Supplementary material for "Variational Zero-inflated Gaussian processes with sparse kernels"

Pashupati Hegde Markus Heinonen Samuel Kaski

Helsinki Institute for Information Technology HIIT
Department of Computer Science, Aalto University

In this supplementary paper we show in detail how we arrived at the evidence lower bounds for the zero-inflated Gaussian process, for the Gaussian process network, and for the sparse Gaussian process network. We also provide some additional details for multi-output prediction experiments with standard GPRN and Sparse GPRN models.

A) The stochastic variational bound of the zero-inflated GP

Here, we will derive the evidence lower bound (ELBO) of the zero-inflated Gaussian process. The joint distribution and priors for the model augmented with inducing points is defined as

$$p(\mathbf{y}, \mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)p(\mathbf{u}_f)p(\mathbf{u}_g) \tag{1}$$
$$p(\mathbf{f}|\mathbf{g}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ K_{fnn}) \tag{2}$$
$$p(\mathbf{g}) = \mathcal{N}(\mathbf{g}|\mathbf{0}, K_{gnn}) \tag{3}$$
$$p(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f|\mathbf{0}, K_{fmm}) \tag{4}$$
$$p(\mathbf{u}_g) = \mathcal{N}(\mathbf{u}_g|\mathbf{0}, K_{gmm}). \tag{5}$$

The the joint GP priors between latent and inducing functions can be written as below:

$$\begin{pmatrix}\mathbf{f}|\mathbf{g}\\ \mathbf{u}_f\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ K_{fnn} & \operatorname{diag}(\Phi(\mathbf{g}))K_{fnm}\\ K_{fmn}\operatorname{diag}(\Phi(\mathbf{g})) & K_{fmm}\end{bmatrix}\right) \tag{6}$$

$$\begin{pmatrix}\mathbf{g}\\ \mathbf{u}_g\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}K_{gnn} & K_{gnm}\\ K_{gmn} & K_{gmm}\end{bmatrix}\right) \tag{7}$$

Now, by conditioning latent functions on respective inducing variables, we arrive at following conditional distributions. The conditional $p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)$ is sparsified by latent $\mathbf{g}$ augmentation.

$$p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f) = \mathcal{N}(\mathbf{f}|\operatorname{diag}(\Phi(\mathbf{g}))K_{fnm}K_{fmm}^{-1}\mathbf{u}_f, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ \widetilde{K}_f) \tag{8}$$
$$p(\mathbf{g}|\mathbf{u}_g) = \mathcal{N}(\mathbf{g}|K_{gnm}K_{gmm}^{-1}\mathbf{u}_g, \widetilde{K}_g) \tag{9}$$
$$\widetilde{K}_f = K_{fnn} - K_{fnm}K_{fmm}^{-1}K_{fmn} \tag{10}$$
$$\widetilde{K}_g = K_{gnn} - K_{gnm}K_{gmm}^{-1}K_{gmn}. \tag{11}$$

For inference, we approximate true posterior $p(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g|\mathbf{y})$ with variational posterior of the form given

below

$$q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g) = p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_f)q(\mathbf{u}_g) \tag{12}$$

$$p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f) = \mathcal{N}(\mathbf{f}|\operatorname{diag}(\Phi(\mathbf{g}))K_{fnm}K_{fmm}^{-1}\mathbf{u}_f, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ \widetilde{K}_f) \tag{13}$$

$$p(\mathbf{g}|\mathbf{u}_g) = \mathcal{N}(\mathbf{g}|K_{gnm}K_{gmm}^{-1}\mathbf{u}_g, \widetilde{K}_g) \tag{14}$$

$$q(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f|\mathbf{m}_f, \mathbf{S}_f) \tag{15}$$

$$q(\mathbf{u}_g) = \mathcal{N}(\mathbf{u}_g|\mathbf{m}_g, \mathbf{S}_g) \tag{16}$$

and where $\mathbf{S}_f, \mathbf{S}_g \in \mathbb{R}^{m\times m}$ are square positive semi-definite matrices.

In variational inference we minimize the Kullback-Leibler divergence between the variational approximation $q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)$ and the true augmented joint distribution $p(\mathbf{y}, \mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)$:

$$\mathrm{KL}[q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)||p(\mathbf{y}, \mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)] = \int q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)}{q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)} d\mathbf{f}d\mathbf{g}d\mathbf{u}_f d\mathbf{u}_g \tag{17}$$

$$= \int q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g) \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)p(\mathbf{u}_f)p(\mathbf{u}_g)}{p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_f)q(\mathbf{u}_g)} d\mathbf{f}d\mathbf{g}d\mathbf{u}_f d\mathbf{u}_g \tag{18}$$

$$= \int q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g) \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u}_f)p(\mathbf{u}_g)}{q(\mathbf{u}_f)q(\mathbf{u}_g)} d\mathbf{f}d\mathbf{g}d\mathbf{u}_f d\mathbf{u}_g \tag{19}$$

$$= \iiiint p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_f)q(\mathbf{u}_g) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{u}_f d\mathbf{u}_g d\mathbf{g} d\mathbf{f} \tag{20}$$

$$- \underbrace{\int q(\mathbf{u}_f) \log q(\mathbf{u}_f) d\mathbf{u}_f}_{\mathrm{KL}[q(\mathbf{u}_f)||p(\mathbf{u}_f)]} - \underbrace{\int q(\mathbf{u}_g) \log q(\mathbf{u}_g) d\mathbf{u}_g}_{\mathrm{KL}[q(\mathbf{u}_g)||p(\mathbf{u}_g)]} \tag{21}$$

