Variational Elliptical Processes

Anonymous authors
Paper under double-blind review

Abstract

We present elliptical processes—a family of non-parametric probabilistic models that subsumes the Gaussian processes and the Student's t processes. This generalization includes a range of new heavy-tailed behaviors while retaining computational tractability. The elliptical processes are based on a representation of elliptical distributions as a continuous mixture of Gaussian distributions. We parameterize this mixture distribution as a spline normalizing flow, which we train using variational inference. The proposed form of the variational posterior enables a sparse variational elliptical process applicable to large-scale problems. We highlight some advantages compared to a Gaussian process through regression and classification experiments. Elliptical processes can replace Gaussian processes in several settings, including cases where the likelihood is non-Gaussian or when accurate tail modeling is essential.

1 Introduction

Non-Gaussian data arise in many real-world settings, for instance in finance (Mandelbrot, 1963), signal processing (Zoubir et al., 2012) and geostatistics (Diggle et al., 1998). While Gaussian processes ($\mathcal{GP}$s) offer a powerful and widely used modeling framework, it can be seriously misleading in such situations. We use a combination of normalizing flows and modern variational inference techniques to extend the modeling capabilities of $\mathcal{GP}$s to the more general class of elliptical processes ($\mathcal{EP}$s).

Elliptical processes. The elliptical processes subsume the Gaussian process and the Student's t process (Shah et al., 2014). It is based on the elliptical distribution—a scale-mixture of Gaussian distributions attractive mainly because it can describe heavy-tailed distributions while retaining most of the Gaussian distribution's computational tractability (Fang et al., 1990). We use a normalizing flow (Papamakarios et al., 2021a) to model the continuous scale-mixture, which provides an added flexibility that can benefit a range of applications. We explore the use of elliptical processes as both a prior (over functions) and a likelihood, as well as the combination thereof. We also explore the use of $\mathcal{EP}$s as a variational posterior that can adapt its shape to match complex posterior distributions (Tran et al., 2016).

Variational inference. Variational inference is a powerful tool for approximate inference that uses optimization to find a member of a predefined family of distributions that is close to the target distribution (Wainwright et al., 2008; Blei et al.,

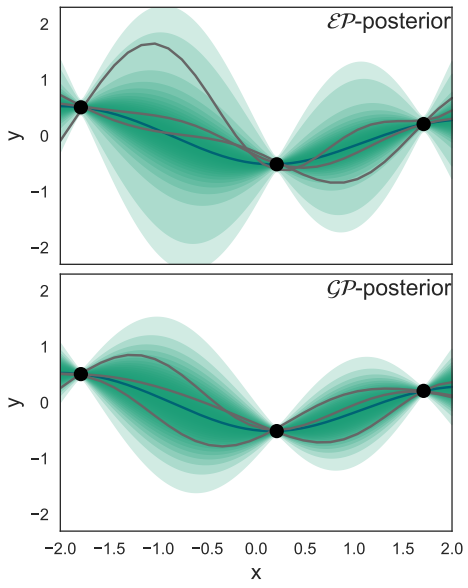

Figure 1: Posterior distributions of an elliptical process and a Gaussian process with equal kernel hyperparameters and covariance. The shaded area are confidence intervals of the posterior processes. The elliptical confidence regions are wider due to the process's heavier tail, which makes the confidence region similar to the Gaussian's close to the mean, but also allows samples further out at the tail.

2017). Significant advances made in the last decade have made variational inference the method of choice for scalable approximate inference in complex parametric models (Ranganath et al., 2014; Hoffman et al., 2013; Kingma & Welling, 2013; Rezende et al., 2014).

It is thus not surprising that the quest for more expressive and scalable variations of Gaussian processes has gone hand-in-hand with the developments in variational inference. For instance, sparse $\mathcal{GP}$s use variational inference to select inducing points to approximate the prior (Titsias, 2009). Inducing points is a common building block in deep probabilistic models such as deep Gaussian processes (Damianou & Lawrence, 2013; Salimbeni et al., 2019) and can also be applied in Bayesian neural networks Maroñas et al. (2021); Ober & Aitchison (2021). Similarly, the combination of inducing points and variational inference enables scalable approximate inference in models with non-Gaussian likelihoods, such as when performing $\mathcal{GP}$ classification (Hensman et al., 2015; Wilson et al., 2016).

However, the closeness of the variational distribution to the target distribution is bounded by the flexibility of the variational distribution. Consequently, the success of deep (neural network) models have inspired various suggestions on flexible yet tractable variational distributions, often based on parameterized transformations of a simple base distribution (Tran et al., 2016). In particular, models using a composition of invertible transformations, known as normalizing flows, have been especially popular (Rezende & Mohamed, 2015; Papamakarios et al., 2021a).

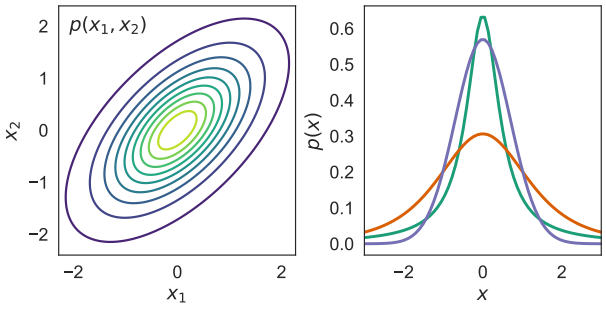

Figure 2: **Left:** A contour plot of an elliptical two-dimensional, correlated distribution with zero means. The name derives from its elliptical level sets. **Right:** Three examples of one-dimensional elliptical distributions with zero means and varying tail-heaviness. Elliptical distributions are symmetric around the mean $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$.

Our contributions. We propose an adaptation of elliptical distributions and processes in the same spirit as modern Gaussian processes. Constructing elliptical distributions based on a normalizing flow provides a high degree of flexibility without sacrificing computational tractability. This makes it possible to sidestep the "curse of Gaussianity", and adapt to heavy-tailed behavior when called for. We thus foresee many synergies between $\mathcal{EP}$s and recently developed $\mathcal{GP}$ methods. We make a first exploration of these, and simultaneously demonstrate the versatility of the elliptical process as a model for the prior and/or the likelihood, or as the variational posterior. In more detail, our contributions are:

- a construction of the elliptical process and the elliptical likelihood as a continuous scale-mixture of Gaussian processes parameterized by a normalizing flow, which offers a natural generalization of the Gaussian and Student's t processes;
- a variational factorization that captures heavy-tailed posteriors, allowing us to create a sparse variational elliptical process applicable to large-scale problems;
- an elliptical process and an elliptical likelihood conditioned on the data, constructed using amortized variational inference. We exemplify how to use this construction when modelling heteroscedastic noise.

2 Background

In this section, we present the necessary background on elliptical distributions, elliptical processes and normalizing flow models. Throughout, we consider the regression problem, where we are given a set of N scalar observations, $\boldsymbol{y} = [y_1, \cdots, y_N]^\top$, at the locations $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^\top$, where $\boldsymbol{x}_n$ is D dimensional. The

measurements y_n are assumed to be noisy measurements, such that,

$$y_n = f(\boldsymbol{x}_n) + \epsilon_n, \tag{1}$$

where ϵ_n is zero mean, i.i.d., noise. The task is to infer the underlying function, $f : \mathbb{R}^D \to \mathbb{R}$.

2.1 Elliptical distributions

The elliptical process is based on elliptical distributions (Figure 2), which include Gaussian distributions as well as more heavy-tailed distributions, such as the Student's t distribution and the Cauchy distribution.

The probability density of a random variable $Y \in \mathbb{R}^N$ that follows the elliptical distribution can be expressed as,

$$p(u;\, \boldsymbol{\eta}) = c_{n,\boldsymbol{\eta}} |\boldsymbol{\Sigma}|^{-1/2} g_N(u;\, \boldsymbol{\eta}), \tag{2}$$

where $u = (\mathbf{y} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ is the squared Mahalanobis distance, $\boldsymbol{\mu}$ is the location vector, $\boldsymbol{\Sigma}$ is the non-negative definite scale matrix, and $c_{N,\boldsymbol{\eta}}$ is a normalization constant. The density generator $g_{N,\boldsymbol{\eta}}(u)$ is a non-negative function with finite integral parameterized by $\boldsymbol{\eta}$ which determines the shape of the distribution.

Elliptical distributions are consistent, i.e., closed under marginalization, if and only if $p(u;\, \boldsymbol{\eta})$ is a scale-mixture of Gaussian distributions (Kano, 1994). The density can be expressed as

$$p(u;\, \boldsymbol{\eta}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_0^\infty \left(\frac{1}{2\pi\xi}\right)^{\frac{n}{2}} e^{\frac{-u}{2\xi}} p(\xi;\, \boldsymbol{\eta}_\xi) d\xi, \tag{3}$$

using a mixing variable $\xi \sim p(\xi;\, \boldsymbol{\eta}_\xi)$. Any mixing distribution $p(\xi;\, \boldsymbol{\eta}_\xi)$ that is strictly positive can be used to define a consistent elliptical process. In particular, we recover the Gaussian distribution if the mixing distribution is a Dirac delta function and the Student's t distribution if it is a scaled inverse chi-square distribution. For more information on the elliptical distribution, see Appendix A

2.2 Elliptical processes

The elliptical process is defined, analogously to a Gaussian process, as:

Definition 1 *An elliptical process ($\mathcal{EP}$) is a collection of random variables such that every finite subset has a consistent elliptical distribution, where the scale matrix is given by a covariance kernel.*

This means that an $\mathcal{EP}$ is specified by a mean function $\mu(\mathbf{x})$, scale matrix (kernel) $k(\mathbf{x}, \mathbf{x})$ and mixing distribution $p(\xi;\, \boldsymbol{\eta}_\xi)$. Since the $\mathcal{EP}$ is built upon consistent elliptical distributions it is closed under marginalization. The marginal mean $\boldsymbol{\mu}$ is the same as the mean for the Gaussian distribution, and the covariance is $\mathrm{Cov}[\mathbf{Y}] = \mathbb{E}[\xi]\, \boldsymbol{\Sigma}$ where $\mathbf{Y}$ is an elliptical random variable, $\boldsymbol{\Sigma}$ is the covariance for a Gaussian distribution and ξ is the mixing variable.

