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

Abstract

Zero-inflated datasets, which have an excess of zero outputs, are commonly encountered in problems such as climate or rare event modelling. Conventional machine learning approaches tend to overestimate the non-zeros leading to poor performance. We propose a novel model family of *zero-inflated Gaussian processes* (ZiGP) for such zero-inflated datasets, produced by *sparse kernels* through learning a latent probit Gaussian process that can zero out kernel rows and columns whenever the signal is absent. The ZiGPs are particularly useful for making the powerful Gaussian process networks more interpretable. We introduce *sparse GP networks* where variable-order latent modelling is achieved through sparse mixing signals. We derive the non-trivial stochastic variational inference tractably for scalable learning of the sparse kernels in both models. The novel output-sparse approach improves both prediction of zero-inflated data and interpretability of latent mixing models.

1 INTRODUCTION

Zero-inflated quantitative datasets with overabundance of zero output observations are common in many domains, such as climate and earth sciences (Enke & Spekat, 1997; Wilby, 1998; Charles et al., 2004), ecology (del Saz-Salazar & Rausell-Köster, 2008; Ancelet et al., 2009), social sciences (Bohning et al., 1997), and in count processes (Barry & Welsh, 2002). Traditional regression modelling of such data tends to underestimate zeros and overestimate nonzeros (Andersen et al., 2014).

A conventional way of forming zero-inflated models is to estimate a mixture of a Bernoulli "on-off" process and a Poisson count distribution (Johnson & Kotz, 1969; Lambert, 1992). In hurdle models a binary "on-off" process determines whether a hurdle is crossed, and the positive responses are governed by a subsequent process (Cragg, 1971; Mullahy, 1986). The hurdle model is analogous to first performing classification and training a continuous predictor on the positive values only, while the zero-inflated model would regress with all observations. Both stages can be combined for simultaneous classification and regression Abraham & Tan (2010).

Gaussian process models have not been proposed for zero-inflated datasets since their posteriors are Gaussian, which are ill-fitted for zero predictions. A suite of Gaussian process models have been proposed for partially related problems, such as mixture models (Tresp, 2001; Rasmussen & Ghahramani, 2002; Lázaro-Gredilla et al., 2012) and change point detection (Herlands et al., 2016). Structured spike-and-slab models place smoothly sparse priors over the structured inputs (Andersen et al., 2014).

In contrast to other approaches, we propose a Bayesian model that learns the underlying latent prediction function, whose covariance is sparsified through another Gaussian process switching between the 'on' and 'off' states, resulting in an zero-inflated Gaussian process model. This approach introduces a tendency of predicting exact zeros to Gaussian processes, which is directly useful in datasets with excess zeros.

A Gaussian process network (GPRN) is a latent signal framework where multi-output data are explained through a set of latent signals and mixing weight Gaussian processes (Wilson et al., 2012). The standard GPRN tends to have dense mixing that combines all latent signals for all latent outputs. By

applying the zero-predicting Gaussian processes to latent mixture models, we introduce sparse GPRNs where latent signals are mixed with sparse instead of dense mixing weight functions. The sparse model induces variable-order mixtures of latent signals resulting in simpler and more interpretable models. We demonstrate both of these properties in our experiments with spatio-temporal and multi-output datasets.

Main contributions. Our contributions include[1]

1. A novel zero-inflated Gaussian process formalism consisting of a latent Gaussian process and a separate 'on-off' probit-linked Gaussian process that can zero out rows and columns of the model covariance. The novel sparse kernel adds to GPs the ability to predict zeros.
2. Novel stochastic variational inference (SVI) for such sparse probit covariances, which in general are intractable due to having to compute expectations of GP covariances with respect to probit-linked processes. We derive the SVI for learning both of the underlying processes.
3. A novel sparse GPRN with an on-off process in the mixing matrices leading to sparse and variable-order mixtures of latent signals.
4. A solution to the stochastic variational inference of sparse GPRN where the SVI is derived for the network of full probit-linked covariances.

2 GAUSSIAN PROCESSES

We begin by introducing the basics of conventional Gaussian processes. Gaussian processes (GP) are a family of non-parametric, non-linear Bayesian models (Rasmussen & Williams, 2006). Assume a dataset of n inputs $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $\mathbf{x}_i \in \mathbb{R}^D$ and noisy outputs $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. The observations $y = f(\mathbf{x}) + \varepsilon$ are assumed to have additive, zero mean noise $\varepsilon \sim \mathcal{N}(0, \sigma_y^2)$ with a zero-mean GP prior on the latent function $f(\mathbf{x})$,

$$f(\mathbf{x}) \sim \mathcal{GP}\left(0, K(\mathbf{x}, \mathbf{x}')\right), \tag{1}$$

which defines a distribution over functions $f(\mathbf{x})$ whose mean and covariance are

$$\mathbb{E}[f(\mathbf{x})] = 0 \tag{2}$$

$$\mathbf{cov}[f(\mathbf{x}), f(\mathbf{x}')] = K(\mathbf{x}, \mathbf{x}'). \tag{3}$$

[1]The `TensorFlow` compatible code will be made publicly available at `https://github.com/hegdepashupati/zero-inflated-gp`

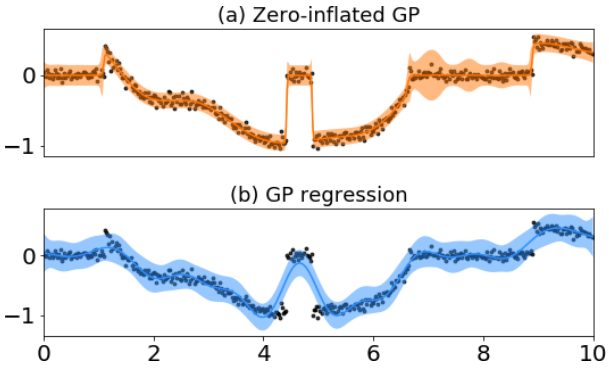

Figure 1: Illustration of a zero-inflated GP **(a)** and standard GP regression **(b)**. The standard approach is unable to model sudden loss of signal (at $4 \ldots 5$) and signal close to zero (at $0 \ldots 1$ and $7 \ldots 9$).