Following the derivation of Hensman et al. (2015), this corresponds to maximizing the evidence lower bound (ELBO):

$$\log p(\mathbf{y}) \geq \iiiint \log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)q(\mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_g) d\mathbf{u}_f d\mathbf{u}_g d\mathbf{g} d\mathbf{f} - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_g)||p(\mathbf{u}_f, \mathbf{u}_g)] \tag{22}$$

$$= \mathbb{E}_{q(\mathbf{f})} \log p(\mathbf{y}|\mathbf{f}) - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_g)||p(\mathbf{u}_f, \mathbf{u}_g)] \tag{23}$$

where we define

$$q(\mathbf{f}) = \iiint p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)q(\mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_g) d\mathbf{u}_f d\mathbf{u}_g d\mathbf{g}$$

$$= \int q(\mathbf{f}|\mathbf{g})q(\mathbf{g}) d\mathbf{g}, \tag{24}$$

where the variational approximations are tractably

$$q(\mathbf{g}) = \int p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_g) d\mathbf{u}_g \tag{25}$$

$$= \int \mathcal{N}(\mathbf{g}|K_{gnm}K_{gmm}^{-1}\mathbf{u}_g, \widetilde{K}_g)\mathcal{N}(\mathbf{u}_g|\mathbf{m}_g, \mathbf{S}_g) d\mathbf{u}_g \tag{26}$$

$$= \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}_g, \Sigma_g) \tag{27}$$

$$q(\mathbf{f}|\mathbf{g}) = \int p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)q(\mathbf{u}_f) d\mathbf{u}_f \tag{28}$$

$$= \int \mathcal{N}(\mathbf{f}|\operatorname{diag}(\Phi(\mathbf{g}))K_{fnm}K_{fmm}^{-1}\mathbf{u}_f, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ \widetilde{K}_f)\mathcal{N}(\mathbf{u}_f|\mathbf{m}_f, \mathbf{S}_f) d\mathbf{u}_f \tag{29}$$

$$= \mathcal{N}(\mathbf{f}|\operatorname{diag}(\Phi(\mathbf{g}))\boldsymbol{\mu}_f, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ \Sigma_f) \tag{30}$$

with

$$\boldsymbol{\mu}_f = K_{fnm}K_{fmm}^{-1}\mathbf{m}_f \tag{31}$$
$$\boldsymbol{\mu}_g = K_{gnm}K_{gmm}^{-1}\mathbf{m}_g \tag{32}$$
$$\Sigma_f = K_{fnn} + K_{fnm}K_{fmm}^{-1}(\mathbf{S}_f - K_{fmm})K_{fmm}^{-1}K_{fmn} \tag{33}$$
$$\Sigma_g = K_{gnn} + K_{gnm}K_{gmm}^{-1}(\mathbf{S}_g - K_{gmm})K_{gmm}^{-1}K_{gmn}. \tag{34}$$

The variational marginalizations $q(\mathbf{g})$ and $q(\mathbf{f}|)$ follow from standard Gaussian identities[1]. Substituting the variational marginalizations back to the ELBO results in

$$\log p(\mathbf{y}) \geq \int\int \log p(\mathbf{y}|\mathbf{f})q(\mathbf{f}|\mathbf{g})q(\mathbf{g})d\mathbf{f}d\mathbf{g} - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_g)||p(\mathbf{u}_f, \mathbf{u}_g)]. \tag{35}$$

Next, we marginalize the $\mathbf{f}$ from the ELBO. We additionally assume the likelihood $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f_i)$ factorises, which results in

$$\int_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f})q(\mathbf{f}|\mathbf{g})d\mathbf{f} = \int \log \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 I)q(\mathbf{f}|\mathbf{g})d\mathbf{f} \tag{36}$$
$$= \sum_{i=1}^{N} \int \log \mathcal{N}(y_i|f_i, \sigma_y^2)q(f_i|g_i)df_i \tag{37}$$
$$= \sum_{i=1}^{N} \log \mathcal{N}(y_i|\Phi(g_i)\mathbf{k}_{fi}^T K_{fmm}^{-1} m_{fi}, \sigma_y^2) - \frac{1}{2\sigma_y^2}\left\{\Phi(g_i)^2(\mathbf{k}_{fii} + \mathbf{k}_{fi}^T K_{fmm}^{-1}(\mathbf{S}_f - K_{fmm})K_{fmm}^{-1}\mathbf{k}_{fi})\right\} \tag{38}$$
$$= \sum_{i=1}^{N} \log \mathcal{N}(y_i|\Phi(g_i)\mu_{fi}, \sigma_y^2) - \frac{1}{2\sigma_y^2}\left\{\Phi(g_i)^2\sigma_{fi}^2\right\}, \tag{39}$$

where

$$\mu_{fi} = [\boldsymbol{\mu}_f]_i = \mathbf{k}_{fi}^T K_{fmm}^{-1} m_{fi} \tag{40}$$
$$\sigma_{fi}^2 = [\Sigma_f]_{ii} = \mathbf{k}_{fii} + \mathbf{k}_{fi}^T K_{fmm}^{-1}(\mathbf{S}_f - K_{fmm})K_{fmm}^{-1}\mathbf{k}_{fi}. \tag{41}$$

Substituting the above result into the ELBO results in

$$\mathbb{E}_{q(\mathbf{f})} \log p(\mathbf{y}|\mathbf{f}) = \int_{\mathbf{g}} q(\mathbf{g}) \int_{\mathbf{f}} q(\mathbf{f}|\mathbf{g}) \log p(\mathbf{y}|\mathbf{f})d\mathbf{f}d\mathbf{g} \tag{42}$$
$$= \int_{\mathbf{g}} \sum_{i=1}^{N} \log \mathcal{N}(y_i|\Phi(g_i)\mu_{fi}, \sigma_y^2) - \frac{1}{2\sigma_y^2}\left\{\Phi(g_i)^2\sigma_{fi}^2\right\}\ q(\mathbf{g})d\mathbf{g} \tag{43}$$
$$= \sum_{i=1}^{N} \int_{g_i} \log \mathcal{N}(y_i|\Phi(g_i)\mu_{fi}, \sigma_y^2)\ q(g_i)dg_i - \frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\int_{g_i}\left\{\Phi(g_i)^2\sigma_{fi}^2\right\}\ q(g_i)dg_i \tag{44}$$
$$= \sum_{i=1}^{N} \log \mathcal{N}(y_i|\langle\Phi(g_i)\rangle_{q(g_i)}\mu_{fi}, \sigma_y^2) - \frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\left\{Var[\Phi(g_i)](\mu_{fi})^2\right\} - \frac{1}{2\sigma_y^2}\sum_{i=1}^{N}\left\{\langle\Phi(g_i)^2\rangle_{q(g_i)}\sigma_{fi}^2\right\}. \tag{45}$$

The expectations $\langle . \rangle_{q(g_i)}$ of CDF transformation of a random variable with univariate Gaussian distribution.