Formally a stochastic process $\{X_t : t \in T\}$ on a probability space $(\Omega, \mathcal{F}, P)$ consists of random maps $X_t : \omega \to S_t$, $t \in T$, for measurable spaces $(S_t, \mathcal{S}_t)$, $t \in T$ (Bhattacharya & Waymire, 2007). We focus on the setting where $S = \mathbb{R}$ and the index set T is a subset of $\mathbb{R}^N$, in particular, the half-line $[0, \infty)$. Due to Kolmogorov's extension theorem, we may construct the $\mathcal{EP}$ from the family of finite-dimensional, consistent, elliptical distributions, which is easy to check due to the restriction to $S = \mathbb{R}$ (which is a Polish space) and Kano's characterization above.

Ergodicity. When using $\mathcal{GP}$ for regression or classification we usually assume that the data originate from a single sample path, which is a single sample from the $\mathcal{GP}$. A stationary Gaussian process is ergodic since all sample paths have the same statistics. An elliptical process, on the other hand, can be viewed as a hierarchical model, constructed by first sampling $\xi \sim p(\xi;\, \boldsymbol{\eta}_\xi)$ and then $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}\, \xi)$. This structure implies that an elliptical process is *not* ergodic: it is not possible to infer the mixing distribution $p(\xi;\, \boldsymbol{\eta}_\xi)$ from a single path. To learn the mixing distribution $p(\xi;\, \boldsymbol{\eta}_\xi)$ we need draws from multiple paths.

Conditional distribution. To use the $\mathcal{EP}$ for predictions, we need the conditional mean and covariance of the corresponding elliptical distribution, which derive next. We partition the data as $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2]$, where $\mathbf{y}_1$ is the N_1 observed data points, $\mathbf{y}_2$ is the N_2 data points to predict, and $N_1 + N_2 = N$. We have the following result:

Proposition 1 *If the data $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2]$ originate from the consistent elliptical distribution in (3), the conditional distribution originates from the distribution*

$$p_{\mathbf{y_2}|u_1}(\mathbf{y}_2) = \frac{c_{N_1,\boldsymbol{\eta}}}{\left|\boldsymbol{\Sigma}_{22|1}\right|^{\frac{1}{2}} (2\pi)^{\frac{N_2}{2}}} \int_0^\infty \xi^{-\frac{n}{2}} e^{-(u_{2|1}+u_1)\frac{1}{2\xi}} p(\xi;\, \boldsymbol{\eta}) d\xi \tag{4}$$

with the conditional mean $\mathbb{E}[\boldsymbol{y_2}|\boldsymbol{y}_1] = \boldsymbol{\mu}_{2|1}$ and the conditional covariance

$$\mathrm{Cov}[\boldsymbol{Y}_2|\boldsymbol{Y}_1 = \boldsymbol{y}_2] = \mathbb{E}[\hat{\xi}]\boldsymbol{\Sigma}_{22|1}, \quad \hat{\xi} \sim \xi|\mathbf{y}_1, \tag{5}$$

where $u_1 = (\mathbf{y}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$, $u_{2|1} = (\mathbf{y}_2 - \boldsymbol{\mu}_{2|1})^\top \Sigma_{22|1}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_{2|1})$, and $c_{N_1,\boldsymbol{\eta}}$ is a normalization constant. The conditional scale matrix $\boldsymbol{\Sigma}_{22|1}$ and the conditional mean vector $\boldsymbol{\mu}_{2|1}$ are the same as the mean and the covariance matrix for a Gaussian distribution. The proof is derived in Appendix B.

The conditional distribution is guaranteed to be a consistent elliptical distribution but not necessarily the same as the original one—the shape depends on the training samples. (Recall that consistency only concerns the marginal distribution.)

We make predictions for unseen data points by first approximating the posterior, and then using its conditional distribution. Gaussian processes are computationally convenient since combining a $\mathcal{GP}$ prior with a Gaussian likelihood gives a Gaussian posterior. Unfortunately, this closure property does *not* hold for elliptical distributions in general. We therefore use variational inference to incorporate (non-Gaussian) noise according to the graphical models in Figure 3.

We aim to model mixing distributions that can capture any shape of the elliptical noise in the data. One way to learn complex probability distributions is to normalize flows, which we will now go through.

2.3 Flow based models

Normalizing flows are a family of generative models that map simple distributions to complex ones through a series of learned transformations (Papamakarios et al., 2021b). Suppose we have a random variable $\boldsymbol{x}$ that follows an unknown probability distribution $p_x(\boldsymbol{x})$. Then, the main idea of a normalizing flow is to express $\boldsymbol{x}$ as a transformation T_γ of a variable $\mathbf{z}$ with a known simple probability distribution $p_z(\mathbf{z})$. The transformation T_γ has to be bijective and invertible, and it can have learnable parameters γ. Both T and its inverse have to be differentiable. The probability density of $\boldsymbol{x}$ is obtained by a change of variables:

$$p_x(\boldsymbol{x}) = p_z(\boldsymbol{z}) \left| \det \left(\frac{\partial T_\gamma(\boldsymbol{z})}{\partial \boldsymbol{z}} \right) \right|^{-1}. \tag{6}$$

We focus on one-dimensional flows, since we are interested in modeling the mixing distribution. In particular, we use *linear rational spline flows* Dolatabadi et al. (2020); Durkan et al. (2019), wherein the mapping T_γ is an elementwise, monotonic linear rational spline: a piecewise function where each piece is a linear rational function. The parameters are the number of pieces (bins) and the knot locations.

To train the model parameters, we use amortized variational inference, which we go through next.

2.4 Amortized variational inference

In *amortized variational inference* (Gershman & Goodman, 2014) we replace the variational parameters, $\boldsymbol{\varphi}$ with a function that maps the input to the variational parameters $\boldsymbol{\varphi} = f(\boldsymbol{x})$. This is convenient for modelling local latent variables, i.e., variables associated directly to individual data points $\boldsymbol{x}_i$ which have corresponding variational parameters $\boldsymbol{\varphi}_i$. By replacing the local parameter with a function, $\boldsymbol{\varphi}_i = f(\boldsymbol{x}_i)$, we reduce the problem to fitting a function f, rather than fitting each $\boldsymbol{\varphi}_i$. Furthermore, it becomes easy to add new data points, since the local variational parameters are then given by the function f.

3 Method

We propose the variational $\mathcal{EP}$ with elliptical noise, where the variational $\mathcal{EP}$ can learn any consistent elliptical process, and the elliptical noise can capture any consistent elliptical noise. The key idea is to model the mixing distributions with a normalizing flow. The joint probability distribution of the model (see Figure 3c) is

$$p(\boldsymbol{y}, \boldsymbol{f}, \omega, \xi; \boldsymbol{\eta}) = \underbrace{p(\boldsymbol{f}|\xi; \boldsymbol{\eta}_{\boldsymbol{f}})p(\xi; \boldsymbol{\eta}_{\xi})}_{\text{prior}} \underbrace{\prod_{i=1}^{N} p(y_i|f_i, \omega)p(\omega; \boldsymbol{\eta}_{\omega})}_{\text{likelihood}}. \tag{7}$$

Here, $p(\boldsymbol{f}|\xi; \boldsymbol{\eta}_{\boldsymbol{f}}) \sim \mathcal{N}(0, K\xi)$ is a regular $\mathcal{EP}$ prior with the covariance kernel K containing the parameters $\boldsymbol{\eta}_{\boldsymbol{f}}$, $p(\xi; \boldsymbol{\eta}_{\xi})$ is the process mixing distribution and $p(\omega; \boldsymbol{\eta}_{\omega})$ is the noise mixing distribution.

To learn the mixing distributions $p(\xi; \boldsymbol{\eta}_{\xi})$ and $p(\omega; \boldsymbol{\eta}_{\omega})$ by gradient-based optimization, they need to be differentiable with respect to the parameters $\boldsymbol{\eta}_{\xi}$ and $\boldsymbol{\eta}_{\omega}$ in addition to being flexible and computationally efficient to sample and evaluate. Based on these criteria, a spline flow (Section 2.3) is a natural fit. We construct the mixing distributions by transforming a sample from a standard normal distribution with a spline flow. The output of the spline flow is then projected onto the positive real axis using a differentiable function such as *Softplus* or *Sigmoid*.

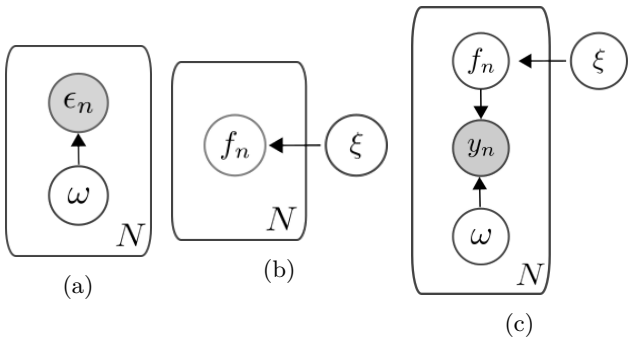

Figure 3: Graphical models of **(a)**, the elliptical likelihood, **(b)** the $\mathcal{EP}$-prior, and **(c)** the $\mathcal{EP}$ with independent elliptical noise.

In the following sections, we detail the construction of the model and show how to train it using variational inference. For clarity, we first describe the likelihood and the prior separately, before combining them and then describing a (computationally efficient) sparse approximation.

