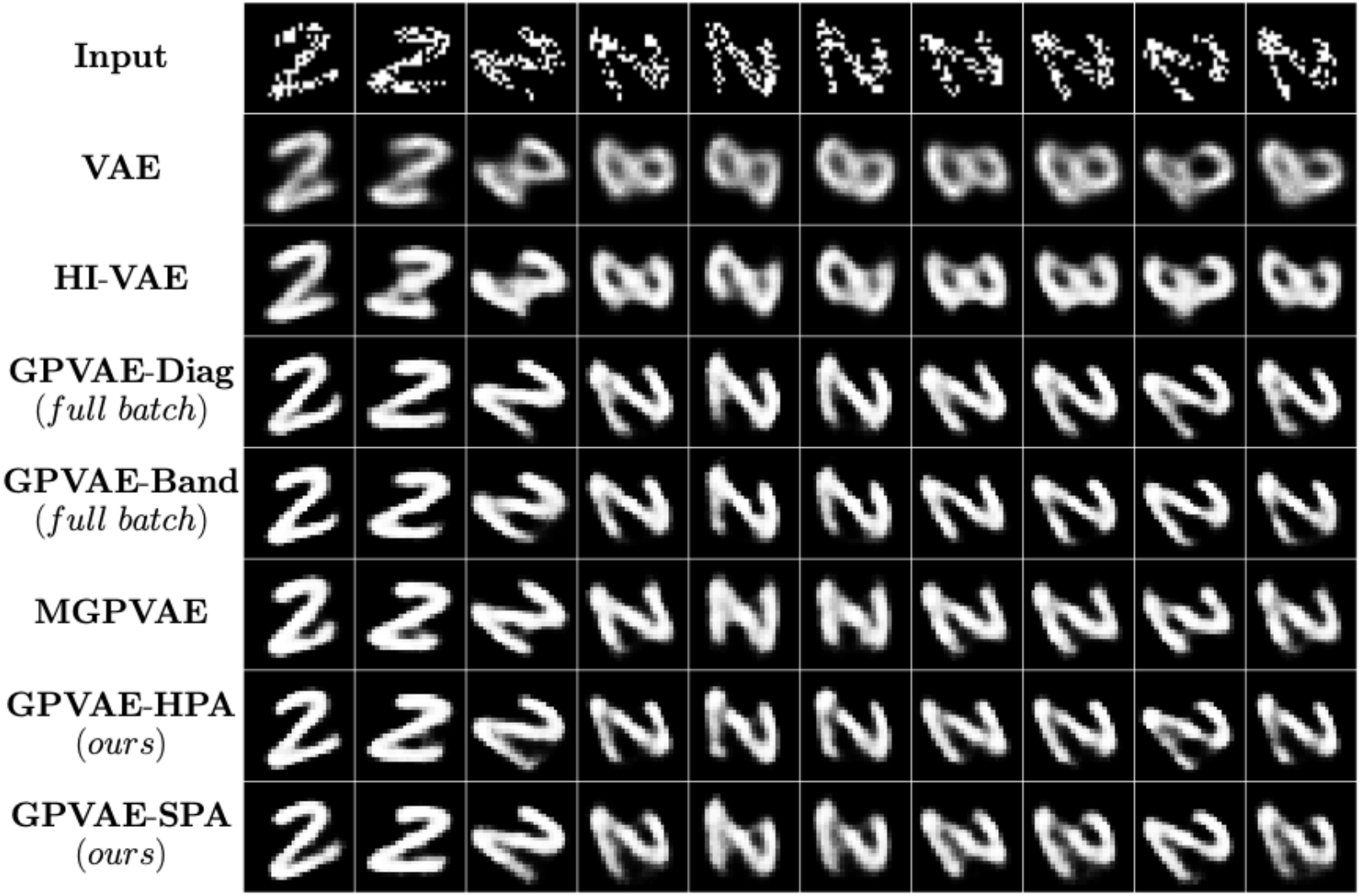
- Thousands of black-and-white videos of 30 frames





