

Towards Robust Prediction Using The Elliptical Process for Regression

A.B. Al Khabori^{*}, M.T. Alodat and Amadou Sarr

Department of Statistics, Sultan Qaboos University,

Al-Khod Muscat, Sultanate of Oman

E-mails: s33859@student.squ.edu.om; m.alodat@squ.edu.om; asarr@squ.edu.om

^{*}Corresponding author: s33859@student.squ.edu.om

Abstract

We present a novel Bayesian family of models, Elliptical Processes (EPs), designed to extend regression modeling framework. Unlike the widely used Gaussian Process (GP) Regression, EPs generalize the GP framework by accommodating non-normal tail distributions and outlier-prone data, where Gaussian assumptions often fail. In this paradigm, the GP is still a special case, but EPs are more accurate, especially when dealing with heavy-tailed data. We use the Laplace approximation technique to ensure scalability and computational efficiency while addressing the analytical complexity inherent in EP inference. We derive the predictive distribution for EPs at new input points and provide a comprehensive performance comparison with GP, T-Process, and other EP models. The proposed EP, particularly under non-Gaussian assumptions, outperform the GP by demonstrating superior performance and flexibility in handling complex data structures across both simulated and real-world scenarios.

Keywords: Bayesian Inference; Elliptical Process; Gaussian Process; Laplace approximation; Predictive distribution; Regression.

MSC: Primary 62J02; Secondary 60G15.

1 Introduction

With the progress of technology and the increasing sophistication of computers, there is a growing potential to utilize models, in Bayesian framework for handling intricate and non Gaussian data sets more effectively. The evolution of Gaussian Processes (GPs) and its applications in Bayesian modeling has seen a number of significant turning points. In 1978, O’Hagan pioneered the use of GP priors as a non-parametric treatment for nonlinear regression within a Bayesian framework. Neal (1995) built upon O’Hagan’s findings and incorporated them into neural networks, demonstrating their utility in probabilistic modeling. The effectiveness of Gaussian Process Regression (GPR) in addressing a range of regression and classification problems was demonstrated by Rasmussen and Williams in their 1996 machine learning description.

It has become evident that not all datasets adhere to distribution assumptions and many real world data collections display variations, like tails and skewness instead of following a normal distribution as traditionally assumed (Neal, 1998). Neal (1998) shares a viewpoint emphasizing the importance of investigating modeling strategies that can handle a wide range of non standard and diverse data structures effectively. In the light of his viewpoint, for the innovation of the family of Elliptical Processes for regression (EPR) comes into play as an option with its capability to provide a Bayesian framework, for addressing the inherent intricacies found in diverse real life datasets.

While GPs are powerful tools for non-parametric Bayesian modeling, their limitations have become increasingly apparent. Neal (1998) identified discontinuity issues at unknown locations in functions modeled by GPs, particularly when applied to complex real-world problems. In addition to discontinuities, GPs face practical limitations, such as high computational complexity, which scales cubically with the number of data points, making them impractical for large datasets (Rasmussen and Williams, 2006). GPs also require careful selection of kernel functions, and poor choices can lead to suboptimal performance, especially when data are non-stationary or exhibit non-Gaussian characteristics (Rasmussen and Williams, 2006). Furthermore, the assumption of Gaussian noise restricts their robustness to outliers or heavy-tailed data, and their inability to model sharp discontinuities limits their applicability in many real-world scenarios (Neal, 1998). These challenges have motivated the development of alternative models that relax some of the restrictive assumptions of GPs.

Several models have been put forth in response to these constraints. Benavoli et al. (2020) introduced the Skew GP through the Unified Skew-Normal (SUN) distribution to effectively model asymmetric data, while Kuss et al. (2006) created resilient models that enhanced performance when noise and outliers were present. Alodat and Aludaat (2008) introduced the Generalized Hyperbolic Process, while Vanhatalo et al. (2009) created a GPR model with a Student-t likelihood that is resilient against outliers. Alodat and Al-Momani (2014) introduced a model referred to as the Skew Gaussian Process for regression to handle skewed data structures, and Alodat and Al-Shakhatreh (2020) suggested a GPR model with skewed errors, which provides more flexibility for handling asymmetric and non-Gaussian data, thereby broadening the toolkit for overcoming the drawbacks of conventional GPs.