3.1 Likelihood

By definition, the likelihood (Figure 3a) describes the measurement noise ϵ_n (Equation (1)). The probability distribution of the independent elliptical likelihood is,

$$p(\epsilon_n; \sigma, \boldsymbol{\eta}_{\omega}) = \int \mathcal{N}(\epsilon_n; 0, \sigma^2\omega)p(\omega; \boldsymbol{\eta}_{\omega})d\omega, \tag{8}$$

where σ can be set to one without loss of generality. In other words, the likelihood is a continuous mixture of Gaussian distributions where, e.g., ϵ_n follows a Student's t distribution if ω is scaled chi-squared distributed. We parameterize $p(\omega; \boldsymbol{\eta}_{\omega})$ as a spline flow,

$$p(\omega; \boldsymbol{\eta}_{\omega}) = p(\zeta) \left| \frac{\partial f(\zeta; \boldsymbol{\eta}_{\omega})}{\partial \zeta} \right|^{-1} \tag{9}$$

although it could, in principle, be any positive, finite probability distribution. Here, $p(\zeta) \sim \mathcal{N}(0, 1)$, and $f(\,\cdot\,; \boldsymbol{\eta}_{\omega})$ represent the spline flow transformation followed by a *Softplus* transformation to guarantee ω to be positive. Now, assume that we observe N independent and identically distributed residuals $\epsilon_n = y_n - f_n$ between the observations $\boldsymbol{y}$ and some function, $\boldsymbol{f}$. We train the likelihood by maximizing the (log) marginal likelihood with respect to the parameters $\boldsymbol{\eta}_{\omega}$, that is

$$\log p(\boldsymbol{\epsilon}; \boldsymbol{\eta}_{\omega}) = \sum_{n=1}^{N} \log \int p(\epsilon_n|f(\zeta; \boldsymbol{\eta}_{\omega})) \left| \frac{\partial f(\zeta; \boldsymbol{\eta}_{\omega})}{\partial \zeta} \right|^{-1} p(\zeta)d\zeta. \tag{10}$$

For general mixing distributions this integral is intractable, but we can use variational inference to approximate it, thereby maximizing the evidence lower bound (ELBO) instead of the marginal likelihood directly. Hence, we approximate the integral (10), by approximating the posterior of the latent variables $q(\zeta_n;\, \boldsymbol{\varphi}_{\zeta_n}) \approx p(\zeta|\epsilon_n)$, where $\boldsymbol{\varphi}_{\zeta_n}$ are the variational parameters, which gives us the following ELBO,

$$\mathcal{L}(\boldsymbol{\eta}_\omega, \boldsymbol{\varphi}_{\zeta_1}, \cdots, \boldsymbol{\varphi}_{\zeta_N}) = \sum_{n=1}^{N} \mathbb{E}_{\zeta_n \sim q(\zeta_n;\, \boldsymbol{\varphi}_{\zeta_n})} \left[\log p(\epsilon_n | f(\zeta;\, \boldsymbol{\eta}_\omega)) \left| \frac{\partial f(\zeta;\, \boldsymbol{\eta}_\omega)}{\partial \zeta} \right|^{-1} p(\zeta) - \log q(\zeta_n;\, \boldsymbol{\varphi}_{\zeta_n}) \right]. \quad (11)$$

We have one set of variational parameters $\boldsymbol{\varphi}_{\zeta_n}$ per observed noise ϵ_n for the variational approximation above. To reduce the complexity of the model we amortize (Section 2.4) the variational parameters by letting $\boldsymbol{\varphi}_{\zeta_i} = f(\epsilon_i;\, \boldsymbol{\gamma}_\zeta)$, which reduces the ELBO to $\mathcal{L}(\boldsymbol{\eta}_\omega, \boldsymbol{\gamma}_\zeta)$.

Since we want a posterior that is similar to the prior distribution $p(\zeta)$, we set it to be normally distributed $q(\zeta_n) = \mathcal{N}(\mu_{\boldsymbol{\gamma}_\zeta}(\epsilon_n)\sigma_{\boldsymbol{\gamma}_\zeta}(\epsilon_n))$. The mean and variance functions, $\mu_{\boldsymbol{\gamma}_\zeta}(\,\cdot\,)$ and $\sigma_{\boldsymbol{\gamma}_\zeta}(\,\cdot\,)$ are parameterized by neural networks. We train the model by gradient-based optimization of the ELBO in parameter space, $\nabla_{\boldsymbol{\eta}_\omega, \boldsymbol{\gamma}_\zeta} \mathcal{L}(\boldsymbol{\eta}_\omega, \boldsymbol{\gamma}_\zeta)$. The gradients are estimated using black-box variational inference (Wingate & Weber, 2013; Ranganath et al., 2014).

3.2 Prior

We use the same idea as with the elliptical likelihood to construct the elliptical process ($\mathcal{EP}$). The joint distribution of the elliptical prior(see Equation (7) and Figure 3b) is

$$p(\boldsymbol{f}, \xi;\, \boldsymbol{\eta}) = p(\boldsymbol{f}|\xi;\, \boldsymbol{\eta}_{\boldsymbol{f}}) p(\xi;\, \boldsymbol{\eta}_\xi). \quad (12)$$

Given N observations $(\boldsymbol{x}_i, y_i)$, which we assume are corrupted by Gaussian noise with known variance σ^2, we get the following marginal likelihood

$$p(\boldsymbol{y};\, \boldsymbol{\eta}_{\boldsymbol{f}}, \boldsymbol{\eta}_\xi) = \int p(\boldsymbol{y}, \boldsymbol{f}, \xi;\, \boldsymbol{\eta}_{\boldsymbol{f}}, \boldsymbol{\eta}_\xi) d\boldsymbol{f} d\xi = \int \prod_{i=1}^{N} p(y_i|\boldsymbol{f};\, \sigma^2) p(\boldsymbol{f}|\xi;\, \boldsymbol{\eta}_{\boldsymbol{f}}) p(\xi;\, \boldsymbol{\eta}_\xi) d\boldsymbol{f} d\xi. \quad (13)$$

In this model we have to marginalize over two latent variables, $\boldsymbol{f}$ and ξ. However, this integral is intractable—just as it was for the elliptical likelihood—since $p(\xi;\, \boldsymbol{\eta}_\xi)$ is parameterized by a spline flow. To overcome this we use the same procedure as for the likelihood model and approximate the marginal likelihood with the ELBO

$$\mathcal{L}(\boldsymbol{\eta}_{\boldsymbol{f}}, \boldsymbol{\eta}_\xi, \boldsymbol{\varphi}_{\boldsymbol{f}}, \boldsymbol{\varphi}_\xi) = \mathbb{E}_{q(\boldsymbol{f}, \xi;\, \boldsymbol{\varphi}_{\boldsymbol{f}}, \boldsymbol{\varphi}_\xi)} \left[\log p(\mathbf{y}, \boldsymbol{f}, \xi;\, \boldsymbol{\eta}_{\boldsymbol{f}}, \boldsymbol{\eta}_\xi) - \log q(\boldsymbol{f}, \xi;\, \boldsymbol{\varphi}_{\boldsymbol{f}}, \boldsymbol{\varphi}_\xi) \right], \quad (14)$$

where $q(\boldsymbol{f}, \xi;\, \boldsymbol{\varphi}_{\boldsymbol{f}}, \boldsymbol{\varphi}_\xi)$ is an approximation of the posterior $p(\boldsymbol{f}, \xi|\mathbf{y};\, \boldsymbol{\eta}_{\boldsymbol{f}}, \boldsymbol{\eta}_\xi)$. Importantly, this approximate posterior is not only a tool for maximizing the marginal likelihood, but it is also to used to make predictions for unseen data points.

We assume that the approximate posterior is elliptical, and we do so for two reasons: first, this makes the approximate posterior similar to the true posterior, and second, we can then use the conditional distribution to make predictions. Specifically, we factorize the posterior as

$$q(\boldsymbol{f}, \xi, \omega;\, \boldsymbol{\varphi}) = q(\boldsymbol{f}|\xi;\, \boldsymbol{\varphi}_{\boldsymbol{f}}) q(\xi;\, \boldsymbol{\varphi}_\xi), \quad (15)$$

where $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_{\boldsymbol{f}}, \boldsymbol{\varphi}_\xi)$ are the variational parameters, $q(\boldsymbol{f}|\xi;\, \boldsymbol{\varphi}_{\boldsymbol{f}}) = \mathcal{N}(\mathbf{m}, \mathbf{S}\xi)$ is a Gaussian distribution with the variational mixing distribution $\xi \sim q(\xi;\, \boldsymbol{\varphi}_\xi)$. Again, $q(\xi;\, \boldsymbol{\varphi}_\xi)$ could be any positive finite distribution, but we parameterize it with a spline flow. This factorization enables predictions on unseen data points, $\boldsymbol{x}^*$, according to

$$p(f^*|\mathbf{y}) = \int p(f_*|\boldsymbol{f}, \xi;\, \boldsymbol{\eta}_{\boldsymbol{f}}) q(\boldsymbol{f}|\xi;\, \boldsymbol{\varphi}_{\boldsymbol{f}}) q(\xi;\, \boldsymbol{\varphi}_\xi) d\boldsymbol{f} d\xi = \mathbb{E}_{q(\xi;\, \boldsymbol{\varphi}_\xi)} \left[\mathcal{N}(f_*|m_*, s_*\xi) \right], \quad (16)$$

where m_* and s_* are derived, in a similar fashion as for the variational Gaussian process, see Appendix C.

Combining the $\mathcal{EP}$ with an elliptical likelihood results in the joint probability in Equation (7), which corresponds to the graphical model in Figure 3c. The posterior is then approximated by

$$p(\boldsymbol{f}, \xi, \boldsymbol{\zeta}|\mathbf{y}; \boldsymbol{\eta_f}, \boldsymbol{\eta}_\xi, \boldsymbol{\eta}_\omega) \approx q(\boldsymbol{f}|\xi; \boldsymbol{\varphi_f}) q(\xi; \boldsymbol{\varphi}_\xi) \prod_{i=1}^{N} q\left(\zeta_i; \boldsymbol{\varphi}_{\zeta_i} = f(y_i, f_i; \boldsymbol{\gamma}_\zeta)\right). \tag{17}$$

The flexibility of this flow-based construction lets us capture a broad range of elliptical processes. We can also specify an appropriate likelihood ourselves. For instance, using a categorical likelihood enables $\mathcal{EP}$ classification. The model is trained with stochastic gradient descent and black-box variational inference (Wingate & Weber, 2013; Ranganath et al., 2014).