Then for any collection of inputs X, the function values follow a multivariate normal distribution $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K_{XX})$, where $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))^T \in \mathbb{R}^n$, and where $K_{XX} \in \mathbb{R}^{n \times n}$ with $[K_{XX}]_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. The key property of Gaussian processes is that they encode functions that predict similar output values $f(\mathbf{x}), f(\mathbf{x}')$ for similar inputs $\mathbf{x}, \mathbf{x}'$, with similarity determined by the kernel $K(\mathbf{x}, \mathbf{x}')$. In this paper we assume the Gaussian ARD kernel

$$K(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{j=1}^{D}\frac{(x_j - x_j')^2}{\ell_j^2}\right), \tag{4}$$

with a signal variance σ_f^2 and dimension-specific lengthscale $\ell_1, \ldots, \ell_D$ parameters.

The inference of the hyperparameters $\theta = (\sigma_y, \sigma_f, \ell_1, \ldots, \ell_D)$ is performed commonly by maximizing the marginal likelihood

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\theta) d\mathbf{f}, \tag{5}$$

which results in a convenient marginal likelihood called evidence, $p(\mathbf{y}|\theta) = N(\mathbf{y}|\mathbf{0}, K_{XX} + \sigma_y^2 I)$ for a Gaussian likelihood.

The Gaussian process defines a univariate normal predictive posterior distribution $f(\mathbf{x})|\mathbf{y}, X \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$ for an arbitrary input $\mathbf{x}$ with the prediction mean and variance[2]

$$\mu(\mathbf{x}) = K_{\mathbf{x}X}(K_{XX} + \sigma_y^2 I)^{-1}\mathbf{y}, \tag{6}$$

$$\sigma^2(\mathbf{x}) = K_{\mathbf{xx}} - K_{\mathbf{x}X}(K_{XX} + \sigma_y^2 I)^{-1}K_{X\mathbf{x}}, \tag{7}$$

[2]In the following we omit the implicit conditioning on data inputs X for clarity.

where $K_{X\mathbf{x}} = K_{\mathbf{x}X}^T \in \mathbb{R}^n$ is the kernel column vector over pairs $X \times \mathbf{x}$, and $K_{\mathbf{xx}} = K(\mathbf{x}, \mathbf{x}) \in \mathbb{R}$ is a scalar. The predictions $\mu(\mathbf{x}) \pm \sigma(\mathbf{x})$ come with uncertainty estimates in GP regression.

3 ZERO-INFLATED GAUSSIAN PROCESSES

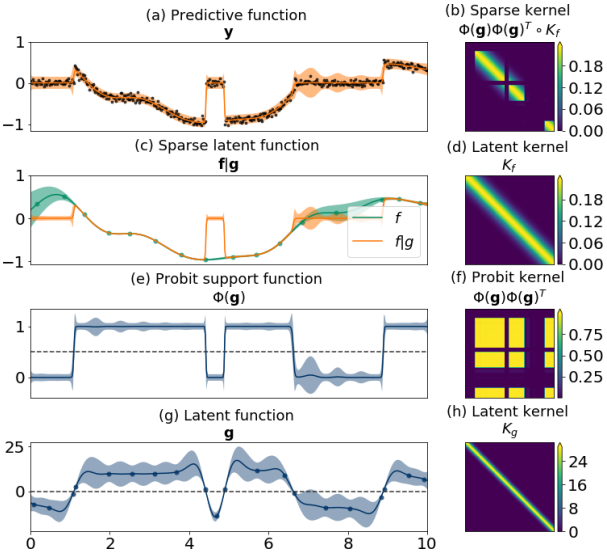

Figure 2: Illustration of the zero-inflated GP **(a)** and the sparse kernel **(b)** composed of a smooth latent function **(c,d)** filtered by a probit support function **(e,f)**, which is induced by the underlying latent sparsity **(g,h)**.

We introduce zero-inflated Gaussian processes that have – in contrast to standard GP's – a tendency to produce exactly zero predictions (See Figure 1). Let $g(\mathbf{x})$ denote the latent "on-off" state of a function $f(\mathbf{x})$. We assume GP priors for both functions with a joint model

$$p(\mathbf{y}, \mathbf{f}, \mathbf{g}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{g})p(\mathbf{g}), \tag{8}$$

where

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 I) \tag{9}$$
$$p(\mathbf{f}|\mathbf{g}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ K_f) \tag{10}$$
$$p(\mathbf{g}) = \mathcal{N}(\mathbf{g}|\beta\mathbf{1}, K_g) \tag{11}$$

The sparsity values $g(\mathbf{x})$ are squashed between 0 and 1 through a standard Normal cumulative distribution, or a probit link function, $\Phi : \mathbb{R} \to [0, 1]$

$$\Phi(g) = \int_{-\infty}^{g} \phi(\tau)d\tau = \frac{1}{2}\left(1 + \text{erf}\left(\frac{g}{\sqrt{2}}\right)\right), \tag{12}$$

where $\phi(\tau) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\tau^2}$ is the standard normal density function. The structured probit sparsity $\Phi(\mathbf{g})$ models the "on-off" smoothly due to the latent sparsity function $\mathbf{g}$ having a GP prior with prior mean β. The latent function $\mathbf{f}$ is modeled throughout but it is only visible during the "on" states. This masking effect has similarities to both zero-inflated and hurdle models. The underlying latent function $\mathbf{f}$ is learned from only non-zero data similarly to in hurdle models, but the function $\mathbf{f}$ is allowed to predict zeros similarly to zero-inflated models.

The key part of our model is the sparse *probit-sparsified covariance* $\Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ K$ where the "on-off" state $\Phi(\mathbf{g})$ has the ability to zero out rows and columns of the kernel matrix at the "off" states (See Figure 2f for the probit pattern $\Phi(\mathbf{g})\Phi(\mathbf{g})^T$ and Figure 2b for the resulting sparse kernel). Since the sparse kernel is represented as Hadamard product between a covariance kernel K and an outer product kernel $\Phi(\mathbf{g})\Phi(\mathbf{g})^T$, Schur product theorem implies that it is a valid kernel. As the sparsity $g(\mathbf{x})$ converges towards minus infinity, the probit link $\Phi(g(\mathbf{x}))$ approaches zero, which leads the function distribution approaching $\mathcal{N}(f_i|0, 0)$, or $f_i = 0$. Numerical problems are avoided since in practice $\Phi(g) > 0$, and due to the conditioning noise variance term $\sigma_y^2 > 0$.