[1] See for instance Bishop (2006): Pattern recognition and Machine learning, Springer, Section 2.3.

The analytical forms for these integrals can be written:

$$\langle \Phi(g_i) \rangle_{q(g_i)} = \int \Phi(g_i) q(g_i) dg_i \tag{46}$$

$$= \int \Phi(g_i) \mathcal{N}(g_i | \mu_{gi}, \sigma_{gi}^2) dg_i \tag{47}$$

$$= \Phi\left(\frac{\mu_{gi}}{\sqrt{1+\sigma_{gi}^2}} \right) \tag{48}$$

$$Var[\Phi(g_i)] = \int (\Phi(g_i) - \langle \Phi(g_i) \rangle_{q(g_i)})^2 q(g_i) dg_i \tag{49}$$

$$= \Phi\left(\frac{\mu_{gi}}{\sqrt{1+\sigma_{gi}^2}} \right) - 2T\left(\frac{\mu_{gi}}{\sqrt{1+\sigma_{gi}^2}}, \frac{1}{\sqrt{1+2\,\sigma_{gi}^2}} \right) - \Phi\left(\frac{\mu_{gi}}{\sqrt{1+\sigma_{gi}^2}} \right)^2 \tag{50}$$

$$\langle \Phi(g_i)^2 \rangle_{q(g_i)} = \int \Phi(g_i)^2 q(g_i) dg_i \tag{51}$$

$$= \Phi\left(\frac{\mu_{gi}}{\sqrt{1+\sigma_{gi}^2}} \right) - 2T\left(\frac{\mu_{gi}}{\sqrt{1+\sigma_{gi}^2}}, \frac{1}{\sqrt{1+2\,\sigma_{gi}^2}} \right) \tag{52}$$

where

$$\mu_{gi} = [\boldsymbol{\mu}_g]_i = \mathbf{k}_{gi}^T K_{gmm}^{-1} m_{gi} \tag{53}$$

$$\sigma_{gi}^2 = [\Sigma_g]_{ii} = K_{gii} + \mathbf{k}_{gi}^T K_{gmm}^{-1} (\mathbf{S}_g - K_{gmm}) K_{gmm}^{-1} \mathbf{k}_{gi}. \tag{54}$$

Owen's T function is defined as $T(h,a) = \phi(h) \int_0^a \frac{\phi(hx)}{1+x^2} dx$.

The final evidence lower bound with the Kullback-Leibler terms is

$$p(\mathbf{y}) \geq \sum_{i=1}^{N} \left\{ \log \mathcal{N}(y_i | \langle \Phi(g_i) \rangle_{q(g_i)} \mu_{fi}, \sigma_y^2) - \frac{1}{2\sigma_y^2} \left(Var[\Phi(g_i)]\, \mu_{fi}^2 + \langle \Phi(g_i)^2 \rangle_{q(g_i)}\, \sigma_{fi}^2 \right) \right\} \tag{55}$$

$$- \left\{ \frac{1}{2} \log |K_{fmm}| - \frac{1}{2} \log |\mathbf{S}_f| + \frac{1}{2} Tr\left[(\mathbf{m}_f {\mathbf{m}_f}^T + \mathbf{S}_f) K_{fmm}^{-1} \right] - \frac{m}{2} \right\} \tag{56}$$

$$- \left\{ \frac{1}{2} \log |K_{gmm}| - \frac{1}{2} \log |\mathbf{S}_g| + \frac{1}{2} Tr\left[(\mathbf{m}_g {\mathbf{m}_g}^T + \mathbf{S}_g) K_{gmm}^{-1} \right] - \frac{m}{2} \right\} \tag{57}$$

$$= \mathcal{L}_{\text{ZIGP}} \tag{58}$$

B) The stochastic variational bound of the Gaussian process network

In GPRN a vector-valued output function $\mathbf{y}(\mathbf{x}) \in \mathbb{R}^P$ with P outputs is modeled using vector-valued latent functions $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^Q$ with Q latent values and mixing weights $W(\mathbf{x}) \in \mathbb{R}^{P \times Q}$ as

$$\mathbf{y}(x) = W(x)[\mathbf{f}(x) + \boldsymbol{\epsilon}] + \boldsymbol{\varepsilon}, \tag{59}$$

where for all $q = 1, \ldots, Q$ and $p = 1, \ldots, P$ we assume GP priors and additive zero-mean noises,

$$f_q(\mathbf{x}) \sim \mathcal{GP}(0, K_f(\mathbf{x}, \mathbf{x}')) \tag{60}$$

$$W_{qp}(\mathbf{x}) \sim \mathcal{GP}(0, K_w(\mathbf{x}, \mathbf{x}')) \tag{61}$$

$$\epsilon_q \sim \mathcal{N}(0, \sigma_f^2) \tag{62}$$

$$\varepsilon_p \sim \mathcal{N}(0, \sigma_y^2). \tag{63}$$

The subscripts are used to denote individual components of $\mathbf{f}$ and W with p and q indicating p^{th} output dimension and q^{th} latent dimension respectively.

We begin by introducing the inducing variable augmentation for latent functions $\mathbf{f}(\mathbf{x})$ and mixing weights $W(\mathbf{x})$ with $\mathbf{u}_f, \mathbf{z}_f = \{\mathbf{u}_{f_q}, \mathbf{z}_{f_q}\}_{q=1}^{Q}$ and $\mathbf{u}_w, \mathbf{z}_w = \{\mathbf{u}_{w_{qp}}, \mathbf{z}_{w_{qp}}\}_{q,p=1}^{Q,P}$:

$$p(\mathbf{y}, \mathbf{f}, W, \mathbf{u}_f, \mathbf{u}_w) = p(\mathbf{y}|\mathbf{f}, W)p(\mathbf{f}|\mathbf{u}_f)p(W|\mathbf{u}_w)p(\mathbf{u}_f)p(\mathbf{u}_w) \tag{64}$$