3.3 Sparse elliptical process

To create a computationally tractable model for large datasets, we derive a sparse version of the model using the variational inference framework. Our particular factorization of the variational posterior (15) makes this straightforward due to the similarity with the sparse variational $\mathcal{GP}$ posterior (Titsias, 2009). Assuming we have a sparse $\mathcal{EP}$ prior and fixed Gaussian noise, we get a joint model

$$p(\mathbf{y}, \boldsymbol{f}, \boldsymbol{u}, \xi; \boldsymbol{\eta}) = p(\mathbf{y}|\boldsymbol{f})(\boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta_f}) p(\boldsymbol{u}|\xi; \boldsymbol{\eta_u}) p(\xi; \boldsymbol{\eta}_\xi), \tag{18}$$

where $\boldsymbol{u}$ are outputs of the elliptical process located at the inducing inputs $\mathbf{x}_u$. Intending to make predictions with the model, we approximate the posterior over the latent variables in the same ways as (15), with

$$p(\boldsymbol{f}, \boldsymbol{u}, \xi|\mathbf{y}; \boldsymbol{\eta}) \approx p(\boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta_f}) q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) q(\xi; \boldsymbol{\varphi}_\xi), \tag{19}$$

where $q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) \sim \mathcal{N}(\mathbf{m}_u, \mathbf{S}_u \xi)$ and$q(\xi; \boldsymbol{\varphi}_\xi)$ are variational distributions. We train the model by minimizing the ELBO defined in (14) and make predictions for unseen data points x^* using the predictive approximated posterior

$$p(f^*|\mathbf{y}) = \int p(f_*|\boldsymbol{u}, \xi) q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) q(\xi; \boldsymbol{\varphi}_\xi) d\boldsymbol{u} d\xi = \mathbb{E}_{q(\xi; \boldsymbol{\varphi}_\xi)} \left[\mathcal{N}(f_*|m^*_{\boldsymbol{u}}, s^*_{\boldsymbol{u}}\xi)\right], \tag{20}$$

where $m^*_{\boldsymbol{u}}$ and $s^*_{\boldsymbol{u}}$ are derived in Appendix C, in a similar fashion as for the $\mathcal{GP}$ in Titsias (2009).

3.4 Extension to heteroscedastic noise

We extend the elliptical likelihood by modeling heteroscedastic noise. First, recall from Section 3.1 that we amortized the variational mixing distribution for the elliptical likelihood. Here, we describe how we can model elliptic heteroscedastic noise by letting the parameters $\boldsymbol{\eta}_\omega$ of the mixing distribution of the likelihood depend on the input location.

In heteroscedastic regression, the noise depends on the input location $\boldsymbol{x}_n$. For example, heteroscedastic elliptical noise can be useful when we have a time series where the noise variance and tail-heaviness change over time. Examples of this can be in statistical finances (Liu et al., 2020) and robotics (Kersting et al., 2007). To model this, we let the spline flow mixing distribution parameters $\boldsymbol{\eta}_\omega$ depend on the input location, $\boldsymbol{x}_n$, such that $\boldsymbol{\eta}_{\omega_n} = f(\boldsymbol{x}_n, \gamma_{p,\omega})$. For the variational spline parameters $\boldsymbol{\varphi}_{\omega,n}$ in addition to making them depend on the noise ϵ_n, we also make them depend on the input location, such that $\boldsymbol{\varphi}_{\omega_n} = f(\boldsymbol{x}_n, \epsilon_n; \gamma_{\omega,q})$. This results in an elliptical likelihood with a location-dependent mixing distribution. Finally, we train the model by minimizing the ELBO

$$\mathcal{L}(\boldsymbol{\gamma}_{\omega,p}, \boldsymbol{\gamma}_{\omega,q}) = \sum_{n=1}^{N} \mathbb{E}_{\zeta_n \sim q(\omega_n; \boldsymbol{\varphi}_{\omega_n})} \left[\log p(\epsilon_n|\omega_n) p(\omega_n; \boldsymbol{\eta}_{\omega_n}) - \log q(\omega_n; \boldsymbol{\varphi}_{\omega_n})\right], \tag{21}$$

where $\boldsymbol{\varphi}_{\omega_n} = f(\epsilon_n, \boldsymbol{x}_n; \boldsymbol{\gamma}_{\omega,q})$ and $\boldsymbol{\eta}_{\omega_n} = f(\boldsymbol{x}_n; \boldsymbol{\gamma}_{\omega,p})$. This model can be extended by adding extra information in the spline flow input.

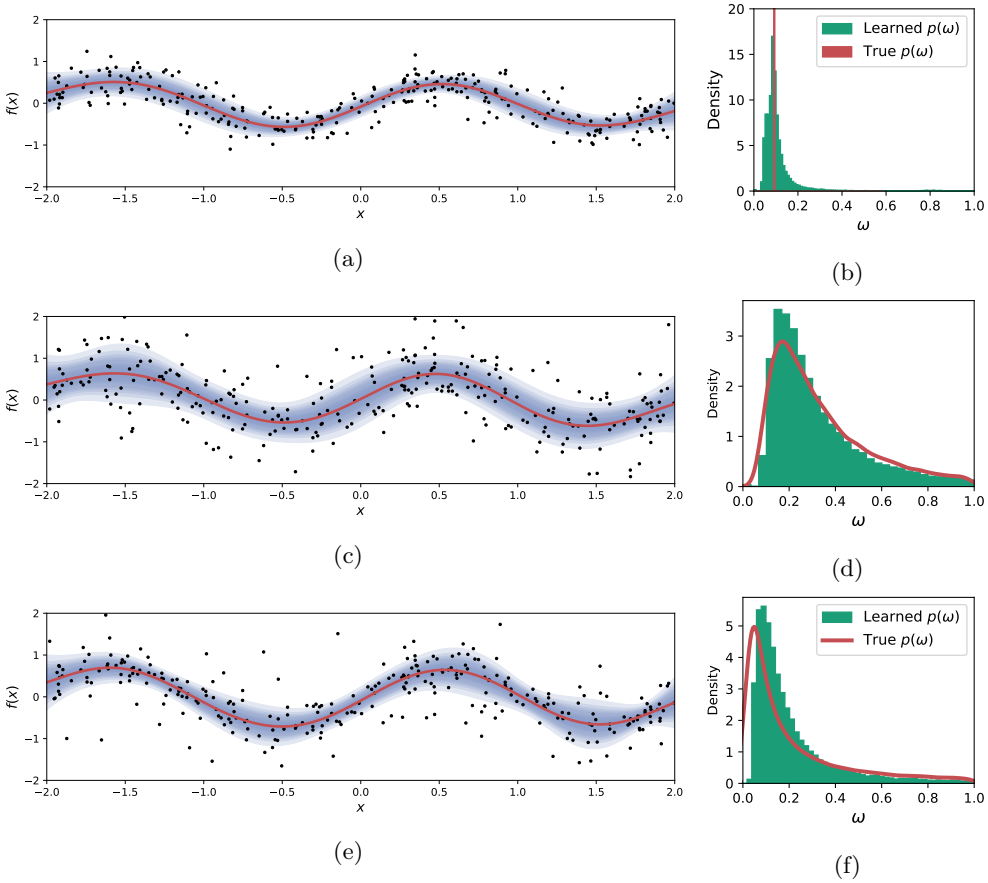

Figure 4: The posterior predictive distribution when using an $\mathcal{GP}$ with elliptical noise, modeled with a spline flow. The histograms show the learned and the true noise mixing distribution.

4 Experiments

We examine the variational elliptical processes using four different experiments. In the first experiment, we investigate how well the elliptical likelihood (Section 3.1) recover known elliptical noise in synthetic data. In the second experiment, we investigate the benefits of using the sparse $\mathcal{EP}$ compared to the sparse $\mathcal{GP}$ for regression on standard benchmarks. In the third experiment, we examine if using a $\mathcal{EP}$ is beneficial in classifications tasks. Finally, in the last experiment, we investigate the amortized elliptical processes described in Section 3.4, to model heteroscedastic noise.

Implementation. The mixing distribution of the variational $\mathcal{EP}$ uses a linear rational spline flow, where we transform the flow using *Softplus* to ensure that it is bounded from below and positive. In all experiments, we use a squared exponential kernel. See Appendix E for further details. The code from the experiments will be published on GitHub if the paper is accepted, with a link added here.

4.1 Noise identification

To examine how well the elliptical likelihood, described in Section 3.1, can capture different types of elliptical noises, we created three equal synthetic datasets, each with $N = 300$ data points, by using the function $y_n = \sin(3\boldsymbol{x}_n)/2$, where $\boldsymbol{x} \in \mathbb{R}$ is uniformly sampled, $\boldsymbol{x}_n \sim U(-2, 2)$. Each of the dataset has its own independent elliptical noise, whose mixing distribution is plotted in Figure 4.

Table 1: Predictive Mean Square Error (MSE) and the predictive log-likelihood (LL) on the hold out sets from the experiments. We show the average of the ten folds and one standard deviation in parenthesis. Bold font indicates the best result obtained.

	CONCR.		**MACHINE**		**MPG**		**CALIFOR.**	
$m = 20$	MSE	neg LL	MSE	neg LL	MSE	neg LL	MSE	neg LL
$\mathcal{GP}$	0.27(0.044)	0.74(0.041)	0.50(0.70)	−0.29(0.24)	0.15(0.54)	0.35(0.12)	0.44(0.021)	0.91(0.022)
$\mathcal{EP}^1$	**0.18**(0.028)	**0.47**(0.049)	**0.33**(0.57)	**−0.30**(0.20)	0.13(0.043)	**0.37**(0.098)	**0.31**(0.016)	**0.67**(0.026)
$\mathcal{EP}^2$	**0.18**(0.024)	**0.47**(0.044)	0.34(0.58)	**−0.30**(0.23)	**0.12**(0.047)	0.21(0.19)	0.32(0.013)	0.71(0.019)
$m = 50$								
$\mathcal{GP}$	0.21(0.032)	0.61(0.038)	0.35(0.62)	0.12(0.21)	0.16(0.054)	0.78(0.044)	0.33(0.016)	0.74(0.024)
$\mathcal{EP}^1$	0.18(0.024)	0.48(0.045)	**0.30**(0.56)	**−0.29**(0.28)	**0.12**(0.042)	**0.35**(0.096)	**0.30**(0.015)	**0.64**(0.024)
$\mathcal{EP}^2$	**0.17**(0.025)	**0.42**(0.048)	0.32(0.60)	−0.24(0.29)	**0.12**(0.039)	**0.35**(0.097)	0.31(0.015)	0.68(0.024)

We trained a sparse variational $\mathcal{GP}$ with a variational elliptical likelihood for each of the datasets. The spline flow is parameterized using nine bins, which we found was suitable using a simple parameter search. We trained the $\mathcal{GP}$-kernel parameters and the likelihood parameters simultaneously by maximizing the ELBO (14). The results from the experiments are illustrated in Figure 4.