The marginal likelihood of the zero-inflated Gaussian process is intractable due to the probit-sparsification of the kernel. We derive a stochastic variational Bayes approximation, which we show to be tractable due to the choice of using the probit link function.

3.1 STOCHASTIC VARIATIONAL INFERENCE

Inference for standard Gaussian process models is difficult to scale as complexity grows with $\mathcal{O}(n^3)$ as a function of the data size n. Titsias (2009) proposed a variational inference approach for GPs using $m < n$ inducing variables, with a reduced computational complexity of $\mathcal{O}(m^3)$ for m inducing points. The novelty of this approach lies in the idea that the locations and values of inducing points can be treated as variational parameters, and optimized. Hensman et al. (2013, 2015) introduced more efficient stochastic variational inference (SVI) with factorised likelihoods that has been demonstrated with up to billion data points (Salimbeni & Deisenroth, 2017). This approach cannot be directly applied to sparse kernels due to having to compute expectation of the probit product in the covariance. We derive

the SVI bound tractably for the zero-inflated model and its sparse kernel, which is necessary in order to apply the efficient parameter estimation techniques with automatic differentiation with frameworks such as TensorFlow (Abadi et al., 2016).

We begin by applying the inducing point augmentations $f(\mathbf{z}_f) = \mathbf{u}_f$ and $g(\mathbf{z}_g) = \mathbf{u}_g$ for both the latent function $f(\cdot)$ and the sparsity function $g(\cdot)$. We place m inducing points $\mathbf{u}_{f1}, \dots \mathbf{u}_{fm}$ and $\mathbf{u}_{g1}, \dots \mathbf{u}_{gm}$ for the two functions. The augmented joint distribution is $p(\mathbf{y}, \mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)p(\mathbf{u}_f)p(\mathbf{u}_g)$, where[3]

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

$$p(\mathbf{g}|\mathbf{u}_g) = \mathcal{N}(\mathbf{g}|Q_g\mathbf{u}_g, \widetilde{K}_g) \tag{14}$$

$$p(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f|\mathbf{0}, K_{fmm}) \tag{15}$$

$$p(\mathbf{u}_g) = \mathcal{N}(\mathbf{u}_g|\mathbf{0}, K_{gmm}) \tag{16}$$

and where

$$Q_f = K_{fnm}K_{fmm}^{-1} \tag{17}$$

$$Q_g = K_{gnm}K_{gmm}^{-1} \tag{18}$$

$$\widetilde{K}_f = K_{fnn} - K_{fnm}K_{fmm}^{-1}K_{fmn} \tag{19}$$

$$\widetilde{K}_g = K_{gnn} - K_{gnm}K_{gmm}^{-1}K_{gmn}. \tag{20}$$

We denote the kernels for functions f and g by the corresponding subscripts. The kernel K_{fnn} is between all n data points, the kernel K_{fnm} is between all n datapoints and m inducing points, and the kernel K_{fmm} is between all m inducing points (similarly for g as well).

The distributions $p(\mathbf{f}|\mathbf{u}_f)$ and $p(\mathbf{g}|\mathbf{u}_g)$ can be obtained by conditioning the joint GP prior between respective latent and inducing functions. Further, the conditional distribution $p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)$ can be obtained by the sparsity augmentation of latent conditional $\mathbf{f}|\mathbf{u}_f$ similar to equation (10) (See Supplements).

Next we use the standard variational approach by introducing approximative variational distributions for the inducing points,

$$q(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f|\mathbf{m}_f, \mathbf{S}_f) \tag{21}$$

$$q(\mathbf{u}_g) = \mathcal{N}(\mathbf{u}_g|\mathbf{m}_g, \mathbf{S}_g) \tag{22}$$

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

$$q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g) = p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_f)q(\mathbf{u}_g). \tag{23}$$

[3]We drop the implicit conditioning on $\mathbf{z}$'s for clarity.

We minimize the Kullback-Leibler divergence between the true augmented posterior $p(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g|\mathbf{y})$ and the variational distribution $q(\mathbf{f}, \mathbf{g}, \mathbf{u}_f, \mathbf{u}_g)$, which is equivalent to solving the following evidence lower bound (as shown by e.g. Hensman et al. (2015)):

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f})} \log p(\mathbf{y}|\mathbf{f}) - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_g)||p(\mathbf{u}_f, \mathbf{u}_g)], \tag{24}$$

where we define

$$\begin{aligned} q(\mathbf{f}) &= \iiint p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)q(\mathbf{u}_f)p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_g)d\mathbf{u}_f d\mathbf{u}_g d\mathbf{g} \\ &= \int q(\mathbf{f}|\mathbf{g})q(\mathbf{g})d\mathbf{g}, \end{aligned} \tag{25}$$

where the variational approximations are tractably

$$\begin{aligned} q(\mathbf{g}) &= \int p(\mathbf{g}|\mathbf{u}_g)q(\mathbf{u}_g)d\mathbf{u}_g \\ &= \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}_g, \Sigma_g) \end{aligned} \tag{26}$$

$$\begin{aligned} q(\mathbf{f}|\mathbf{g}) &= \int p(\mathbf{f}|\mathbf{g}, \mathbf{u}_f)q(\mathbf{u}_f)d\mathbf{u}_f \\ &= \mathcal{N}(\mathbf{f}|\operatorname{diag}(\Phi(\mathbf{g}))\boldsymbol{\mu}_f, \Phi(\mathbf{g})\Phi(\mathbf{g})^T \circ \Sigma_f) \end{aligned} \tag{27}$$

with

$$\boldsymbol{\mu}_f = Q_f\mathbf{m}_f \tag{28}$$

$$\boldsymbol{\mu}_g = Q_g\mathbf{m}_g \tag{29}$$

$$\Sigma_f = K_{fnn} + Q_f(\mathbf{S}_f - K_{fmm})Q_f^T \tag{30}$$

$$\Sigma_g = K_{gnn} + Q_g(\mathbf{S}_g - K_{gmm})Q_g^T. \tag{31}$$

We additionally assume the likelihood $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N p(y_i|f_i)$ factorises.