Other models that researchers have developed include the Extended Skew GP for Regression (Alodat and Rawwash, 2014), which offers more flexibility in modeling skewed data, and the Logistic Gaussian Process using Laplace Approximation (Riihimäki et al., 2014), which improves Bayesian model optimization. More recent developments provide strong solutions for managing data with heavy tails, such as Tang et al. (2017)'s Student-t Process Regression with Student-t Likelihood. The variety of processes accessible for sophisticated modeling was further increased by Tracey and Wolpert's (2018) switch from GPs to Student-T Processes. The EPs, a new family of flexible stochastic processes, were most recently described by Bânkestad et al. (2020). They provided an additional extension by using a squeeze box process (SBP) as an elliptical process to accommodate a variety of data formats in Bayesian modelling.

In this study, we develop an enhanced approach for using EPs in regression models within a Bayesian framework. We present an improved version of the EP model as a scale mixture of GP and evaluate its performance through applications to both simulated and real-world datasets. Our findings illustrate the model's effectiveness in handling non-Gaussian noise, heavy tails, and data irregularities, showcasing its robustness in Bayesian regression tasks.

The rest of this paper is organized as follows: Section 2 review the mathematical background necessary for understanding GPs and background of kernel functions. In Section 3, we introduce the EP model as a scale mixture of GP, including its stochastic representation and key properties. Section 4 present the methodology for applying the EP

to regression tasks as a novel Bayesian family of models. Section 5 outlines the Inference with Laplace approximation and Learning Hyper-parameters. Section 6 describes the experimental setup, datasets and the results of applying EPs to real-world datasets, illustrating their performance relative to alternative models. Finally, Section 7 provides concluding remarks and discusses potential avenues for future research.

2 The Gaussian Process for Regression

To facilitate a deeper understanding of the EP model for regression, a concise overview of GPR is essential, as detailed below.

For $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, the general regression model is given by:

$$y = f(\mathbf{x}) + \epsilon, \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ represents a normally distributed noise term with zero mean and variance σ^2 , assumed to be independent across observations. Here, $f(\cdot)$ is the unknown function to be estimated.

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the set of observed input-output data pairs, where $\mathfrak{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is the set of input data points, and $\mathbf{y} = \{y_1, \dots, y_n\}$ is the set of corresponding outputs. For the observed data \mathcal{D} , the observed version of (1) becomes:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \dots, n, \quad (2)$$

where the noise terms ϵ_i 's are independently distributed as $\mathcal{N}(0, \sigma^2)$.

In GPR model, the function $f(\cdot)$ is modeled as a GP prior over the space of all possible functions f . This GP prior is fully specified by a mean function $\mu(\cdot)$ and a covariance (or kernel) function $\mathcal{C}_\theta(\cdot, \cdot)$, where θ represents the hyperparameters of the kernel. We adopt the notation for GPs, i.e., $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\cdot), \mathcal{C}_\theta(\cdot, \cdot))$. The main goal is to predict f at new input points $\mathfrak{X}^* = \{\mathbf{x}_1^*, \dots, \mathbf{x}_m^*\}$, with the predictive distribution remaining Gaussian.

For convenience, we use the following notation:

$$\begin{aligned} \mathbf{f} = f(\mathfrak{X}) &= [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^T, & \mathbf{f}^* = f(\mathfrak{X}^*) &= [f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_m^*)]^T, \\ \mu(\mathfrak{X}) &= [\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n)]^T, & \mu(\mathfrak{X}^*) &= [\mu(\mathbf{x}_1^*), \dots, \mu(\mathbf{x}_m^*)]^T. \end{aligned}$$

The joint distribution of the function values at both observed and new input points, under the GP model, is:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}_{n+m} \left(\begin{bmatrix} \mu(\mathfrak{X}) \\ \mu(\mathfrak{X}^*) \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

where $\Sigma_{11} = (\mathcal{C}_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$, $\Sigma_{12} = (\mathcal{C}_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j^*))_{i=1,j=1}^{n,m}$, $\Sigma_{21} = \Sigma_{12}^T$, and $\Sigma_{22} = (\mathcal{C}_{\boldsymbol{\theta}}(\mathbf{x}_i^*, \mathbf{x}_j^*))_{i,j=1}^m$.