The figures show the histogram of the true elliptical noise mixing distribution and the learned mixing distribution next to each other. We see that the learned distribution follows the shape of the true mixing distribution. We also plot the resulting predictive $\mathcal{GP}$-posterior to show that the model learned suitable kernel parameters simultaneously as learning the likelihood mixing distribution.

4.2 Regression

We investigated two versions of the sparse variational $\mathcal{EP}$ (see Section 3.3) together with the variational $\mathcal{GP}$ on four different regression datasets. The two $\mathcal{EP}$ versions were both modeled with variational elliptical noise. The two versions differ in the prior processes. For the **first model** ($\mathcal{EP}^1$) we used a fixed prior process mixing distribution that followed an inverse chi-square distribution with 20 degrees of freedom and a posterior mixing distribution $q(\xi; \boldsymbol{\varphi}_\xi)$, modeled with a spline flow. The **second model**, ($\mathcal{EP}^2$), was constructed using both a $\mathcal{GP}$-prior and a $\mathcal{GP}$-posterior.

We can see the first model ($\mathcal{EP}^1$) as having a prior process close to Gaussian and a general elliptical posterior process, which could transform from the information in the likelihood and the data. The second model ($\mathcal{EP}^2$) corresponded to a $\mathcal{GP}$ process with general elliptical noise.

We run the models using 20 and 50 inducing points. The data came from Kibler et al. (1989) and the spatial interpolation data came from Dubois et al. (2003) (see Appendix G for further information about the data). We performed ten-fold cross-validation on the datasets. Table 1 gives the mean squared error (MSE) and the test log-likelihood (LL) for the hold-out datasets.

For these datasets, the $\mathcal{EP}^1$ and $\mathcal{EP}^2$ are almost always performing better then the $\mathcal{GP}$, especially when the number of inducing points is 20. This is according to our hypothesis: the $\mathcal{GP}$, with its thin tail, overfits more often to the outliers while a heavier tail perceives the outliers as noise. For most datasets, it seems sufficient to add elliptical noise to a variational $\mathcal{GP}$ prior. Only for the small and noisy datasets, such as the machine dataset, we see an improvement with adding a general $\mathcal{EP}$ posterior.

4.3 Binary classification

To evaluate the $\mathcal{EP}$ on classifications tasks we perform variational $\mathcal{EP}$ and $\mathcal{GP}$ classification by simply replacing the likelihood with a binary one. This realization is interesting since here, we do not have a likelihood that captures the noise in the data, but instead, the process itself has to do it. Therefore, we can indicate the value of the elliptical process itself without the elliptical noise. We compared two sparse

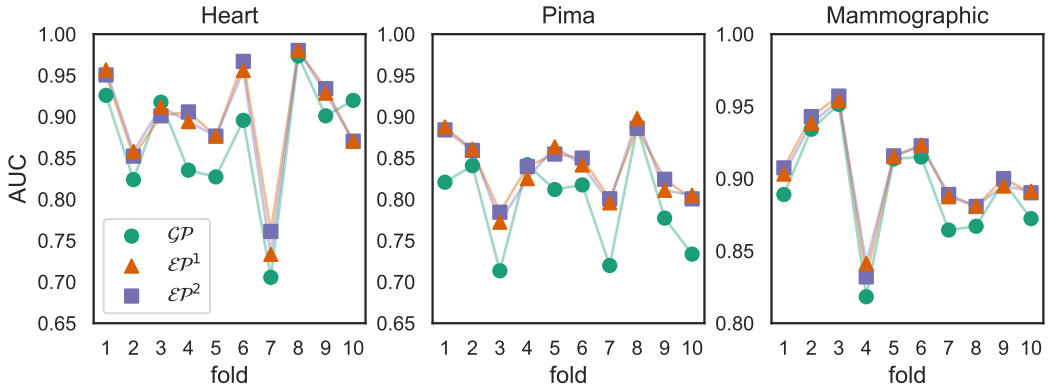

Figure 5: The classification AUC score from the five-fold cross validation. The $\mathcal{EP}$ outperforms the two other models on both datasets, also when the input is corrupted with 20% noise.

$\mathcal{EP}$ models with a sparse $\mathcal{GP}$ model using 20 inducing points. The two $\mathcal{EP}$s differed in the prior mixing distribution. For the first model ($\mathcal{EP}^1$) we used a fixed prior mixing distribution set to a scaled inverse chi-square distribution with ten degrees of freedom. For the second model ($\mathcal{EP}^2$), we set the prior mixing distribution to a spline flow with six bins which means that we train it. We can see the trainable prior mixing distribution as using a continuously scaled mixture of Gaussian processes, which can be more expressive than a single $\mathcal{GP}$.

We trained the models on three classification datasets, described in Appendix G. The results from a ten-fold cross-validation is presented in Figure 5. From the area under the curve (AUC) score, we see that the $\mathcal{EP}$ separates the two classes better. It seems that mainly the variational elliptical distribution contributes to the higher AUC score. Training the mixing distribution of the $\mathcal{EP}$ prior did not improve the score.

4.4 Elliptic heteroskedastic noise

In this experiment, we aimed to learn heteroscedastic noise as described in Section 3.4 on a synthetic dataset of 150 samples, see Figure 6. We created the dataset using the function $f(x) = \sin(5x) + x$. We then added Student's t noise, $\epsilon(x) \sim St(\nu(x), \sigma(x)$, where we decreased the noise scale by $\nu(x) = 25 - 11|x + 1|^{0.9}$, and the increased the standard deviation by $\sigma(x) = 0.5|x + 1|^{1.6} + 0.001$. We used a variational sparse $\mathcal{GP}$ with heteroscedastic noise as described in Section 3.4.

We used six bins for the prior mixing distribution and eight bins for the posterior mixing distribution, which resulted in 19 and 35 parameters to predict, respectively. We had more bins for the posterior mixing distribution since we wanted the approximate posterior to be as flexible as possible to fit the true posterior.

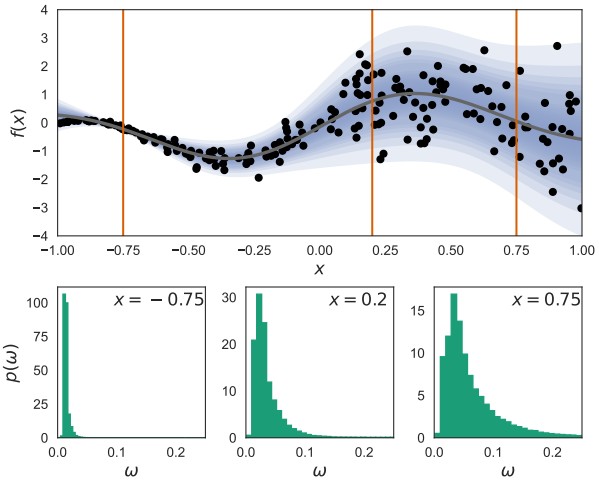

Figure 6: The result form training a $\mathcal{GP}$ process with heteroscedastic elliptical noise on a synthetic dataset. The histogram shows the noise resulting mixing distributions ad different $\boldsymbol{x}_i$.

The results from the experiments are depicted in Figure 6 and show that the model was able to capture the varying noise, both in term of the scale and the increasing heaviness of the tail. A single spike in the mixing distribution indicates that the noise is Gaussian, and the *wider* the mixing distribution is, the heavier tailed the noise is.

5 Related work

In general, attempts at modeling heavy-tailed stochastic processes modify either the likelihood or the stochastic process prior—rarely both. Approximate inference is typically needed when going beyond Gaussian likelihoods (Neal, 1997; Jylänki et al., 2011), e.g., for robust regression, but approximations that preserve analytical tractability have been proposed (Shah et al., 2014).

Ma et al. (2019) describes a class of stochastic processes where the finite-dimensional distributions are only defined implicitly as a parameterized transformation of some base distribution, thereby generalizing earlier work on warped Gaussian processes (Snelson et al., 2004; Rios & Tobar, 2019). However, the price of this generality is that standard variational inference is no longer possible. Based on an assumption of a Gaussian likelihood, they describe an alternative based on the wake-sleep algorithm by Hinton et al. (1995).

Other attempts at creating more expressive $\mathcal{GP}$ priors include Maroñas et al. (2021), who used a $\mathcal{GP}$ in combination with a normalizing flow, and Luo & Sun (2017), who used a discrete mixture of Gaussian processes. Similar ideas combining mixtures and normalizing flows have also been proposed to create more expressive likelihoods (Abdelhamed et al., 2019; Daemi et al., 2019; Winkler et al., 2019; Rivero & Dvorkin, 2020) and variational posteriors (Nguyen & Bonilla, 2014). Non-stationary extensions of Gaussian processes, such as when modeling heteroscedastic noise, are quite rare but the mixture model of Li et al. (2021) and the variational model of Lázaro-Gredilla & Titsias (2011) are two examples.

In the statistics literature, it is well-known that the elliptical processes can be defined as scale-mixtures of Gaussian processes (Huang & Cambanis, 1979; O'Hagan et al., 1999). However, unlike in machine learning, little emphasis is placed on building the models from data (i.e., training). These models have found applications in environmental statistics because of the field's inherent interest in modeling spatial extremes (Davison et al., 2012). Several works take the mixing distribution as the starting point, like us, and make localized predictions of quantiles (Maume-Deschamps et al., 2017) or other tail-risk measures (Opitz, 2016).