We solve the final ELBO of equations (24) and (25) as (See Supplements for detailed derivation)

$$\begin{aligned} \mathcal{L}_{\mathrm{ZI}} = \sum_{i=1}^N \Big\{ &\log \mathcal{N}(y_i|\langle\Phi(g_i)\rangle_{q(g_i)}\mu_{fi}, \sigma_y^2) \\ &- \frac{1}{2\sigma_y^2}\left(\mathrm{Var}[\Phi(g_i)]\mu_{fi}^2 + \langle\Phi(g_i)^2\rangle_{q(g_i)}\sigma_{fi}^2\right)\Big\} \\ &- \mathrm{KL}[q(\mathbf{u}_f)||p(\mathbf{u}_f)] - \mathrm{KL}[q(\mathbf{u}_g)||p(\mathbf{u}_g)], \end{aligned} \tag{32}$$

where μ_{fi} is the i'th element of $\boldsymbol{\mu}_f$ and σ_{fi}^2 is the i'th diagonal element of Σ_f (similarly with g).

The expectations are tractable,

$$\langle \Phi(g_i) \rangle_{q(g_i)} = \Phi(\lambda_{gi}), \qquad \lambda_{gi} = \frac{\mu_{gi}}{\sqrt{1+\sigma_{gi}^2}} \tag{33}$$

$$\langle \Phi(g_i)^2 \rangle_{q(g_i)} = \Phi(\lambda_{gi}) - 2T\left(\lambda_{gi}, \frac{\lambda_{gi}}{\mu_{gi}}\right) \tag{34}$$

$$\mathrm{Var}[\Phi(g_i)] = \Phi(\lambda_{gi}) - 2T\left(\lambda_{gi}, \frac{\lambda_{gi}}{\mu_{gi}}\right) - \Phi(\lambda_{gi})^2. \tag{35}$$

The Owen's T function $T(a,b) = \phi(a)\int_0^b \frac{\phi(a\tau)}{1+\tau^2} d\tau$ (Owen, 1956) has efficient numerical solutions in practise (Patefield & Tandy, 2000).

The ELBO is considerably more complex than the standard stochastic variational bound of a Gaussian process (Hensman et al., 2013), due to the probit-sparsified covariance.

The bound is likely only tractable for the choice of probit link function $\Phi(\mathbf{g})$, while other link functions such as the logit would lead to intractable bounds necessitating slower numerical integration (Hensman et al., 2015).

We optimize the $\mathcal{L}_{\text{ZI}}$ with stochastic gradient ascent techniques with respect to the inducing locations $\mathbf{z}_g, \mathbf{z}_f$, inducing value means $\mathbf{m}_f, \mathbf{m}_g$ and covariances $\mathbf{S}_f, \mathbf{S}_g$, the sparsity prior mean β, the noise variance σ_y^2, the signal variances σ_f, σ_g, and finally the dimensions-specific lengthscales $\ell_{f1}, \ldots, \ell_{fD}; \ell_{g1}, \ldots, \ell_{gD}$ of the Gaussian ARD kernel.

4 GAUSSIAN PROCESS NETWORK

The Gaussian Process Regression Networks (GPRN) framework by Wilson et al. (2012) is an efficient model for multi-target regression problems, where each individual output is a linear but non-stationary combination of shared latent functions. Formally, 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

$$\mathbf{y}(x) = W(x)[\mathbf{f}(x) + \boldsymbol{\epsilon}] + \boldsymbol{\varepsilon}, \tag{36}$$

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

$$f_q(\mathbf{x}) \sim \mathcal{GP}(0, K_f(\mathbf{x}, \mathbf{x}')) \tag{37}$$

$$W_{qp}(\mathbf{x}) \sim \mathcal{GP}(0, K_w(\mathbf{x}, \mathbf{x}')) \tag{38}$$

$$\epsilon_q \sim \mathcal{N}(0, \sigma_f^2) \tag{39}$$

$$\varepsilon_p \sim \mathcal{N}(0, \sigma_y^2). \tag{40}$$

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 assume shared latent and output noise variances σ_f^2, σ_y^2 without loss of generality. The distributions of both functions $\mathbf{f}$ and W have been inferred either with variational EM (Wilson et al., 2012) or by variational mean-field approximation with diagonalized latent and mixing functions (Nguyen & Bonilla, 2013).

4.1 STOCHASTIC VARIATIONAL INFERENCE

Variational inference for GPRN has been proposed earlier with diagonalized mean-field approximation by (Nguyen & Bonilla, 2013). Further, stochastic variational inference by introducing inducing variables has been proposed for GPRN (Nguyen et al., 2014). In this section we rederive the SVI bound for standard GPRN for completeness and then propose the novel sparse GPRN model, and solve its SVI bounds as well, in the following section.

We begin by introducing the inducing variable augmentation technique 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 $\mathbf{u}_w, \mathbf{z}_w = \{\mathbf{u}_{w_{qp}}, \mathbf{z}_{w_{qp}}\}_{q,p=1}^{Q,P}$:

$$\begin{aligned} p(\mathbf{y}, \mathbf{f}, W, \mathbf{u}_f, \mathbf{u}_w) & \\ = p(\mathbf{y}|\mathbf{f}, W) p(\mathbf{f}|\mathbf{u}_f) p(W|\mathbf{u}_w) p(\mathbf{u}_f) p(\mathbf{u}_w) & \end{aligned} \tag{41}$$

$$p(\mathbf{f}|\mathbf{u}_f) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{f}_q | Q_{f_q} \mathbf{u}_{f_q}, \widetilde{K}_{f_q}) \tag{42}$$

$$p(W|\mathbf{u}_w) = \prod_{q,p=1}^{Q,P} \mathcal{N}(\mathbf{w}_{qp} | Q_{w_{qp}} \mathbf{u}_{w_{qp}}, \widetilde{K}_{w_{qp}}) \tag{43}$$

$$p(\mathbf{u}_f) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{u}_{f_q} | \mathbf{0}, K_{f_q,mm}) \tag{44}$$

$$p(\mathbf{u}_w) = \prod_{q,p=1}^{Q,P} \mathcal{N}(\mathbf{u}_{w_{qp}} | \mathbf{0}, K_{w_{qp},mm}), \tag{45}$$