In the presence of noise, we seek to predict \mathbf{f}^* given the observations \mathbf{y} . Learning the hyperparameters $\boldsymbol{\theta}$ involves maximizing the likelihood of \mathbf{y} under the model, given by $\mathbf{y}|\mathfrak{X} \sim \mathcal{N}_n(\mu(\mathfrak{X}), \Sigma_{11} + \sigma^2 \mathbf{I}_n)$. The conditional distribution of \mathbf{f}^* given \mathbf{y}, \mathfrak{X} and \mathfrak{X}^* is then:

$$\mathbf{f}^*|\mathbf{y}, \mathfrak{X}, \mathfrak{X}^* \sim \mathcal{N}_m \left(\mu(\mathfrak{X}^*) + \Sigma_{21}(\Sigma_{11} + \sigma^2 \mathbf{I}_n)^{-1}(\mathbf{y} - \mu(\mathfrak{X})), \Sigma_{22} - \Sigma_{21}(\Sigma_{11} + \sigma^2 \mathbf{I}_n)^{-1}\Sigma_{12} \right).$$

The predictor of \mathbf{f}^* is:

$$\widehat{\mathbf{f}}^* = E(\mathbf{f}^*|\mathbf{y}, \mathfrak{X}, \mathfrak{X}^*) = \mu(\mathfrak{X}^*) + \Sigma_{21}(\Sigma_{11} + \sigma^2 \mathbf{I}_n)^{-1}(\mathbf{y} - \mu(\mathfrak{X})),$$

with uncertainty given by:

$$\text{cov}(\mathbf{f}^*|\mathbf{y}, \mathfrak{X}, \mathfrak{X}^*) = \Sigma_{22} - \Sigma_{21}(\Sigma_{11} + \sigma^2 \mathbf{I}_n)^{-1}\Sigma_{12}. \text{ See (Rasmussen, 2006) for more details.}$$

For new observations, we predict \mathbf{y}^* at \mathbf{x}^* :

$$\mathbf{y}^*|\mathbf{y}, \mathfrak{X}, \mathfrak{X}^* \sim \mathcal{N}_m(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),$$

where

$$\boldsymbol{\mu}^* = \mu(\mathfrak{X}^*) + (\Sigma_{21})(\Sigma_{11} + \sigma^2 \mathbf{I}_n)^{-1}(\mathbf{y} - \mu(\mathfrak{X})),$$

$$\boldsymbol{\Sigma}^* = (\Sigma_{22} + \sigma^2 \mathbf{I}_m) - (\Sigma_{21})(\Sigma_{11} + \sigma^2 \mathbf{I}_n)^{-1}(\Sigma_{12}).$$

This formulation allows for exact predictions with uncertainty estimates for a new observation by leveraging the properties of Gaussian distributions.

2.1 Kernel functions

GPs are frequently used in Bayesian non-parametric regression techniques as priors, as a GP is identified by its mean function ($\mu(\cdot)$) and covariance (kernel) function ($\mathcal{C}_{\boldsymbol{\theta}}(\cdot, \cdot; \boldsymbol{\theta})$), where $\boldsymbol{\theta}$ is the vector of hyperparameters related to the kernel function. Non-parametric regression relies heavily on kernel functions since they define covariance structures and have a big impact on model behavior. Because it directly affects the model's ability to

capture relationships within the dataset, selecting the right kernel is essential. To describe different data patterns, GPs use kernels as similarity metrics that regulate function complexity and smoothness. To establish a viable GP framework, the kernel selection procedure should be guided by prior information and must meet specific requirements, such as positive semi-definiteness and symmetry.

The Matérn kernel, the squared exponential (or radial basis function RBF) kernel, and the linear kernel are a few examples of frequently used kernels. These kernels offer versatility in capturing a variety of data structures and can be used singly or in combination. For instance, the smoothness characteristics and efficacy of the squared exponential kernel in simulating correlations over different distances are well established (Bishop, 2006). It is defined as:

$$\mathcal{C}_{SE}(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \psi^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right), \quad (3)$$

where l is the length scale parameter and ψ controls the variance.

Constructing new kernels from existing ones is necessary to improve model flexibility and capturing complex patterns in data. Here are common methods to construct new kernels from old ones:

Table 1: Techniques for Constructing New Kernels

Technique	New Kernel Construction
Positive scalar (a) multiplication	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = a \cdot \mathcal{C}_1(\mathbf{x}, \mathbf{x}')$
Adding a positive constant (b)	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = \mathcal{C}_1(\mathbf{x}, \mathbf{x}') + b$
Linear combination of positive weights (c_i)	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^m c_i \cdot \mathcal{C}_i(\mathbf{x}, \mathbf{x}')$
Polynomial function (q) with positive coefficients	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = q(\mathcal{C}_1(\mathbf{x}, \mathbf{x}'))$
Exponential function	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = \exp(\mathcal{C}_1(\mathbf{x}, \mathbf{x}'))$
Direct sum	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = \mathcal{C}_1(\mathbf{x}_1, \mathbf{x}'_1) + \mathcal{C}_2(\mathbf{x}_2, \mathbf{x}'_2)$
Tensor product	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = \mathcal{C}_1(\mathbf{x}_1, \mathbf{x}'_1) \cdot \mathcal{C}_2(\mathbf{x}_2, \mathbf{x}'_2)$
Vertical rescaling with scaling function (d)	$\mathcal{C}(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}) \cdot \mathcal{C}_1(\mathbf{x}, \mathbf{x}') \cdot d(\mathbf{x}')$