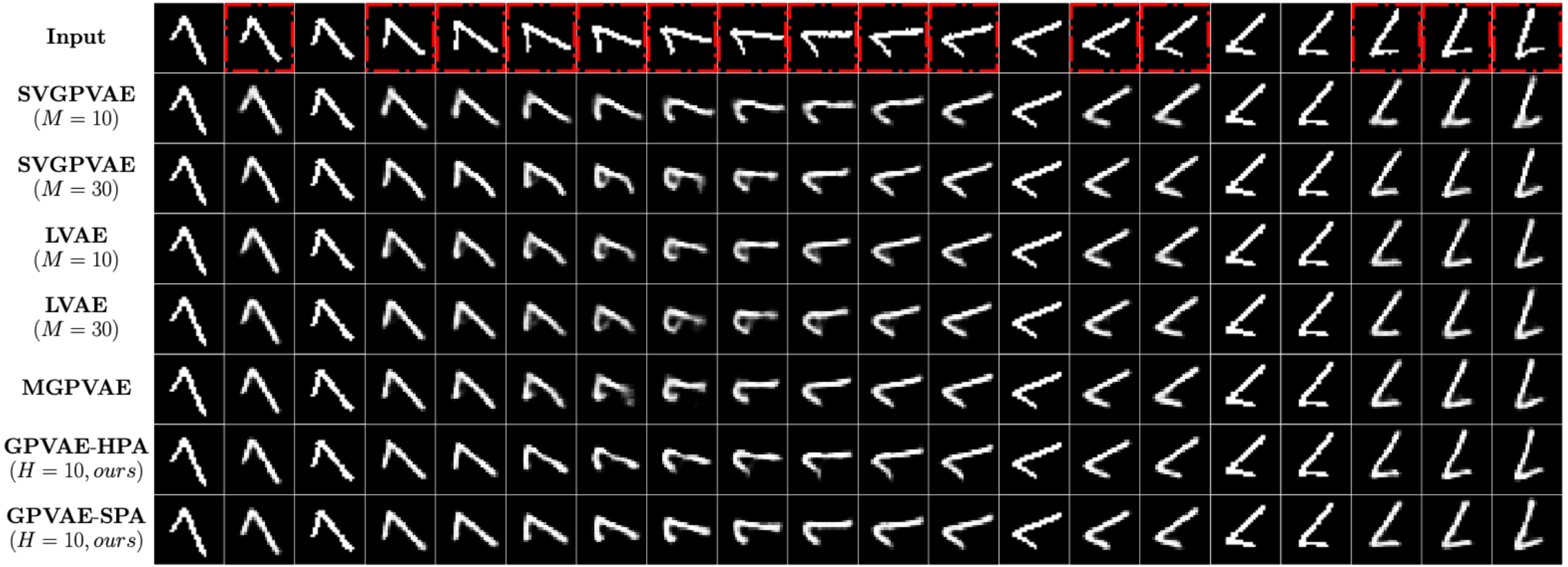
Experiments: Impute Corrupted Frames in Rotated MNIST

- 50,000 training sequences of 10 frames with 60% pixels absent



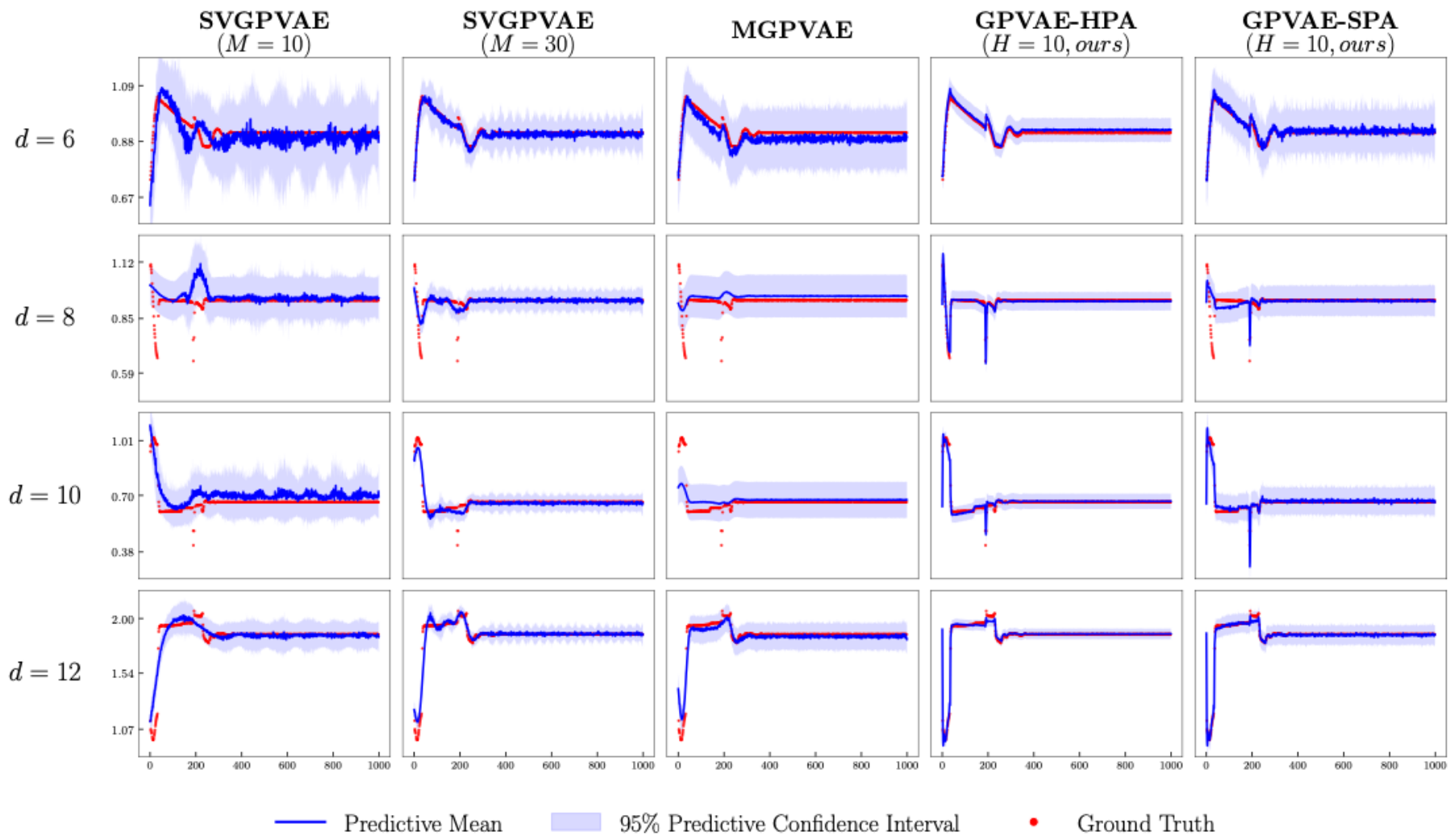
Experiments: Generate Missing Frames in Rotated MNIST