$$p(\mathbf{f}|\mathbf{u}_f) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_q|Q_{f_q}\mathbf{u}_{f_q}, \widetilde{K}_{f_q}) \tag{65}$$

$$p(W|\mathbf{u}_w) = \prod_{q,p=1}^{Q,P} \mathcal{N}(\mathbf{w}_{qp}|Q_{w_{qp}}\mathbf{u}_{w_{qp}}, \widetilde{K}_{w_{qp}}) \tag{66}$$

$$p(\mathbf{u}_f) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{u}_{f_q}|\mathbf{0}, K_{f_q,mm}) \tag{67}$$

$$p(\mathbf{u}_w) = \prod_{q,p=1}^{Q,P} \mathcal{N}(\mathbf{u}_{w_{qp}}|\mathbf{0}, K_{w_{qp},mm}), \tag{68}$$

where we have separate kernels K and extrapolation matrices Q for each component of $W(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})$ that are of the form as given below:

$$Q_f = K_{f_q nm} K_{f_q mm}^{-1} \tag{69}$$

$$Q_g = K_{w_{qp} nm} K_{w_{qp} mm}^{-1} \tag{70}$$

$$\widetilde{K}_f = K_{f_q nn} - K_{f_q nm} K_{f_q mm}^{-1} K_{f_q mn} \tag{71}$$

$$\widetilde{K}_{w_{qp}} = K_{w_{qp} nn} - K_{w_{qp} nm} K_{w_{qp} mm}^{-1} K_{w_{qp} mn}. \tag{72}$$

Following the variational inference framework, we define the variational joint distribution as,

$$q(\mathbf{f}, W, \mathbf{u}_f, \mathbf{u}_w) = p(\mathbf{f}|\mathbf{u}_f)p(W|\mathbf{u}_w)q(\mathbf{u}_f)q(\mathbf{u}_w) \tag{73}$$

$$q(\mathbf{u}_{f_q}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{u}_{f_q}|\mathbf{m}_{f_q}\mathbf{S}_{f_q}) \tag{74}$$

$$q(\mathbf{u}_{w_{qp}}) = \prod_{q,p=1}^{Q,P} \mathcal{N}(\mathbf{u}_{w_{qp}}|\mathbf{m}_{w_{qp}}, \mathbf{S}_{w_{qp}}), \tag{75}$$

where $\mathbf{u}_{w_{qp}}$ and $\mathbf{u}_{f_q}$ indicate the inducing points for functions $W_{qp}(\mathbf{x})$ and $f_q(\mathbf{x})$, respectively. The ELBO can be now stated as

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f},W,\mathbf{u}_f,\mathbf{u}_w)} \log p(\mathbf{y}|\mathbf{f}, W) - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_w)||p(\mathbf{u}_f, \mathbf{u}_w)] \tag{76}$$

$$= \iiiint q(\mathbf{f}, W, \mathbf{u}_f, \mathbf{u}_w) \log p(\mathbf{y}|\mathbf{f}, W) d\mathbf{f} dW d\mathbf{u}_f d\mathbf{u}_w - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_w)||p(\mathbf{u}_f, \mathbf{u}_w)] \tag{77}$$

Since the variational joint posterior decomposes as equation (73), we begin by marginalizing the inducing distributions $\mathbf{u}_f$ and $\mathbf{u}_w$,

$$\iiiint q(\mathbf{f}, W, \mathbf{u}_f, \mathbf{u}_w) d\mathbf{f} dW d\mathbf{u}_f d\mathbf{u}_w = \int_{\mathbf{f}} \int_{\mathbf{u}_f} p(\mathbf{f}|\mathbf{u}_f)q(\mathbf{u}_f) d\mathbf{f} d\mathbf{u}_f \int_W \int_{\mathbf{u}_w} p(W|\mathbf{u}_w)q(\mathbf{u}_w) dW d\mathbf{u}_w \tag{78}$$

$$= \int_{\mathbf{f}} q(\mathbf{f}) d\mathbf{f} \int_W q(W) dW \tag{79}$$

where

$$q(\mathbf{f}) = \int p(\mathbf{f}|\mathbf{u}_f)q(\mathbf{u}_f)d\mathbf{u}_f \tag{80}$$

$$= \prod_{q=1}^{Q} \int \mathcal{N}(\mathbf{f}_q|K_{f_q nm}K_{f_q mm}^{-1}\mathbf{u}_{f_q}, \widetilde{K}_{f_q})\mathcal{N}(\mathbf{u}_{f_q}|\mathbf{m}_{f_q}, \mathbf{S}_{f_q})d\mathbf{u}_{f_q} \tag{81}$$

$$= \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_q|\boldsymbol{\mu}_{f_q}, \Sigma_{f_q}) \tag{82}$$

$$q(W) = \int p(W|\mathbf{u}_w)q(\mathbf{u}_w)d\mathbf{u}_w \tag{83}$$

$$= \prod_{q,p=1}^{Q,P} \int \mathcal{N}(W_{qp}|K_{w_{qp}nm}K_{w_{qp}mm}^{-1}\mathbf{u}_{w_{qp}}, \widetilde{K}_{w_{qp}})\mathcal{N}(\mathbf{u}_{w_{qp}}|\mathbf{m}_{w_{qp}}, \mathbf{S}_{w_{qp}})d\mathbf{u}_{w_{qp}} \tag{84}$$

$$= \prod_{q,p=1}^{Q,P} \mathcal{N}(W_{qp}|\boldsymbol{\mu}_{w_{qp}}, \Sigma_{w_{qp}}) \tag{85}$$

with

$$\boldsymbol{\mu}_{f_q} = K_{f_q nm}K_{f_q mm}^{-1}\mathbf{m}_{f_q} \tag{86}$$

$$\boldsymbol{\mu}_{w_{qp}} = K_{w_{qp}nm}K_{w_{qp}mm}^{-1}\mathbf{m}_{w_{qp}} \tag{87}$$

$$\Sigma_{f_q} = K_{f_q nn} + K_{f_q nm}K_{f_q mm}^{-1}(\mathbf{S}_{f_q} - K_{f_q mm})K_{f_q mm}^{-1}K_{f_q mn} \tag{88}$$

$$\Sigma_{w_{qp}} = K_{w_{qp}nn} + K_{w_{qp}nm}K_{w_{qp}mm}^{-1}(\mathbf{S}_{w_{qp}} - K_{w_{qp}mm})K_{w_{qp}mm}^{-1}K_{w_{qp}mn}. \tag{89}$$