where \mathcal{C}_1 , \mathcal{C}_2 and \mathcal{C}_i are base kernels (Bishop, 2006).

To fit a GP model for regression, the kernel’s hyperparameters are optimized using the maximum a posteriori (MAP) method, which combines prior knowledge with observed

data to improve prediction accuracy. Once the model is optimized, we can make predictions on new data while providing an estimate of the prediction’s uncertainty. Further details on hyperparameters optimization and computing the posterior will be discussed in the next sections. Additional insights can be found in (Rasmussen, 2006) and (Bishop, 2006).

3 The Elliptical Process

Elliptical distributions form a family of probability distributions that include multivariate normal distributions as a special case. Understanding the key properties of elliptical distributions is crucial for applications in statistical modeling, including regression modeling and machine learning.

In Bayesian non-parametric regression, the unknown output function $f(\cdot)$ is typically modeled using a GP as a prior in non-parametric methods. Here, we propose a robust and powerful extension called the Elliptical Process for Regression model (EPR). With its broader set of finite-dimensional distributions, the EPR emerges as a compelling competitor to the GPR in certain aspects. Specifically, we assume $f(\mathbf{x})$ in the previous regression model (2) follows an EP $v(\mathbf{x})$, which is a generalization of the GP, created by multiplying the GP by the square root of a positive random variable. This expands the horizons of regression modeling:

$$y_i = v(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \dots, n$$

The noise terms ϵ_i are considered as Gaussian white noise, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Additionally, the EP is defined as a scale mixture of GP:

$$v(\mathbf{x}) = \sqrt{R}f(\mathbf{x}),$$

where R is a positive random variable with a probability density function (PDF) $p_R(r)$, independent of the GP $f(\mathbf{x})$.

One of the defining characteristics of the EP is that any finite set of observations follows an elliptical distribution \mathcal{E} . For any set of input points $\mathfrak{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ and corresponding function values $\mathbf{v} = v(\mathfrak{X}) = [v(\mathbf{x}_1), \dots, v(\mathbf{x}_n)]^T$, the joint distribution of these function values is a multivariate elliptical distribution:

$$\mathbf{v} \mid \mathfrak{X} \sim \mathcal{E}_n(\mu(\mathfrak{X}), \Sigma_{11}, p_R),$$

where

$$\boldsymbol{\mu}(\mathcal{X}) = [\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n)]^T, \quad \boldsymbol{\Sigma}_{11} = (\mathcal{C}_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n.$$

3.1 Stochastic Representation

The elliptical distribution is a linear transformation of a spherically distributed random vector centered at the origin. Specifically, an n -dimensional random vector \mathbf{v} is said to be elliptically distributed if and only if:

$$\mathbf{v} \stackrel{d}{=} \boldsymbol{\mu} + R\mathbf{A}\mathbf{U}, \quad (4)$$

where $\boldsymbol{\mu}$ is an n -vector, $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$, R is a positive random variable, and \mathbf{U} is uniformly distributed over the unit sphere. Using the stochastic representation in Equation (4), we can write:

$$\mathbf{v} \sim \mathcal{E}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, p_R).$$

If it exists, the probability density function (PDF) of the elliptical distribution is:

$$p_{\mathbf{V}}(\mathbf{v}) = C_n |\boldsymbol{\Sigma}|^{-\frac{1}{2}} g\left((\mathbf{v} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{v} - \boldsymbol{\mu})\right), \quad (5)$$

where C_n is a normalizing constant, $\boldsymbol{\Sigma}$ is the covariance matrix, $\boldsymbol{\mu}$ is the mean vector, and $g(\cdot)$ is the density generator.