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 same form as in equations (17–20). The $\mathbf{w}$ is a vectorised form of W. The variational approximation is then

$$q(\mathbf{f}, W, \mathbf{u}_f, \mathbf{u}_w) = p(\mathbf{f}|\mathbf{u}_f)p(W|\mathbf{u}_w)q(\mathbf{u}_f)q(\mathbf{u}_w) \tag{46}$$

$$q(\mathbf{u}_{f_q}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{u}_{f_q}|\mathbf{m}_{f_q}, \mathbf{S}_{f_q}) \tag{47}$$

$$q(\mathbf{u}_{w_{qp}}) = \prod_{q,p=1}^{Q,P} \mathcal{N}(\mathbf{u}_{w_{qp}}|\mathbf{m}_{w_{qp}}, \mathbf{S}_{w_{qp}}), \tag{48}$$

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

$$\begin{aligned}\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f},W)} \log p(\mathbf{y}|\mathbf{f}, W) \\ - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_w)||p(\mathbf{u}_f, \mathbf{u}_w)],\end{aligned} \tag{49}$$

where the variational distributions decompose as $q(\mathbf{f}, W) = q(\mathbf{f})q(W)$ with marginals of the same form as in equations (28–31),

$$q(\mathbf{f}) = \int q(\mathbf{f}|\mathbf{u}_f)q(\mathbf{u}_f)d\mathbf{u}_f = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_f, \Sigma_f) \tag{50}$$

$$q(W) = \int q(W|\mathbf{u}_w)q(\mathbf{u}_w)d\mathbf{u}_w = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_w, \Sigma_w). \tag{51}$$

Since the noise term $\boldsymbol{\varepsilon}$ is assumed to be isotropic Gaussian, the density $p(\mathbf{y}|W, \mathbf{f})$ factorises across all target observations and dimensions. The expectation term in equation (49) then reduces to solving the following integral for the i^{th} observation and p^{th} target dimension,

$$\sum_{i,p=1}^{N,P} \iint \log \mathcal{N}(y_{p,i}|\mathbf{w}_{p,i}^T\mathbf{f}_i, \sigma_y^2)q(\mathbf{f}_i, \mathbf{w}_{p,i})d\mathbf{w}_{p,i}d\mathbf{f}_i. \tag{52}$$

The above integral has a closed form solution resulting in the final ELBO as (See Supplements)

$$\begin{aligned}\mathcal{L}_{\text{GPRN}} = \sum_{i=1}^{N}\Bigg\{\sum_{p=1}^{P}\log\mathcal{N}\Big(y_{p,i}|\sum_{q=1}^{Q}\mu_{w_{qp},i}\mu_{f_q,i}, \sigma_y^2\Big) \\ - \frac{1}{2\sigma_y^2}\sum_{q,p=1}^{Q,P}\Big(\mu_{w_{qp},i}^2\sigma_{f_q,i}^2 + \mu_{f_q,i}^2\sigma_{w_{qp},i}^2 + \sigma_{w_{qp},i}^2\sigma_{f_q,i}^2\Big)\Bigg\} \\ - \sum_{q,p}^{Q,P}\mathrm{KL}[q(\mathbf{u}_{w_{qp}}, \mathbf{u}_{f_q})||p(\mathbf{u}_{w_{qp}}, \mathbf{u}_{f_q})],\end{aligned} \tag{53}$$

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 the W_{qp}'s).

5 SPARSE GAUSSIAN PROCESS NETWORK

In this section we demonstrate how zero-inflated GPs can be used as plug-in components in other standard models. In particular, we propose a significant modification to GPRN by adding sparsity to the mixing matrix components. This corresponds to each of the p outputs being a sparse mixture of the latent Q functions, i.e. they can effectively use any subset of the Q latent dimensions by having zeros for the rest in the mixing functions. This makes the mixture more easily interpretable, and induces a variable number of latent functions to explain the output of each input $\mathbf{x}$. The latent function $\mathbf{f}$ can also be sparsified, with a derivation analogous to the derivation below.

We extend the GPRN with probit sparsity for the mixing matrix W, resulting in a joint model

$$p(\mathbf{y}, \mathbf{f}, W, \mathbf{g}) = p(\mathbf{y}|\mathbf{f}, W)p(\mathbf{f})p(W|\mathbf{g})p(\mathbf{g}), \tag{54}$$

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

$$p(W_{qp}|\mathbf{g}_{qp}) = \mathcal{N}(\mathbf{w}_{qp}|\mathbf{0}, \Phi(\mathbf{g}_{qp})\Phi(\mathbf{g}_{qp})^T \circ K_w). \tag{55}$$

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 as in equations (25–27), the marginal likelihood can be written as

$$\begin{aligned}\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f},W,\mathbf{g})} \log p(\mathbf{y}|\mathbf{f}, W) \\ - \mathrm{KL}[q(\mathbf{u}_f, \mathbf{u}_w, \mathbf{u}_g)||p(\mathbf{u}_f, \mathbf{u}_w, \mathbf{u}_g)].\end{aligned} \tag{56}$$

The joint distribution in the variational expectation factorizes as $q(\mathbf{f}, W, \mathbf{g}) = q(\mathbf{f})q(W|\mathbf{g})q(\mathbf{g})$. Also, with a Gaussian noise assumption, the expectation term factories across all the observations and target dimensions. The key step reduces to solving the following integrals:

$$\begin{aligned}\sum_{i,p=1}^{N,P}\iiint \log\mathcal{N}(y_{p,i}|(\mathbf{w}_{p,i} \circ \mathbf{g}_{p,i})^T\mathbf{f}_i, \sigma_y^2) \\ \cdot q(\mathbf{f}_i, \mathbf{w}_{p,i}, \mathbf{g}_{p,i})d\mathbf{w}_{p,i}d\mathbf{f}_i d\mathbf{g}_{p,i}.\end{aligned} \tag{57}$$