Since the noise term $\boldsymbol{\varepsilon}$ is assumed to be isotropic Gaussian, density $p(\mathbf{y}|W, \mathbf{f})$ factorises across all target observations and dimensions. The expectation term in the ELBO then reduces to,

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(W)}\mathbb{E}_{q(\mathbf{f})} \log p(\mathbf{y}|\mathbf{f}, W) \tag{90}$$

$$= \sum_{i,p=1}^{N,P} \iint \log \mathcal{N}(y_{p,i}|\mathbf{w}_{p,i}^T\mathbf{f}_i, \varepsilon_p^2)q(\mathbf{f}_i, \mathbf{w}_{p,i})d\mathbf{w}_{p,i}d\mathbf{f}_i - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_w)||p(\mathbf{u}_f, \mathbf{u}_w)]. \tag{91}$$

The integral with respect to $\mathbf{f}$ can be now solved as

$$\int \log \mathcal{N}(y_{p,i}|\mathbf{w}_{p,i}^T\mathbf{f}_i, \varepsilon_p^2)q(\mathbf{f}_i)d\mathbf{f}_i = \log \mathcal{N}(y_{p,i}|\mathbf{w}_{p,i}^T\boldsymbol{\mu}_{f_i}, \varepsilon_p^2) - \frac{1}{2\varepsilon_p^2}Tr\big[\mathbf{w}_{p,i}^T\Sigma_{f_i}\mathbf{w}_{p,i}\big] \tag{92}$$

$$= \log \mathcal{N}(y_{p,i}|\mathbf{w}_{p,i}^T\boldsymbol{\mu}_{f_i}, \varepsilon_p^2) - \frac{1}{2\varepsilon_p^2}Tr\big[\Sigma_{f_i}\mathbf{w}_{p,i}\mathbf{w}_{p,i}^T\big]. \tag{93}$$

Next we can marginalize W from the above terms,

$$\int \log \mathcal{N}(y_{p,i}|\mathbf{w}_{p,i}^T\boldsymbol{\mu}_{f_i}, \varepsilon_p^2)q(\mathbf{w}_{p,i})d\mathbf{w}_{p,i} = \log \mathcal{N}(y_{p,i}|\boldsymbol{\mu}_{w_{p,i}}^T\boldsymbol{\mu}_{f_i}, \varepsilon_p^2) - \frac{1}{2\varepsilon_p^2}Tr\big[\boldsymbol{\mu}_{f_i}^T\Sigma_{w_{q,i}}\boldsymbol{\mu}_{f_i}\big] \tag{94}$$

$$= \log \mathcal{N}(y_{p,i}|\boldsymbol{\mu}_{w_{p,i}}^T\boldsymbol{\mu}_{f_i}, \varepsilon_p^2) - \frac{1}{2\varepsilon_p^2}\sum_{q=1}^{Q}\mu_{f_q,i}^2\sigma_{w_{qp},i}^2 \tag{95}$$

$$\int \frac{1}{2\varepsilon_p^2}Tr\big[\Sigma_{f_i}\mathbf{w}_{p,i}\mathbf{w}_{p,i}^T\big]q(\mathbf{w}_{p,i})d\mathbf{w}_{p,i} = \frac{1}{2\varepsilon_p^2}Tr\big[\Sigma_{f_i}(\boldsymbol{\mu}_{w_{p,i}}\boldsymbol{\mu}_{w_{p,i}}^T + \Sigma_{w_{q,i}})\big] \tag{96}$$

$$= \frac{1}{2\varepsilon_p^2}\sum_{q=1}^{Q}\left(\mu_{w_{qp},i}^2\sigma_{f_q,i}^2 + \sigma_{w_{qp},i}^2\sigma_{f_q,i}^2\right). \tag{97}$$

Finally, adding the above results across all N observations and response dimensions P along with Gaussian KL divergence terms, we get the final lowerbound:

$$\log p(\mathbf{y}) \geq \sum_{i=1}^{N} \left\{ \sum_{p=1}^{P} \log \mathcal{N}\Big(y_{p,i} | \sum_{q=1}^{Q} \mu_{w_{qp},i} \mu_{f_q,i}, \varepsilon_p^2\Big) - \frac{1}{2\varepsilon_p^2} \sum_{q,p=1}^{Q,P} \Big(\mu_{w_{qp},i}^2 \sigma_{f_q,i}^2 + \mu_{f_q,i}^2 \sigma_{w_{qp},i}^2 + \sigma_{w_{qp},i}^2 \sigma_{f_q,i}^2\Big) \right\} \tag{98}$$

$$- \sum_{q,p}^{Q,P} \mathrm{KL}[q(\mathbf{u}_{w_{qp}}, \mathbf{u}_{f_q}) || p(\mathbf{u}_{w_{qp}}, \mathbf{u}_{f_q})] \tag{99}$$

$$= \mathcal{L}_{\text{GPRN}}, \tag{100}$$

where $\mu_{f_q,i}$ is the i'th element of $\boldsymbol{\mu}_{f_q}$ and $\sigma_{f_q,i}^2$ is the i'th diagonal element of Σ_{f_q} (similarly for W_{qp}'s).