The choice of the density generator $g(\cdot)$ plays a pivotal role in determining the specific characteristics of the elliptical distribution. The following are a few noteworthy selections that will be utilized later:

1. Normal Distribution: When $g(u) = \exp(-u/2)$, the elliptical distribution becomes the multivariate normal distribution.
2. Student-t Distribution: When $g(u) = \left(1 + \frac{u}{m}\right)^{-(n+m)/2}$, where m is the degrees of freedom, the elliptical distribution becomes a multivariate Student-t distribution, well-suited for modeling data with heavy tails.
3. Cauchy Distribution: If $g(u) = (1 + u)^{-(n+1)/2}$, we obtain the Cauchy distribution, known for its heavy tails and robustness to outliers.
4. Laplace (L) Distribution: If $g(u) = \exp(-\sqrt{u})$, this leads to the Laplace distribution, which is suitable for data with sharp peaks and heavier tails than the Gaussian.

5. Power Exponential (PE) Distribution: If $g(u) = \exp(-u^{s/2})$, where $s > 1$, this defines the Power Exponential distribution, providing a flexible way to model data that interpolates between normal ($s = 2$) and Laplace ($s = 1$) distributions.

In summary, modeling a variety of data patterns and behaviors within the framework of elliptical distributions is made possible by the freedom in selecting the density generator $g(\cdot)$. For more details, see (Pu et al., 2024).

3.2 Some Properties of the Elliptical Process

As the finite-dimensional distributions of the EP are all elliptical, the EP inherits several key properties from elliptical distributions. These include characterization by location and scale parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, closure under linear transformations, symmetry around the mean, invariance under affine transformations, and stability under marginalization. Additionally, a multivariate normal distribution can be used to depict elliptical distributions, providing a wide range of tail behaviors, from lighter to heavier than those of the normal distribution. Examples of this adaptability include the multivariate normal, multivariate t , multivariate Cauchy and other distributions. Because of these characteristics, the EP can be used to model data with different levels of symmetry and tail behavior in a versatile and effective manner.

Furthermore, elliptical distributions exhibit attractive conditional properties, such as the ability to define conditional distributions, which provide insights into the distribution given observed values. Moreover, elliptical distributions' conditional mean and covariance can be represented in closed form. Elliptical distributions are useful for capturing intricate relationships in data because of these conditional characteristics (Fang et al., 2018).

Let $\mathbf{V} \in \mathbb{R}^d$, where $\mathbf{V} \sim \mathcal{E}_d(\boldsymbol{\gamma}, \boldsymbol{\Omega}, h)$, and partition \mathbf{V} as $\mathbf{V}^T = (\mathbf{V}_1^T, \mathbf{V}_2^T)$, with the covariance matrix and location vector partitioned as:

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}.$$

The following conditional properties hold:

- (1) $\mathbf{V}_1 \sim \mathcal{E}_{d_1}(\boldsymbol{\gamma}_1, \boldsymbol{\Omega}_{11}, h_1)$.

(2) $\mathbf{V}_2|\mathbf{V}_1 = \mathbf{v}_1 \sim \mathcal{E}_{d_2}(\gamma_{2|1}, \boldsymbol{\Omega}_{22|1}, h_{22|1})$, where:

$$\gamma_{2|1} = \gamma_2 + \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(\mathbf{v}_1 - \gamma_1),$$

$$\boldsymbol{\Omega}_{22|1} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12},$$

$$h_{22|1}(t) = h(t + q_1),$$

where

$$q_1 = (\mathbf{v}_1 - \gamma_1)^\top \boldsymbol{\Omega}_{11}^{-1}(\mathbf{v}_1 - \gamma_1).$$

For further details, see (Gómez et al., 2003).

Understanding these properties is crucial for advanced statistical modeling, particularly in regression analysis and predictive modeling, where flexibility in accommodating various data structures is essential.

4 Elliptical Process for Regression

In this section, we will apply the EP defined as a scale mixture of GP in section 3 to conduct a non-parametric regression model defined in equation (2). Since $v \sim \mathcal{E}(\mu(\cdot), \mathcal{C}_\theta(\cdot, \cdot), p_R)$, $v(\mathbf{x}) \sim \mathcal{E}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}, p_R)$ and $\mathbf{v} = [v(\mathbf{x}_1), \dots, v(\mathbf{x}_n)]^T = \sqrt{R}\mathbf{f}$, then the conditional distribution of \mathbf{v} given $R = r$:

$$\mathbf{v}|R = r \sim \mathcal{N}_n(r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}), r\boldsymbol{\Sigma}_{11}).$$