The above integral has a tractable solution leading to the final sparse GPRN evidence lower bound (See Supplements)

$$\mathcal{L}_{\text{sGPRN}} = \sum_{i=1}^{N} \Bigg\{ \sum_{p=1}^{P} \log \mathcal{N}\Big(y_{p,i} | \sum_{q=1}^{Q} \mu_{w_{qp},i} \mu_{g_{qp},i} \mu_{f_q,i}, \sigma_y^2\Big) - \frac{1}{2\sigma_y^2} \sum_{q,p=1}^{Q,P} \Big((\mu_{g_{qp},i}^2 + \sigma_{g_{qp},i}^2) \cdot (\mu_{w_{qp},i}^2 \sigma_{f_q,i}^2 + \mu_{f_q,i}^2 \sigma_{w_{qp},i}^2 + \sigma_{w_{qp},i}^2 \sigma_{f_q,i}^2) \Big) - \frac{1}{2\sigma_y^2} \sum_{q,p=1}^{Q,P} \Big(\sigma_{g_{qp},i}^2 \mu_{f_q,i}^2 \mu_{w_{qp},i}^2 \Big) \Bigg\} - \sum_{q,p}^{Q,P} \text{KL}[q(\mathbf{u}_{f_q}, \mathbf{u}_{w_{qp}}, \mathbf{u}_{g_{qp}}) || p(\mathbf{u}_{f_q}, \mathbf{u}_{w_{qp}}, \mathbf{u}_{g_{qp}})], \tag{58}$$

where $\mu_{f_q,i}, \mu_{w_{qp},i}$ are the variational expectation means for $\mathbf{f}(\cdot), W(\cdot)$ as in equations (28, 29), $\mu_{g_{qp},i}$ is the variational expectation mean of $g(\cdot)$ as in equation (33), and analogously for the variances.

6 EXPERIMENTS

First we demonstrate how the proposed method can be used for regression problems with zero-inflated targets. We do that both on a simulated dataset and for real-world climate modeling scenarios on a Finnish rain precipitation dataset with approximately 90% zeros. Finally, we demonstrate the GPRN model and how it improves both the interpretability and predictive performance in the JURA geological dataset.

We use the squared exponential kernel with ARD in all experiments. All the parameters including inducing locations, values and variances and kernel parameters were learned through stochastic Adam optimization (Kingma & Ba, 2014) on the TensorFlow (Abadi et al., 2016) platform.

We compare our approach ZIGP to baseline ZERO voting, to conventional Gaussian process regression (GPR) and classification (GPC) with SVI approximations from the GPflow package (Matthews et al., 2017). Finally, we also compare to first classifying the non-zeros, and successively applying regression either to all data points (GPCR), or to only predicted non-zeros ($\text{GPCR}_{\neq 0}$, hurdle model).

We record the predictive performance by considering mean squared error and mean absolute error. We also compare the models' ability to predict true zeros with F1, accuracy, precision, and recall of the optimal models.

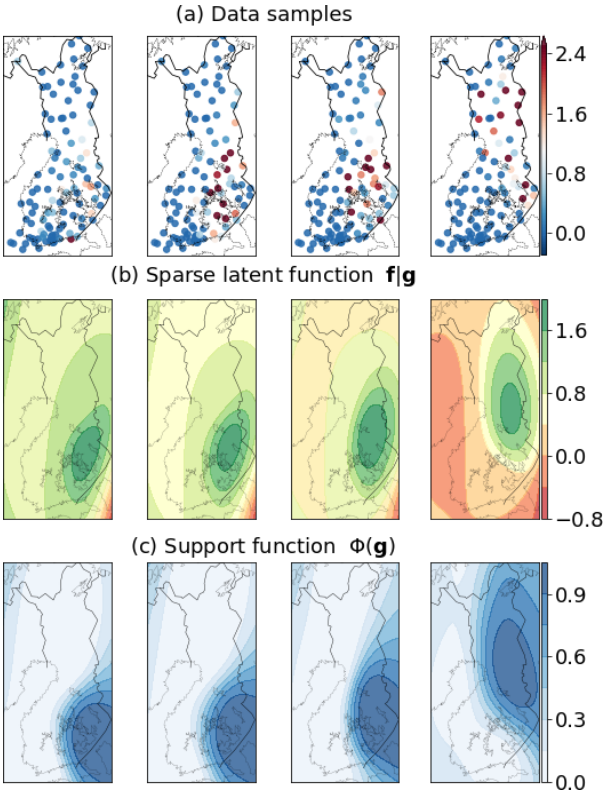

Figure 3: ZIGP model fit on the precipitation dataset. Sample of the actual data **(a)** against the sparse rain function estimate **(b)**, with the probit support function **(c)** showing the rain progress.

6.1 SPATIO-TEMPORAL DATASET

Zero-inflated cases are commonly found in climatology and ecology domains. In this experiment we demonstrate the proposed method by modeling precipitation in Finland[4]. The dataset consists of hourly quantitative non-negative observations of precipitation amount across 105 observatory locations in Finland for the month of July 2018. The dataset contains 113015 datapoints with approximately 90% zero precipitation observations. The data inputs are three-dimensional: latitude, longitude and time. Due to the size of the data, this experiment illustrates the scalability of the variational inference.

We randomly split 80% of the data for training and the rest 20% for testing purposes. We split across time only, such that at a single measurement time, all locations are simultaneously either in the training set, or in the test set.

[4]Data can be found at `http://en.ilmatieteenlaitos.fi/`

Table 1: Results for the precipitation dataset over baseline (Zero; majority voting), four competing methods and the proposed method ZIGP on test data. The columns list both quantitative and qualitative performance criteria, best performance is boldfaced.

MODEL	RMSE	MAE	F1	ACC.	PREC.	RECALL
ZERO	0.615	0.104	0.000	0.898	0.000	0.000
GPC	-	-	0.367	**0.911**	0.675	0.252
GPR	0.569	0.159	0.401	0.750	0.266	**0.817**
GPCR	0.589	**0.102**	0.366	**0.911**	0.679	0.251
$\text{GPCR}_{\neq 0}$	0.575	**0.101**	0.358	**0.912**	**0.712**	0.240
ZIGP	**0.561**	0.121	**0.448**	0.861	0.381	0.558

We further utilize the underlying spatio-temporal grid structure of the data to perform inference in an efficient manner by Kronecker techniques (Saatchi, 2011). All the kernels for latent processes are assumed to factorise as $\mathbf{K} = \mathbf{K}_{space} \otimes \mathbf{K}_{time}$ which allows placing inducing points independently on spatial and temporal grids.