6 Conclusions

The Gaussian distribution is the default choice in statistical modeling for good reasons. Even so, far from everything is Gaussian—casually pretending it is, comes at a risk. The elliptical distribution offers a computationally tractable alternative that can capture heavy-tailed distributions. The same reasoning applies when comparing the Gaussian process to the elliptical process. We believe that a sensible approach in many applications would be to start from the weaker assumptions of the elliptical process and let the data decide whether the evidence supports gaussianity.

We constructed the elliptical processes as a scale mixture of Gaussian distributions. By parameterizing the mixing distribution using a normalizing flow, we show how a corresponding elliptical process can be trained using variational inference. The variational approximation we propose enables us to capture heavy-tailed posteriors and makes it straightforward to create a sparse variational elliptical process that scales to large datasets.

We performed experiments on robust regression and classification. In addition, we compared the elliptical processes with the Gaussian process. Our experiments show that, as expected, the elliptical process was more accurate in the presence of outliers or heavy-tailed noise.

The added flexibility of the elliptical processes could benefit a range of applications, both classical and new. However, advanced statistical models are not a cure-all, and one needs to avoid over-reliance on such models, especially in safety-critical applications.

References

Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3165–3173, 2019.

Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, and Francisco Herrera. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011.

Rabindra Nath Bhattacharya and Edward C Waymire. *A basic course in probability theory*, volume 69. Springer, 2007.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Atefeh Daemi, Hariprasad Kodamana, and Biao Huang. Gaussian process modelling with Gaussian mixture likelihood. *Journal of Process Control*, 81:209–220, 2019.

Andreas Damianou and Neil D Lawrence. Deep Gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 207–215, 2013.

Anthony C Davison, Simone A Padoan, Mathieu Ribatet, et al. Statistical modeling of spatial extremes. *Statistical science*, 27(2):161–186, 2012.

Peter J Diggle, Jonathan A Tawn, and Rana A Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. Invertible generative modeling using linear rational splines. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4236–4246, 2020.

Grégoire Dubois, Jacek Malczewski, and Marc De Cort. *Mapping radioactivity in the environment: Spatial interpolation comparison 97*. Office for Official Publications of the European Communities, 2003.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric multivariate and related distributions*. Chapman and Hall, 1990.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 351–360, 2015.

Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Steel T Huang and Stamatis Cambanis. Spherically invariant processes: Their nonlinear structure, discrimination, and estimation. *Journal of Multivariate Analysis*, 9(1):59–83, 1979.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12(Nov):3227–3257, 2011.

Yutaka Kano. Consistency property of elliptic probability density functions. *Journal of Multivariate Analysis*, 51(1):139–147, 1994.

Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic Gaussian process regression. In *International Conference on Machine Learning (ICML)*, pp. 393–400, 2007.

Dennis Kibler, David W Aha, and Marc K Albert. Instance-based prediction of real-valued attributes. *Computational Intelligence*, 5(2):51–57, 1989.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic Gaussian process regression. In *ICML*, 2011.

Tao Li, Di Wu, and Jinwen Ma. Mixture of robust Gaussian processes and its hard-cut EM algorithm with variational bounding approximation. *Neurocomputing*, 452:224–238, 2021.

Bingqing Liu, Ivan Kiskin, and Stephen Roberts. An overview of Gaussian process regression for volatility forecasting. In *International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 681–686, 2020.

Chen Luo and Shiliang Sun. Variational mixtures of Gaussian processes for classification. In *IJCAI*, volume 357, pp. 4603–4609, 2017.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In *International Conference on Machine Learning (ICML)*, pp. 4222–4233, 2019.

Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419, 1963.

Juan Maroñas, Oliver Hamelijnck, Jeremias Knoblauch, and Theodoros Damoulas. Transforming Gaussian processes with normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pp. 1081–1089, 2021.

Véronique Maume-Deschamps, Didier Rullière, and Antoine Usseglio-Carleve. Quantile predictions for elliptical random fields. *Journal of Multivariate Analysis*, 159:1–17, 2017.

Radford M Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report 9702, Department of Statistics, University of Toronto, 1997.

Trung V Nguyen and Edwin V Bonilla. Automated variational inference for Gaussian process models. *Advances in Neural Information Processing Systems*, 27, 2014.

Sebastian W Ober and Laurence Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In *International Conference on Machine Learning (ICML)*, pp. 8248–8259, 2021.

Anthony O'Hagan, Marc C Kennedy, and Jeremy E Oakley. Uncertainty analysis and other inference tools for complex computer codes. In *Bayesian statistics 6*, pp. 503–524. Oxford University Press, 1999.

Thomas Opitz. Modeling asymptotically independent spatial extremes based on Laplace random fields. *Spatial Statistics*, 16:1–18, 2016.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3): 291–297, 1997.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021a.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021b.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822, 2014.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, pp. 1530–1538, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, pp. 1278–1286, 2014.

Gonzalo Rios and Felipe Tobar. Compositionally-warped Gaussian processes. *Neural Networks*, 118:235–246, 2019.

Ana Diaz Rivero and Cora Dvorkin. Flow-based likelihoods for non-Gaussian inference. *Physical Review D*, 102(10):103507, 2020.

Hugh Salimbeni, Vincent Dutordoir, James Hensman, and Marc Deisenroth. Deep Gaussian processes with importance-weighted variational inference. In *International Conference on Machine Learning (ICML)*, pp. 5589–5598, 2019.

Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 877–885, 2014.

Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, pp. 261. American Medical Informatics Association, 1988.

Edward Snelson, Carl Edward Rasmussen, and Zoubin Ghahramani. Warped Gaussian processes. *Advances in neural information processing systems*, 16:337–344, 2004.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics*, pp. 567–574, 2009.

Dustin Tran, Rajesh Ranganath, and David M Blei. Variational Gaussian process. In *International Conference on Learning Representations (ICLR)*, 2016.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Andrew G Wilson, Zhiting Hu, Russ R Salakhutdinov, and Eric P Xing. Stochastic variational deep kernel learning. *Advances in Neural Information Processing Systems*, 29, 2016.

David Wingate and Theophane Weber. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*, 2013.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

I-C Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.

Abdelhak M Zoubir, Visa Koivunen, Yacine Chakhchoukh, and Michael Muma. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *IEEE Signal Processing Magazine*, 29(4): 61–80, 2012.

A The elliptical distribution

The Gaussian distribution—the basic building block of Gaussian processes—has several attractive properties that we wish the elliptical process to inherit, namely (i) closure under marginalization, (ii) closure under conditioning, and (iii) straightforward sampling. This leads us to consider the family of *consistent* elliptical distributions. Following Kano (1994), we say that a family of elliptical distributions $\{p(u(\mathbf{y}_N);\, \boldsymbol{\eta}) \,|\, N \in \mathbb{N}\}$ is consistent if and only if

$$\int_{-\infty}^{\infty} p\left(u(\mathbf{y}_{N+1});\, \boldsymbol{\eta}\right) dy_{N+1} = p\left(u(\mathbf{y}_N);\, \boldsymbol{\eta}\right). \tag{22}$$

In other words, a consistent elliptical distribution is closed under marginalization.

Far from all elliptical distributions are consistent, but the complete characterization of those that are is provided by the following theorem (Kano, 1994).

Theorem 1 *An elliptical distribution is consistent if and only if it originates from the integral*

$$p(u;\, \boldsymbol{\eta}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_0^{\infty} \left(\frac{1}{\xi 2\pi}\right)^{\frac{N}{2}} e^{\frac{-u}{2\xi}} p(\xi;\, \boldsymbol{\eta}_\xi) d\xi, \tag{23}$$

where ξ is a mixing variable with the corresponding, strictly positive finite, mixing distribution $p(\xi;\, \boldsymbol{\eta})$, that is independent of N.

This shows that consistent elliptical distributions $p(u;\, \boldsymbol{\eta})$ are scale-mixtures of Gaussian distributions, with a mixing variable $\xi \sim p(\xi;\, \boldsymbol{\eta})$. Note that any mixing distribution fulfilling Theorem 1 can be used to define a consistent elliptical process. We recover the Gaussian distribution if the mixing distribution is a Dirac delta function and the Student's t distribution if it is a scaled chi-square distribution.

If $p(u;\, \boldsymbol{\eta})$ is a scale-mixture of normal distributions, it has the stochastic representation

$$\mathbf{Y}|\, \xi \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}\xi), \quad \xi \sim p(\xi;\, \boldsymbol{\eta}). \tag{24}$$

By using the following representation of the elliptical distribution,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Z}\xi^{1/2}, \tag{25}$$

where $\mathbf{Z}$ follows the standard normal distribution, we get the mean

$$\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbb{E}\left[\mathbf{Z}\right]\ \ \mathbb{E}[\xi^{1/2}] = \boldsymbol{\mu} \tag{26}$$

and the covariance

$$\begin{aligned}
\mathrm{Cov}(\mathbf{Y}) &= \mathbb{E}\left[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^\top\right] \\
&= \mathbb{E}\left[(\boldsymbol{\Sigma}^{1/2}\mathbf{Z}\sqrt{\xi})(\boldsymbol{\Sigma}^{1/2}\mathbf{Z}\sqrt{\xi})^\top\right] \\
&= \mathbb{E}\left[\xi\boldsymbol{\Sigma}^{1/2}\mathbf{Z}\mathbf{Z}^\top(\boldsymbol{\Sigma}^{1/2})^\top\right] \\
&= \mathbb{E}\left[\xi\right]\boldsymbol{\Sigma}.
\end{aligned} \tag{27}$$

The variance is a scale factor of the scale matrix $\boldsymbol{\Sigma}$. To get the variance we have to derive $\mathbb{E}\left[\xi\right]$. Note that if ξ follows the inverse chi-square distribution, $E[\xi] = \nu/(\nu-2)$. We recognize form the Student's t distribution, where $\mathrm{Cov}(\mathbf{Y}) = \nu/(\nu-2)\boldsymbol{\Sigma}$.

B Proof of Proposition

To prove Proposition 1, we partition the data $\mathbf{y}$ as $[\mathbf{y}_1, \mathbf{y}_2]$, so n_1 data points belong to $\mathbf{y}_1$, n_2 data points belong to $\mathbf{y}_2$ and $n_1 + n_2 = n$.