C) The stochastic variational bound of the sparse Gaussian process network

Sparse GPRN is a modification to standard GPRN where sparsity is added to the mixing matrix components. This corresponds to the p'th output being a sparse mixture of the latent Q functions, i.e. it can effectively use any subset of the Q latent dimensions by having zeros in the mixing functions. The joint distribution for the model can be written as,

$$p(\mathbf{y}, \mathbf{f}, W, \mathbf{g}) = p(\mathbf{y}|\mathbf{f}, W) p(\mathbf{f}) p(W|\mathbf{g}) p(\mathbf{g}), \tag{101}$$

where all individual components of latent function $\mathbf{f}$ and mixing matrix W are given GP priors. We encode the sparsity terms $\mathbf{g}$ for all $Q \times P$ mixing functions $W_{qp}(\mathbf{x})$ functions as

$$p(W_{qp}|\mathbf{g}_{qp}) = \mathcal{N}(\mathbf{w}_{qp}|\mathbf{0}, \Phi(\mathbf{g}_{qp})\Phi(\mathbf{g}_{qp})^T \circ K_w). \tag{102}$$

To introduce variational inference, the joint model is augmented with three sets of inducing variables for $\mathbf{f}$, W and $\mathbf{g}$. After marginalizing out the inducing variables similar to SVI for standard GPRN, the lower bund for marginal likelihood can be written as

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f},W,\mathbf{g})} \log p(\mathbf{y}|\mathbf{f}, W) - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_w, \mathbf{u}_g) || p(\mathbf{u}_f, \mathbf{u}_w, \mathbf{u}_g)]. \tag{103}$$

Where the joint distribution in the variational expectation factorizes as $q(\mathbf{f}, W, \mathbf{g}) = q(\mathbf{f})q(W|\mathbf{g})q(\mathbf{g})$. The variational posterior after marginalizing inducing variables is written as,

$$q(\mathbf{f}) = \int q(\mathbf{f}|\mathbf{u}_f) q(\mathbf{u}_f) d\mathbf{u}_f \tag{104}$$

$$= \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_q | \boldsymbol{\mu}_{f_q}, \Sigma_{f_q}) \tag{105}$$

$$q(W) = \int q(W|\mathbf{u}_w) q(\mathbf{u}_w) d\mathbf{u}_w \tag{106}$$

$$= \prod_{q,p=1}^{Q,P} \mathcal{N}(W_{qp} | \boldsymbol{\mu}_{w_{qp}}, \Sigma_{w_{qp}}) \tag{107}$$

$$q(\mathbf{g}) = \int q(\mathbf{g}|\mathbf{u}_g) q(\mathbf{u}_g) d\mathbf{u}_g \tag{108}$$

$$= \prod_{q,p=1}^{Q,P} \mathcal{N}(g_{qp} | \boldsymbol{\mu}_{g_{qp}}, \Sigma_{g_{qp}}) \tag{109}$$

with

$$\boldsymbol{\mu}_{f_q} = K_{f_q nm} K^{-1}_{f_q mm} \mathbf{m}_{f_q} \tag{110}$$

$$\boldsymbol{\mu}_{w_{qp}} = K_{w_{qp} nm} K^{-1}_{w_{qp} mm} \mathbf{m}_{w_{qp}} \tag{111}$$

$$\boldsymbol{\mu}_{g_{qp}} = K_{g_{qp} nm} K^{-1}_{g_{qp} mm} \mathbf{m}_{g_{qp}} \tag{112}$$

$$\Sigma_{f_q} = K_{f_q nn} + K_{f_q nm} K^{-1}_{f_q mm} (\mathbf{S}_{f_q} - K_{f_q mm}) K^{-1}_{f_q mm} K_{f_q mn} \tag{113}$$

$$\Sigma_{w_{qp}} = K_{w_{qp} nn} + K_{w_{qp} nm} K^{-1}_{w_{qp} mm} (\mathbf{S}_{w_{qp}} - K_{w_{qp} mm}) K^{-1}_{w_{qp} mm} K_{w_{qp} mn} \tag{114}$$

$$\Sigma_{g_{qp}} = K_{g_{qp} nn} + K_{g_{qp} nm} K^{-1}_{g_{qp} mm} (\mathbf{S}_{g_{qp}} - K_{g_{qp} mm}) K^{-1}_{g_{qp} mm} K_{g_{qp} mn}. \tag{115}$$

Similar to standard GPRN, with the isotropic Gaussian, density $p(\mathbf{y}|W, \mathbf{f})$ factorizes across all target observations and dimensions. The expectation term in the ELBO then reduces to

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(W|\mathbf{g})} \mathbb{E}_{q(\mathbf{f})} \mathbb{E}_{q(\mathbf{g})} \log p(\mathbf{y}|\mathbf{f}, W) \tag{116}$$

$$= \sum_{i,p=1}^{N,P} \iiint \log \mathcal{N}(y_{p,i}|(\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})^T \mathbf{f}_i, \varepsilon_p^2) q(\mathbf{f}_i) q(\mathbf{w}_{p,i}|q(\mathbf{g}_{p,i}) q(\mathbf{g}_{p,i}) d\mathbf{w}_{p,i} d\mathbf{f}_i d\mathbf{g}_i - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_w)||p(\mathbf{u}_f, \mathbf{u}_w)]. \tag{117}$$

The integral with respect to $\mathbf{f}$ can be now solved as

$$\int \log \mathcal{N}(y_{p,i}|(\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})^T \mathbf{f}_i, \varepsilon_p^2) q(\mathbf{f}_i) d\mathbf{f}_i = \log \mathcal{N}(y_{p,i}|(\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})^T \boldsymbol{\mu}_{f_i}, \varepsilon_p^2) - \frac{1}{2\varepsilon_p^2} Tr\big[(\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})^T \Sigma_{f_i} (\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})\big] \tag{118}$$

$$= \log \mathcal{N}(y_{p,i}|(\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})^T \boldsymbol{\mu}_{f_i}, \varepsilon_p^2) - \frac{1}{2\varepsilon_p^2} Tr\big[\Sigma_{f_i} (\mathbf{g}_{p,i} \mathbf{g}_{p,i}^T \circ \mathbf{w}_{p,i} \mathbf{w}_{p,i}^T)\big]. \tag{119}$$