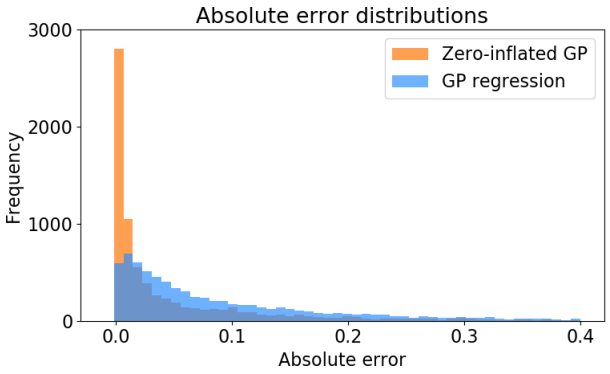

Figure 4: The distribution of errors with the rain dataset with the ZIGP and the GPR. The zero-inflated GP achieves much higher number of perfect (zero) predictions.

Figure 3 depicts the components of the zero-inflated GP model on the precipitation dataset. As shown in panel (c), the latent support function models the presence or absence of rainfall. It smoothly follows the change in rain patterns across hourly observations. The amount of precipitation is modeled by the other latent process and the combination of these two results in sparse predictions. Figure 4 shows that the absolute error distribution is remarkably better with the ZIGP model due to it identifying the absence of rain exactly. While both models fit the high rainfall regions well, for zero and near-zero regions GPR does not refine its small errors. Table 1 indicates that the ZIGP model achieves the lowest mean square error, while also achieving the highest F1 score that takes into account the class imbalance,

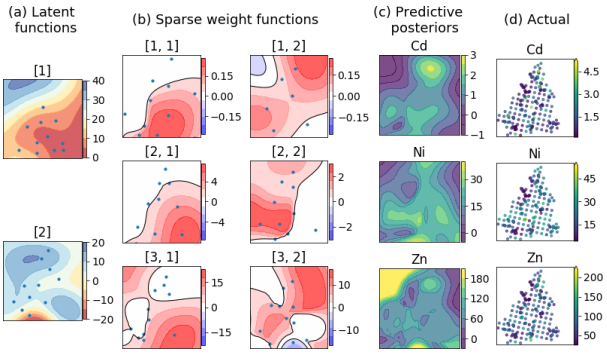

Figure 5: The sparse GPRN model fit on the Jura dataset with 11 inducing points. The $Q = 2$ (dense) latent functions **(a)** are combined with the 3×2 sparse mixing functions **(b)** into the $P = 3$ output predictions **(c)**. The real data are shown in **(d)**. The white mixing regions are estimated 'off'.

which biases the elementary accuracy, precision and recall quantities towards the majority class.

6.2 MULTI-OUTPUT PREDICTION - JURA

In this experiment we model the multi-response Jura dataset with the sparse Gaussian process regression network SGPRN model and compare it with standard GPRN as baseline. Jura contains concentration measurements of cadmium, nickel and zinc metals in the region of Swiss Jura. We follow the experimental procedure of Wilson et al. (2012) and Nguyen & Bonilla (2013). The training set consists of $n = 259$ observations across $D = 2$ dimensional geo-spatial locations, and the test set consists of 100 separate locations. For both models we use $Q = 2$ latent functions with the stochastic variational inference techniques proposed in this paper. Sparse GPRN uses a sparsity inducing kernel in the mixing weights. The locations of inducing points for the weights $W(\mathbf{x})$ and the support $g(\mathbf{x})$ are shared. The kernel length-scales are given a gamma prior with the shape parameter $\alpha = 0.3$ and rate parameter $\beta = 1.0$ to induce smoothness. We train both the models 30 times with random initialization.

Table 2 shows that our model performs better than the state-of-the-art SVI-GPRN, both with $m = 5$ and $m = 10$ inducing points. Figure 5 visualises the optimized sparse GPRN model, while Figure 6 indicates the sparsity pattern in the mixing weights. The weights have considerable smooth 'on' regions (black) and 'off' regions (white). The 'off' regions indicate that for certain locations, only one of the two latent functions is adaptively utilised.

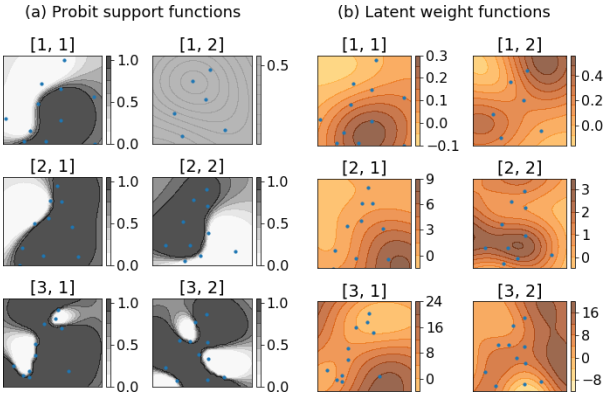

Figure 6: The sparse probit support **(a)** and latent functions **(b)** of the weight function $W(\mathbf{x})$ of the optimized sparse GPRN model. The black regions of **(a)** show regional activations, while the white regions show where the latent functions are 'off'. The elementwise product of the support and weight functions is indicated in the Figure 5**b)**.

Table 2: Results for the Jura dataset for sparse GPRN and vanilla GPRN models with test data. Best performance is with boldface. We do not report RMSE and MAE values GPC, since its a classification method.