The write joint distribution of $[\mathbf{y}_1, \mathbf{y}_2]$ as $p(\mathbf{y}_1, \mathbf{y}_2|\xi)p(\xi; \boldsymbol{\eta})$. The conditional distribution of $\mathbf{y}_2$, given $\mathbf{y}_1$ is then $p(\mathbf{y}_2|\mathbf{y}_1, \xi)p(\xi|\mathbf{y}_1 M; \boldsymbol{\eta})$.

For a given ξ, $p(\mathbf{y}_2|\mathbf{y}_1, \xi)$ is the conditional normal distribution and so

$$p(\mathbf{y}_2|\mathbf{y}_1, \xi) \sim \mathcal{N}(\boldsymbol{\mu}_{2|1}, \Sigma_{22|1}\hat{\xi}), \quad \hat{\xi} \sim p(\xi|\mathbf{y}_1; \boldsymbol{\eta}) \tag{28}$$

where,

$$\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1) \tag{29}$$

$$\Sigma_{22|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}, \tag{30}$$

the same as for the conditional Gaussian distribution. We obtain the conditional distribution $p(\xi|\mathbf{y}_1; \boldsymbol{\eta})$ by remembering that

$$p(\mathbf{y}_1|\xi) \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11}\xi). \tag{31}$$

Using Bayes' Theorem we get

$$\begin{aligned} p(\xi|\mathbf{y}_1; \boldsymbol{\eta}) &\propto p(\mathbf{y}_1|\xi)p(\xi; \boldsymbol{\eta}) \\ &\propto |\Sigma_{11}\xi|^{-1/2} \exp\left\{-\frac{u_1}{2\xi}\right\} p(\xi; \boldsymbol{\eta}) \\ &\propto \xi^{-N_1/2} \exp\left\{-\xi\frac{u_1}{2}\right\} p(\xi; \boldsymbol{\eta}). \end{aligned} \tag{32}$$

Recall that $u_1 = (\mathbf{y} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y} - \boldsymbol{\mu}_1))$. We normalize the distribution by

$$c_{N_1,\boldsymbol{\eta}}^{-1} = \int_0^\infty \xi^{-N_1/2} \exp\left\{-\frac{u_1}{2\xi}\right\} p(\xi; \boldsymbol{\eta}) d\xi \tag{33}$$

The conditional mixing distribution is

$$p(\xi|\mathbf{y}_1; \boldsymbol{\eta}) = c_{N_1,\boldsymbol{\eta}}\xi^{-N_1/2} \exp\left\{-\frac{u_1}{2\xi}\right\} p(\xi; \boldsymbol{\eta}) \tag{34}$$

The conditional distribution of $\mathbf{y}_2$ given $\mathbf{y}_1$ is derived by using the consistency formula

$$p(\mathbf{y}_2|\mathbf{y}_1) = \frac{1}{|\boldsymbol{\Sigma}_{22|1}|^{1/2}(2\pi)^{N_2/2}} \int_0^\infty \xi^{-N_2/2} \exp -\frac{u_{2|1}}{2\xi} p(\xi|\mathbf{y}_1) d\xi, \tag{35}$$

where $u_{2|1} = (\mathbf{y}_2 - \boldsymbol{\mu}_{2|1})^\top \Sigma_{22|1}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_{2|1})$. Using (34) we get

$$p(\mathbf{y}_2|\mathbf{y}_1) = \frac{c_{N_1,\boldsymbol{\eta}}}{|\boldsymbol{\Sigma}_{22|1}|^{1/2}(2\pi)^{N_2/2}} \int_0^\infty \xi^{-n/2} e^{-(u_{2|1}+u_1)/(2\xi)} p(\xi; \boldsymbol{\eta}) d\xi \tag{36}$$

C Training with variational inference

For a Gaussian process the posterior of the latent variables $\boldsymbol{f}$ is

$$p(\boldsymbol{f}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{f})p(\boldsymbol{f}). \tag{37}$$

Here, the prior $p(\boldsymbol{f}|\mathbf{x}) \sim \mathcal{N}(0, K)$, is a Gaussian process with kernel K and the likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{f}) \sim \mathcal{N}(\boldsymbol{f}, \sigma^2\mathbf{I})$ is Gaussian. The posterior derives to

$$p(\boldsymbol{f}|\mathbf{y}) \sim \mathcal{N}\left(\boldsymbol{f}|K\left(K + \sigma^2 I\right)^{-1}\mathbf{y}, \left(\mathbf{K}^{-1} + \sigma^{-2}\mathbf{I}\right)^{-1}\right) \tag{38}$$

and we can derive the predictive distribution of an arbitrary input location x^* by

$$p(f^*|\mathbf{y}) = \int p(f_*|\boldsymbol{f})p(\boldsymbol{f}|\mathbf{y})d\boldsymbol{f}, \tag{39}$$

where $p(f_*|\boldsymbol{f}, \mathbf{x}, \mathbf{x}_*)$ is the conditional distribution, which is again Gaussian with

$$\mathcal{N}\left(f_*|\mathbf{k}_*^\top(\mathbf{k}+\sigma^2\mathbf{I})^{-1}\mathbf{y}, k_{**}-\mathbf{k}_*^\top(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{k}_*\right). \tag{40}$$

We want to derive the predictive distribution for the elliptical process, but the problem is that the posterior is intractable. In order to get a tractable posterior, we train the model using variational inference, where we approximate the intractable posterior with a tractable one,

$$p(\boldsymbol{f}, \xi, \omega|\mathbf{y}; \boldsymbol{\eta}) \approx q(\boldsymbol{f}, \xi, \boldsymbol{\omega}; \boldsymbol{\varphi}) = q(\boldsymbol{f}|\xi; \boldsymbol{\varphi_f})q(\xi; \boldsymbol{\varphi}_\xi)q(\boldsymbol{\omega}; \boldsymbol{\varphi}_\omega). \tag{41}$$

Here, $q(\boldsymbol{f}|\xi; \boldsymbol{\varphi}) \sim \mathcal{N}(\mathbf{m}, \mathbf{S}\xi)$, where $\mathbf{m}$ and $\mathbf{S}$ are variational parameters, and $q(\xi; \boldsymbol{\varphi}_\xi)$ and $q(\omega; \boldsymbol{\varphi}_\omega)$ are parameterized with any positive distribution such as a normalizing flow. We use this approximation when we derive the predictive distribution

$$\begin{aligned}
p(f^*|\mathbf{y}) &= \int p(f_*|\boldsymbol{f}, \xi, \boldsymbol{\omega}; \boldsymbol{\eta})p(\boldsymbol{f}, \xi, \boldsymbol{\omega}|\mathbf{y}; \boldsymbol{\eta})d\boldsymbol{f}d\xi d\boldsymbol{\omega} && (42)\\
&= \int p(f_*|\boldsymbol{f}, \xi; \boldsymbol{\eta_f})p(\boldsymbol{f}, \xi, \boldsymbol{\omega}|\mathbf{y}; \boldsymbol{\eta})d\boldsymbol{f}d\xi d\boldsymbol{\omega} && (43)\\
&\approx \int p(f_*|\boldsymbol{f}, \xi; \boldsymbol{\eta_f})q(\boldsymbol{f}|\xi; \boldsymbol{\varphi_f})q(\xi; \boldsymbol{\varphi}_\xi)q(\boldsymbol{\omega}; \boldsymbol{\varphi}_\omega)d\boldsymbol{f}d\xi d\boldsymbol{\omega} && (44)\\
&= \int p(f_*|\boldsymbol{f}, \xi; \boldsymbol{\eta})q(\boldsymbol{f}|\xi; \boldsymbol{\varphi_f})q(\xi; \boldsymbol{\varphi}_\xi)d\boldsymbol{f}d\xi. && (45)
\end{aligned}$$

If we first take a look at the prior distribution $p(f^*, \boldsymbol{f}|\xi)$ which is, when ξ is deterministic, a $\mathcal{GP}$-prior,

$$\begin{bmatrix} f^* \\ \boldsymbol{f} \end{bmatrix} \xi \sim \mathcal{N}\left(0, \begin{bmatrix} k_{**} & \mathbf{k}_*^\top \\ \mathbf{k}_* & \mathbf{K} \end{bmatrix} \xi\right), \tag{46}$$

with the the conditional distribution

$$\begin{aligned}
p(f^*|\boldsymbol{f}, \xi; \boldsymbol{\eta}) &= \mathcal{N}\left(\mathbf{k}_*^\top\mathbf{K}^{-1}\boldsymbol{f}, \left(k_{**}-\mathbf{k}_*^\top\mathbf{K}^{-1}\mathbf{k}_*\right)\xi\right) && (47)\\
&= \mathcal{N}\left(\mathbf{a}^\top\boldsymbol{f}, b\xi\right). && (48)
\end{aligned}$$

Here, $\mathbf{a}^\top = \mathbf{k}_*^\top\mathbf{K}^{-1}$ and $b = k_{**} - \mathbf{k}_*^\top\mathbf{K}^{-1}\mathbf{k}_*$. We use this expression and the variational approximation when we derive the posterior predictive distribution,

$$\begin{aligned}
p(f^*|\mathbf{y}) &= \int p(f_*|\boldsymbol{f}, \xi; \boldsymbol{\eta})q(\boldsymbol{f}|\xi; \varphi_{\boldsymbol{f}})q(\xi; \varphi_\xi)d\boldsymbol{f}d\xi && (49)\\
&= \mathbb{E}_{q(\xi;\boldsymbol{\varphi}_\xi)}\left[\int p(f_*|\boldsymbol{f}, \xi)q(\boldsymbol{f}|\xi; \varphi_{\boldsymbol{f}})d\boldsymbol{f}\right] && (50)\\
&= \mathbb{E}_{q(\xi;\boldsymbol{\varphi}_\xi)}\left[\int \mathcal{N}\left(f_*|\mathbf{a}^\top\boldsymbol{f}, b\xi\right)\mathcal{N}\left(\boldsymbol{f}|\mathbf{m}, \mathbf{S}\xi\right)d\boldsymbol{f}\right] && (51)\\
&= \mathbb{E}_{q(\xi;\boldsymbol{\varphi}_\xi)}\left[\int \mathcal{N}\left(f_*|\mathbf{a}^\top\mathbf{m}, \mathbf{a}^\top\mathbf{S}\mathbf{a}\xi + b\xi\right)\right] && (52)\\
&= \mathbb{E}_{q(\xi;\boldsymbol{\varphi}_\xi)}\left[\mathcal{N}(f_*|m_*, s_*\xi)\right] && (53)
\end{aligned}$$

where

$$\begin{aligned}
m_* &= \mathbf{a}^\top\mathbf{m} && (54)\\
s_* &= \mathbf{a}^\top\mathbf{S}\mathbf{a} + b && (55)
\end{aligned}$$

and we get the covariance by $\mathbb{E}[\xi]s_*$.