Next, by integrating individual terms with respect to W we get

$$\int \log \mathcal{N}(y_{p,i}|(\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})^T \boldsymbol{\mu}_{f_i}, \varepsilon_p^2) q(\mathbf{w}_{p,i}) d\mathbf{w}_{p,i} = \log \mathcal{N}(y_{p,i}|(\boldsymbol{\mu}_{w_{p,i}} \circ \mathbf{g}_{p,i})^T \boldsymbol{\mu}_{f_i}, \varepsilon_p^2) - \frac{1}{2\varepsilon_p^2} Tr\big[\boldsymbol{\mu}_{f_i}^T (\mathbf{g}_{p,i} \mathbf{g}_{p,i}^T \circ \Sigma_{w_{q,i}}) \boldsymbol{\mu}_{f_i}\big] \tag{120}$$

$$\int \frac{1}{2\varepsilon_p^2} Tr\big[\Sigma_{f_i} (\mathbf{g}_{p,i} \mathbf{g}_{p,i}^T \circ \mathbf{w}_{p,i} \mathbf{w}_{p,i}^T)\big] = \frac{1}{2\varepsilon_p^2} Tr\big[\Sigma_{f_i} (\mathbf{g}_{p,i} \mathbf{g}_{p,i}^T \circ (\boldsymbol{\mu}_{w_{p,i}} \boldsymbol{\mu}_{w_{p,i}}^T + \Sigma_{w_{q,i}}))\big] \tag{121}$$

Finally, integrating all the above terms with respect to $\mathbf{g}$, we get

$$\int \log \mathcal{N}(y_{p,i}|(\boldsymbol{\mu}_{w_{p,i}} \circ \mathbf{g}_{p,i})^T \boldsymbol{\mu}_{f_i}, \varepsilon_p^2) q(\mathbf{g}_{p,i}) d\mathbf{g}_{p,i} = \log \mathcal{N}(y_{p,i}|(\boldsymbol{\mu}_{w_{p,i}} \circ \langle \Phi(\mathbf{g}_{p,i}) \rangle)^T \boldsymbol{\mu}_{f_i}, \varepsilon_p^2) \tag{122}$$

$$- \frac{1}{2\varepsilon_p^2} Tr\big[\boldsymbol{\mu}_{f_i}^T (\boldsymbol{\mu}_{w_{p,i}} \boldsymbol{\mu}_{w_{p,i}}^T \circ Var[\Phi(\mathbf{g}_{p,i})]) \boldsymbol{\mu}_{f_i}\big] \tag{123}$$

$$= \log \mathcal{N}\Big(y_{p,i}|\sum_{q=1}^{Q} \mu_{w_{qp},i} \mu_{g_{qp},i} \mu_{f_q,i}, \varepsilon_p^2\Big) \tag{124}$$

$$- \frac{1}{2\varepsilon_p^2} \sum_{q=1}^{Q} \Big(\sigma_{g_{qp},i}^2 \mu_{f_q,i}^2 \mu_{w_{qp},i}^2\Big) \tag{125}$$

$$\int \frac{1}{2\varepsilon_p^2} Tr\big[\boldsymbol{\mu}_{f_i}^T (\mathbf{g}_{p,i} \mathbf{g}_{p,i}^T \circ \Sigma_{w_{q,i}}) \boldsymbol{\mu}_{f_i}\big] q(\mathbf{g}_{p,i}) d\mathbf{g}_{p,i} = \frac{1}{2\varepsilon_p^2} Tr\big[\boldsymbol{\mu}_{f_i}^T \left(\langle \Phi(\mathbf{g}_{p,i}) \rangle \langle \Phi(\mathbf{g}_{p,i}) \rangle^T + Var[\Phi(\mathbf{g}_{p,i})]\right) \circ \Sigma_{w_{q,i}}) \boldsymbol{\mu}_{f_i}\big] \tag{126}$$

$$= \frac{1}{2\varepsilon_p^2} \sum_{q=1}^{Q} \Big((\mu_{g_{qp},i}^2 + \sigma_{g_{qp},i}^2) \mu_{f_q,i}^2 \sigma_{w_{qp},i}^2\Big) \tag{127}$$

$$\int \frac{1}{2\varepsilon_p^2} Tr\big[\Sigma_{f_i} (\mathbf{g}_{p,i} \mathbf{g}_{p,i}^T \circ (\boldsymbol{\mu}_{w_{p,i}} \boldsymbol{\mu}_{w_{p,i}}^T + \Sigma_{w_{q,i}}))\big] q(\mathbf{g}_{p,i}) d\mathbf{g}_{p,i} = \frac{1}{2\varepsilon_p^2} Tr\big[\Sigma_{f_i} \Big(\left(\langle \Phi(\mathbf{g}_{p,i}) \rangle \langle \Phi(\mathbf{g}_{p,i}) \rangle^T + Var[\Phi(\mathbf{g}_{p,i})]\right) \circ \boldsymbol{\mu}_{w_{p,i}} \boldsymbol{\mu}_{w_{p,i}}^T\Big)\big] \tag{128}$$

$$+ \frac{1}{2\varepsilon_p^2} Tr\big[\Sigma_{f_i} \left(\left(\langle \Phi(\mathbf{g}_{p,i}) \rangle \langle \Phi(\mathbf{g}_{p,i}) \rangle^T + Var[\Phi(\mathbf{g}_{p,i})]\right) \circ \Sigma_{w_{q,i}}\right)\big] \tag{129}$$

$$= \frac{1}{2\varepsilon_p^2} \sum_{q=1}^{Q} \Big((\mu_{g_{qp},i}^2 + \sigma_{g_{qp},i}^2)(\mu_{w_{qp},i}^2 \sigma_{f_q,i}^2 + \sigma_{w_{qp},i}^2 \sigma_{f_q,i}^2)\Big). \tag{130}$$

Adding above results across all the observations N and output dimensions P, we retrieve the final evidence lower bound

$$p(\mathbf{y}) \geq \sum_{i=1}^{N} \Bigg\{ \sum_{p=1}^{P} \log \mathcal{N}\Big(y_{p,i}|\sum_{q=1}^{Q} \mu_{w_{qp},i} \mu_{g_{qp},i} \mu_{f_q,i}, \varepsilon_p^2\Big) \tag{131}$$

$$- \sum_{q,p=1}^{Q,P} \Big((\mu_{g_{qp},i}^2 + \sigma_{g_{qp},i}^2) \cdot (\mu_{w_{qp},i}^2 \sigma_{f_q,i}^2 + \mu_{f_q,i}^2 \sigma_{w_{qp},i}^2 + \sigma_{w_{qp},i}^2 \sigma_{f_q,i}^2)\Big) - \sum_{q,p=1}^{Q,P} \Big(\sigma_{g_{qp},i}^2 \mu_{f_q,i}^2 \sigma_{w_{qp},i}^2\Big) \Bigg\} \tag{132}$$