- 100 frames, each with a 0.6 probability of being dropped.



Experiments: Generate Missing States in MuJoCo

- 1000 timesteps, each with a 0.6 probability of being dropped



Experiments: Impute Missing Locations in SPE10

- ~10^5 locations of reservoir properties, high spatial heterogeneity

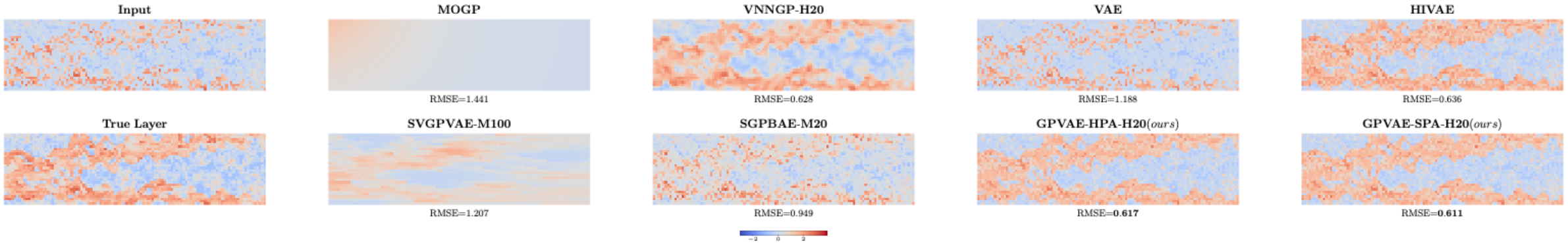


Table 3: Imputation results of various models on SPE10 dataset.

Models	NLL	RMSE	Training Time (s/epoch)
MOGP-M1000 (Hensman et al., 2013)	6.034 ± 0.003	1.230 ± 0.000	3.820 ± 0.021
VNNGP-H20 (Wu et al., 2022)	1.052 ± 0.000	0.683 ± 0.000	2.490 ± 0.084
VAE (Kingma & Welling, 2014)	3.436 ± 0.015	1.298 ± 0.002	**1.416 ± 0.020**
HI-VAE (Nazabal et al., 2020)	0.752 ± 0.034	0.647 ± 0.008	1.448 ± 0.011
SVGPVAE-M100 (Jazbec et al., 2021)	2.025 ± 0.020	0.979 ± 0.006	4.755 ± 0.097
SGPBAE-M20 (Tran et al., 2023)	1.335 ± 0.106	0.898 ± 0.125	1585.6 ± 21.3
GPVAE-HPA-H20	**0.735 ± 0.007**	**0.638 ± 0.007**	4.019 ± 0.036
GPVAE-SPA-H20	**0.736 ± 0.004**	**0.638 ± 0.004**	3.365 ± 0.117

Conclusion

- SVGP → SVGPVAE
  - Inducing variables for latent GPs
  - A reconstruction-minus-KL objective for scalable training
  - No special kernel structures

- NNGP → NNGPVAE
  - HPA & SPA: covariance & precision sparsity in GPVAE
  - No special kernel structures
  - Do not rely on large inducing points