		Cadmium		Nickel		Zinc	
Model	m	RMSE	MAE	RMSE	MAE	RMSE	MAE
GPRN	5	0.724	0.566	**6.469**	**4.958**	33.729	21.959
	10	0.736	0.574	6.626	5.109	34.923	22.544
	15	0.749	0.590	6.526	5.033	35.033	22.670
sGPRN	5	**0.719**	**0.565**	6.553	5.054	**33.475**	**21.774**
	10	**0.727**	**0.567**	**6.520**	**5.062**	**34.225**	**22.114**
	15	**0.725**	**0.569**	**6.479**	**5.033**	**34.308**	**22.288**

6.3 MULTI-OUTPUT PREDICTION - SARCOS

In this experiment we tackle the problem of learning inverse dynamics for seven degrees of freedom of SARCOS anthropomorphic robot arm (Vijayakumar et al., 2005). The dataset consists of 48,933 observations with an input space of 21 dimensions (7 joints positions, 7 joint velocities, 7 joint accelerations). The multi-output prediction task is to learn a mapping from these input variables to the corresponding 7 joint torques of the robot. Multi-output GP has been previously used for inverse dynamics modeling (Williams et al., 2009), but in a different model setting and on a smaller dataset. GPRN with stochastic inference framework has also been explored to model SARCOS dataset (Nguyen et al., 2014), however, they use a different experimental setting and consider 2 of the 7 joint torques as multi-outptut targets.

Table 3: Normalized MSE results on the SARCOS test data for sparse GPRN and standard GPRN models. Best performance is mentioned with boldface.

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

We consider 80%+20% random split of the full dataset for training and testing respectively. Both GPRN and SGPRN model are trained with $m = 50, 100$ and 150 inducing points and $Q = 2$ and 3 latent functions. We repeat the experiment 20 times and report normalized-MSE in Table 3. Sparse GPRN gives better results than standard GPRN in all our experimental settings. Moreover, sparse model (nMSE= 0.0096) gives gives 12% improvement over the standard model (nMSE= 0.0108) for the best test performance with $Q = 3$ latent functions and $m = 150$.

7 DISCUSSION

We proposed a novel paradigm of zero-inflated Gaussian processes with a novel sparse kernel. The sparsity in the kernel is modeled with smooth probit filtering of the covariance rows and columns. This model induces zeros in the prediction function outputs, which is highly useful for zero-inflated datasets with excess of zero observations. Furthermore, we showed how the zero-inflated GP can be used to model sparse mixtures of latent signals with the proposed sparse Gaussian process network. The latent mixture model with sparse mixing coefficients leads to locally using only a subset of the latent functions, which improves interpretability and reduces model complexity. We demonstrated tractable solutions to stochastic variational inference of the sparse probit kernel for the zero-inflated GP, conventional GPRN, and sparse GPRN models, which lends to efficient exploration of the parameter space of the model.

Acknowledgements

We would like to thank anonymous reviewers for their helpful suggestions and comments. This work has been supported by the Academy of Finland grant no. 294238 and 292334.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pp. 265–283, 2016.

Abraham, Z. and Tan, P-N. An integrated framework for simultaneous classification and regression of time-series data. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 653–664. SIAM, 2010.

Ancelet, S., Etienne, M-P., Benot, H., and Parent, E. Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. *Environmental and Ecological Statistics*, 2009.

Andersen, M., Winther, O., and Hansen, L. Bayesian inference for structured spike and slab priors. In *NIPS*, pp. 1745–1753, 2014.

Barry, S. and Welsh, A. H. Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157:179–188, 2002.

Bohning, D., Dierz, E., and Schlattmann, P. Zero-inflated count models and their applications in public health and social science. In Rost, J. and Langeheine, R. (eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences*. Waxman Publishing Co, 1997.

Charles, S., Bates, B., Smith, I., and Hughes, J. Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes*, pp. 1373–1394, 2004.

Cragg, J.G. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39:829–844, 1971.

del Saz-Salazar, S. and Rausell-Köster, P. A double-hurdle model of urban green areas valuation: dealing with zero responses. *Landscape and urban planning*, 84(3-4):241–251, 2008.

Enke, W. and Spekat, A. Downscaling climate model outputs into local and regional weather elements by classification and regression. *Climate Research*, 8:195–207, 1997.

Hensman, J., Fusi, N., and Lawrence, N. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 282–290. AUAI Press, 2013.

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

Herlands, W., Wilson, A., Nickisch, H., Flaxman, S., Neill, D., van Panhuis, W., and Xing, E. Scalable Gaussian processes for characterizing multidimensional change surfaces. In *AISTATS*, volume 51 of *PMLR*, pp. 1013–1021, 2016.

Johnson, N. and Kotz, S. *Distributions in Statistics: Discrete Distributions*. Houghton MiZin, Boston, 1969.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Lambert, D. Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, 34:1–14, 1992.

Lázaro-Gredilla, M., Van Vaerenbergh, S., and Lawrence, N. Overlapping mixtures of Gaussian processes for the data association problem. *Pattern Recognition*, 45(4):1386–1395, 2012.

Matthews, A., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., Leon-Villagra, P., Ghahramani, Z., and Hensman, J. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

Mullahy, J. Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365, 1986.

Nguyen, T. and Bonilla, E. Efficient variational inference for Gaussian process regression networks. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pp. 472–480. PMLR, 2013.

Nguyen, Trung V, Bonilla, Edwin V, et al. Collaborative multi-output gaussian processes. In *UAI*, pp. 643–652, 2014.

Owen, D.B. Tables for computing bivariate normal probabilities. *Annals of Mathematical Statistics*, 27:1075–1090, 1956.

Patefield, M. and Tandy, D. Fast and accurate calculation of owens t-function. *Journal of Statistical Software*, 5:1–25, 2000.

Rasmussen, C. and Ghahramani, Z. Infinite mixtures of Gaussian process experts. In *NIPS*, pp. 881–888, 2002.

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

Saatchi, Y. *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge, 2011.

Salimbeni, H. and Deisenroth, M. Doubly stochastic variational inference for deep Gaussian processes. In *NIPS*, volume 30, 2017.

Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Tresp, V. Mixtures of Gaussian processes. In *NIPS*, pp. 654–660, 2001.

Vijayakumar, Sethu, D'Souza, Aaron, and Schaal, Stefan. Lwpr: A scalable method for incremental online learning in high dimensions. 2005.

Wilby, R.L. Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection. *Climate Research*, 10:163–178, 1998.

Williams, Christopher, Klanke, Stefan, Vijayakumar, Sethu, and Chai, Kian M. Multi-task gaussian process learning of robot inverse dynamics. In *Advances in Neural Information Processing Systems*, pp. 265–272, 2009.

Wilson, A. G., Knowles, D., and Ghahramani, Z. Gaussian process regression networks. In *ICML*, 2012.