Optimizing the ELBO

We train the model by optimizing the evidence lower bound (ELBO) given by

$$\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\eta}) = \mathbb{E}_{q(\boldsymbol{f}|\xi;\boldsymbol{\varphi_f})q(\xi;\boldsymbol{\varphi}_\xi)q(\omega;\boldsymbol{\varphi}_\omega)}\left[\log p(\mathbf{y}, \boldsymbol{f}, \xi, \omega; \boldsymbol{\eta}) - \log\left(q(\boldsymbol{f}|\xi; \boldsymbol{\varphi_f})q(\xi; \boldsymbol{\varphi}_\xi)q(\omega; \boldsymbol{\varphi}_\omega)\right)\right]. \tag{56}$$

The model is implemented in Pyro (Bingham et al., 2018), see Section E for details.

D Sparse elliptical processes

With the variational inference framework we can create a sparse version of the model

$$\int p(\boldsymbol{f}, \boldsymbol{u}, \xi; \boldsymbol{\eta}) d\xi = \int p(\boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta_f}) p(\boldsymbol{u}|\xi; \boldsymbol{\eta_u}) p(\xi; \boldsymbol{\eta}_\xi) d\xi, \tag{57}$$

where $\boldsymbol{u}$ are outputs of the elliptical process, located at the inducing inputs $\mathbf{x}_u$. We approximate the posterior with

$$p(\boldsymbol{f}, \boldsymbol{u}, \xi|\mathbf{y}; \boldsymbol{\eta}) \approx p(\boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta_f}) q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) q(\xi; \boldsymbol{\varphi}_\xi) \tag{58}$$

The posterior of the distribution is given by

$$\begin{aligned} p(f^*|\mathbf{y}) &= \int p(f_*|\boldsymbol{f}, \boldsymbol{u}, \xi; \boldsymbol{\eta}) p(\boldsymbol{f}, \boldsymbol{u}, \xi|\mathbf{y}; \boldsymbol{\eta}) d\boldsymbol{f} d\boldsymbol{u} d\xi \\ &\approx \int p(f_*|\boldsymbol{f}, \boldsymbol{u}, \xi; \boldsymbol{\eta}) p(\boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta_f}) q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) q(\xi; \boldsymbol{\varphi}_\xi) d\boldsymbol{f} d\boldsymbol{u} d\xi \\ &= \int \left[\int p(f_*|\boldsymbol{f}, \boldsymbol{u}, \xi; \boldsymbol{\eta}) p(\boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta_f}) d\boldsymbol{f} \right] q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) q(\xi; \boldsymbol{\varphi}_\xi) d\boldsymbol{u} d\xi \end{aligned} \tag{59}$$

We can simplify the inner expression by using the fact that the elliptical distribution is consistent,

$$\int p(f_*|\boldsymbol{f}, \boldsymbol{u}, \xi; \boldsymbol{\eta}) p(\boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta}) d\boldsymbol{f} = \int p(f_*, \boldsymbol{f}|\boldsymbol{u}, \xi; \boldsymbol{\eta}) d\boldsymbol{f} = p(f_*|\boldsymbol{u}, \xi; \boldsymbol{\eta}). \tag{60}$$

Hence, Equation (59) is simplifies to

$$p(f^*|\mathbf{y}) = \int p(f_*|\boldsymbol{u}, \xi; \boldsymbol{\eta}) q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) q(\xi; \boldsymbol{\varphi}_\xi) d\boldsymbol{u} d\xi, \tag{61}$$

where $q(\boldsymbol{u}|\xi; \boldsymbol{\varphi_u}) = \mathcal{N}(\mathbf{m}_u, \mathbf{S}_u \xi)$ with the variational parameters $\mathbf{m}_u$ and $\mathbf{S}_u$, and ξ is parameterized, e.g., by a normalizing flow

Finally, we obtain the predictive posterior

$$m_* = \mathbf{a}_u^T \mathbf{m}_u \tag{62}$$

$$s_* = \mathbf{a}_u^T \mathbf{S}_u \mathbf{a}_u + b_u \tag{63}$$

where $\mathbf{a}_u^\top = \mathbf{k}_{*u}^\top \mathbf{K}_{uu}^{-1}$ and $b_u = k_{**} - \mathbf{k}_{*u}^\top \mathbf{K}_{uu}^{-1} \mathbf{k}_{*u}$.

E Implementation: variational inference

We used the Pyro library (Bingham et al., 2018), which is a universal probabilistic programming language (PPL) written in Python and supported by PyTorch on the backend.

In Pyro, we trained a model with variational inference (Kingma & Welling, 2013) by creating "stochastic functions" called **model** and a **guide**, where the **model** samples from the prior latent distributions $p(\boldsymbol{f}, \xi, \omega; \boldsymbol{\eta})$, and the observed distribution $p(\mathbf{y}|\boldsymbol{f}, \omega)$, and the **guide** samples the approximate posterior $q(\boldsymbol{f}|\xi; \boldsymbol{\varphi_f}) q(\xi; \boldsymbol{\varphi}_\xi) q(\omega; \varphi_\omega)$. We then trained the model by minimizing the ELBO, where we simultaneously optimized the model parameters $\boldsymbol{\eta}$ and the variational parameters $\boldsymbol{\varphi}$. (See more details here, `https://pyro.ai/examples/svi_part_i.html`.)

To implemented the model in Pyro, we created the guide and the model (see Algorithm 3), which we did by building upon the already implemented variational Gaussian process. We used the guide and the model to derive the evidence lower bound (ELBO), which we then optimized with stochastic gradient descent using the Adam optimizer (Kingma & Ba, 2015).

We used the already implemented rational linear spline flow for the normalizing flow in Pyro.

Algorithm 1 PyTorch implementation of the variational elliptical process.

procedure MODEL($\mathbf{X}, \mathbf{y}$)
  $\mathbf{K} = kernel(\mathbf{X}) + \mathbf{I}\sigma^2$
  $\mathbf{L} = \mathbf{K}.cholesky()$
  sample(name $= f$, dist $= \mathcal{N}(\mathbf{0}, \mathbf{K})$) ▷ Take a sample from the latent $\boldsymbol{f}, \xi$ and ω
  $\xi =$ sample(name $= \xi$, dist $=$ Normalizing flow)
  $\omega =$ sample(name $= \omega$, dist $=$ NF)
  Get variational parameters $\mathbf{m}, \mathbf{S}$
  Derive $p(\mathbf{y}|\mathbf{m}, \mathbf{S}\xi + \mathbf{I}\sigma\omega)$ ▷ Sample $\mathbf{y}$ with the obs statement in Pyro.
end procedure
procedure GUIDE
  sample(name $= f$, dist $= \mathcal{N}(\mathbf{m}, \mathbf{S})$) ▷ Take a sample from the variational latent $\boldsymbol{f}, \xi$ and ω
  sample(name $= \xi$, dist $=$ Variational NF)
  sample(name $= \omega$, dist $=$ Variational NF)
end procedure

F Derivation of the confidence regions of the elliptical process

We derive the confidence region of the elliptical process, by using Monte Carlo approximation of the integral, as

$$p(-z\sigma < x < z\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-z\sigma}^{z\sigma} \int_0^{\infty} \xi^{-1/2} e^{-x^2/(\xi 2\sigma^2)} p(\xi) d\xi dx \tag{64}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-z\sigma}^{z\sigma} \frac{1}{m} \sum_{i=1}^{m} \xi_i^{-1/2} e^{-x^2/(2\xi_i\sigma^2)} dx \tag{65}$$

$$= \frac{1}{\sigma m\sqrt{2\pi}} \sum_{i=1}^{m} \xi_i^{-1/2} \int_{-z\sigma}^{z\sigma} e^{-x^2/(2\xi_i\sigma^2)} dx \tag{66}$$

$$= \frac{2}{m\sqrt{\pi}} \sum_{i=1}^{m} \int_0^{\frac{z}{\sqrt{2\xi_i}}} e^{-u^2} du \tag{67}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \text{erf}\left(\frac{z}{\sqrt{2\xi_i}}\right) \tag{68}$$

For every mixing distribution we can derive the confidence of the prediction. It is the number of samples m we take that decides the accuracy of the confidence.

G Datasets

California housing dataset was originally published by Pace & Barry (1997). There are 20 640 samples and 9 feature variables in this dataset. The targets are prices on houses in the California area.

The Concrete dataset (Yeh, 1998) has 8 input variables and 1030 observations. The target variables are the concrete compressive strength.

Machine CPU dataset (Kibler et al., 1989) where the target value is the relative performance of the CPU. The dataset consist of 209 samples with nine attributes.

Auto MPG dataset (Alcalá-Fdez et al., 2011) originally from the StatLib library which is maintained at Carnegie Mellon University. The data concerns city-cycle fuel consumption in miles per gallon and consists of 392 samples with five features each.

Pima Indians Diabetes Database (Smith et al., 1988) originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a

patient has diabetes, based on certain diagnostic measurements included in the dataset. The dataset consist of 768 samples with eight attributes.

The Cleveland Heart Disease dataset consists of 13 input variables and 270 samples. The target classifies whether a person is suffering from heart disease or not.

The Mammographic Mass dataset predicts the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. This dataset consists of 961 with six attributes.