$$- \sum_{q,p}^{Q,P} \mathrm{KL}[q(\mathbf{u}_{f_q}, \mathbf{u}_{w_{qp}}, \mathbf{u}_{g_{qp}})||p(\mathbf{u}_{f_q}, \mathbf{u}_{w_{qp}}, \mathbf{u}_{g_{qp}})] \tag{133}$$

$$= \mathcal{L}_{s\text{GPRN}}. \tag{134}$$

D) Results for multi-output prediction experiments

JURA

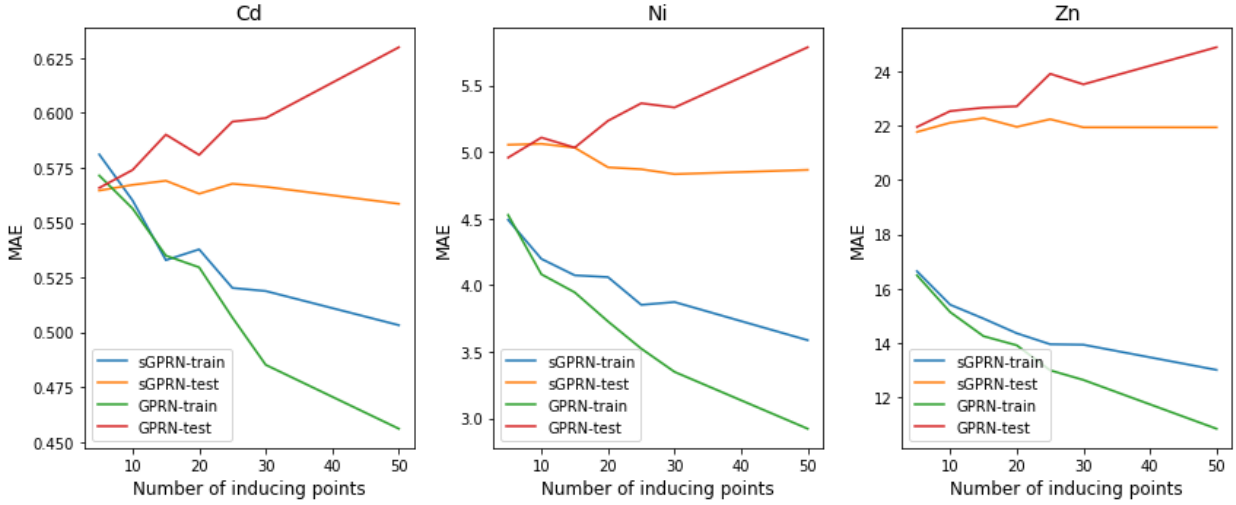

Figure 1: Mean Absolute Error across varying number of inducing points for sparse GPRN and standard GPRN models on Jura dataset. All numbers are average over 30 repetitions. It is surprising that with increase in number of inducing points, the train error decreases, whereas the test error increases. However, Sparse GPRN manages to achieve lower test error because of additional regularization due to sparse latent space.

Table 1: Results for the Jura dataset for sparse GPRN and standard GPRN models on test data. Numbers inside the bracket indicate one standard-error over 30 random initialization.

		Cadmium		Nickel		Zinc	
Model	m	RMSE	MAE	RMSE	MAE	RMSE	MAE
GPRN	5	0.724(0.008)	0.566(0.007)	6.469(0.161)	4.958(0.157)	33.729(0.500)	21.959(0.371)
	10	0.736(0.017)	0.574(0.010)	6.626(0.217)	5.109(0.158)	34.923(0.819)	22.544(0.544)
	15	0.749(0.030)	0.590(0.021)	6.526(0.154)	5.033(0.121)	35.033(1.170)	22.670(0.751)
	20	0.739(0.024)	0.581(0.016)	6.693(0.181)	5.234(0.153)	35.012(1.050)	22.719(0.937)
	25	0.753(0.033)	0.596(0.024)	6.860(0.233)	5.366(0.167)	36.007(1.264)	23.919(1.289)
	30	0.756(0.035)	0.598(0.022)	6.830(0.284)	5.335(0.232)	35.434(1.402)	23.530(0.888)
	50	0.804(0.083)	0.630(0.053)	7.403(0.830)	5.787(0.516)	36.914(3.603)	24.897(2.823)
sGPRN	5	0.719(0.011)	0.565(0.008)	6.553(0.157)	5.054(0.142)	33.475(0.456)	21.774(0.561)
	10	0.727(0.012)	0.567(0.010)	6.520(0.156)	5.062(0.179)	34.225(0.878)	22.114(0.697)
	15	0.725(0.022)	0.569(0.019)	6.479(0.184)	5.033(0.190)	34.308(1.030)	22.288(0.968)
	20	0.714(0.019)	0.563(0.014)	6.325(0.188)	4.885(0.172)	34.033(1.220)	21.962(0.851)
	25	0.722(0.027)	0.568(0.019)	6.311(0.159)	4.871(0.159)	34.440(1.399)	22.243(0.939)
	30	0.722(0.022)	0.566(0.015)	6.263(0.112)	4.834(0.103)	33.929(1.007)	21.945(0.611)
	50	0.704(0.032)	0.559(0.027)	6.307(0.256)	4.866(0.197)	33.665(0.795)	21.944(0.629)

SARCOS

Table 2: Normalized MSE results on the SARCOS test data for sparse GPRN and standard GPRN models. Numbers inside the bracket indicate one standard-error over 20 random splits.

Model		$m = 50$	$m = 100$	$m = 150$
GPRN	$Q = 2$	0.0167(0.0001)	0.0145(0.0011)	0.0127(0.0004)
	$Q = 3$	0.0146(0.0001)	0.0121(0.0001)	0.0108(0.0001)
sGPRN	$Q = 2$	**0.0159(0.0002)**	**0.0131(0.0001)**	**0.0125(0.0011)**
	$Q = 3$	**0.0140(0.0002)**	**0.0117(0.0002)**	**0.0096(0.0002)**

References

Hensman, J., Matthews, A., and Ghahramani, Z. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics*, pp. 351–360, 2015.