The overarching goal remains to predict the output function at new inputs i.e. $\mathbf{v}^* = v(\boldsymbol{\mathfrak{X}}^*) = [v(\mathbf{x}_1^*), \dots, v(\mathbf{x}_m^*)]^T$. Beginning with the Elliptical Process for Regression (EPR), where $v(\mathbf{x})$ follows an EP model with free noise i.e. $\epsilon_i = 0$, the joint distribution of \mathbf{v} and \mathbf{v}^* is:

$$\begin{bmatrix} \mathbf{v} \\ \mathbf{v}^* \end{bmatrix} = \sqrt{R} \begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{E}_{n+m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, p_R).$$

Now, the conditional distribution of \mathbf{v} and \mathbf{v}^* , given $R = r$, is:

$$\begin{bmatrix} \mathbf{v} \\ \mathbf{v}^* \end{bmatrix} \Big|_{R=r} \sim \mathcal{N}_{n+m} \left(r \begin{bmatrix} \boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}) \\ \boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}^*) \end{bmatrix}, r \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

The predictive distribution of $\mathbf{v}^*|\mathbf{v}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R = r$ is obtained by applying the conditional distribution of a multivariate normal distribution as:

$$\mathbf{v}^*|\mathbf{v}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R = r \sim \mathcal{N}_m(r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}^*) + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{v} - r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}})), r\boldsymbol{\Sigma}_{22} - r\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}).$$

Typically, in EPR, as in GPR, $\mu(\mathfrak{X}) = 0$ is often assumed, so here we shall assume that $\mu(\mathfrak{X}) = 0$ for all choices of \mathfrak{X} .

In order to obtain the predictive distribution of $\mathbf{v}^*|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*$, we marginalize over the PDF of $R|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*$. Using the Bays' Theorem, the posterior distribution of R is given by:

$$p_R(r|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*) = \frac{p(\mathbf{v}|\mathfrak{X}, \mathfrak{X}^*, R=r)p_R(r)}{\int_0^\infty p(\mathbf{v}|\mathfrak{X}, \mathfrak{X}^*, R=r)p_R(r) dr}.$$

The predictive distribution of $\mathbf{v}^*|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*$ is:

$$p(\mathbf{v}^*|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*) = \int_0^\infty p(\mathbf{v}^*|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*, R=r) p_R(r|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*) dr,$$

which can be obtained through simulation.

So, we use the following MCMC methods to approximate $p(\mathbf{v}^*|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*)$ as:

$$p(\mathbf{v}^*|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*) \approx \frac{1}{N} \sum_{i=1}^N p(\mathbf{v}^*|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*, R^{(i)}),$$

where $R^{(1)}, R^{(2)}, \dots, R^{(N)}$ is a large sample from $p_R(r|\mathbf{v}, \mathfrak{X}, \mathfrak{X}^*)$.

It is typical for more realistic modeling situations that we do not have access to function values themselves, but only noisy versions, i.e., $y_i = v(\mathbf{x}_i) + \epsilon_i$. Assuming additive independent identically distributed Gaussian noise ϵ_i with variance σ^2 .

Then using the conditional formula of the multivariate normal distribution, we get:

$$\begin{bmatrix} \mathbf{v} \\ \boldsymbol{\epsilon} \\ \mathbf{v}^* \end{bmatrix} \Big|_{R=r} \sim \mathcal{N}_{2n+m} \left(r \begin{bmatrix} \mu(\mathfrak{X}) \\ \mathbf{0} \\ \mu(\mathfrak{X}^*) \end{bmatrix}, \begin{bmatrix} r\boldsymbol{\Sigma}_{11} & \mathbf{0} & r\boldsymbol{\Sigma}_{12} \\ \mathbf{0} & \sigma^2\mathbf{I}_n & \mathbf{0} \\ r\boldsymbol{\Sigma}_{21} & \mathbf{0} & r\boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Introducing the noise term, the joint distribution of the observed target values and the

function values at the test locations under the prior can be written as $\begin{bmatrix} \mathbf{y} \\ \mathbf{v}^* \end{bmatrix} = A \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\epsilon} \\ \mathbf{v}^* \end{bmatrix}$,

so the joint distribution given $R=r$ becomes:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{v}^* \end{bmatrix} \Big|_{R=r} \sim \mathcal{N}_{n+m} \left(\begin{bmatrix} r\mu(\mathfrak{X}) \\ r\mu(\mathfrak{X}^*) \end{bmatrix}, \begin{bmatrix} r\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n & r\boldsymbol{\Sigma}_{12} \\ r\boldsymbol{\Sigma}_{21} & r\boldsymbol{\Sigma}_{22} \end{bmatrix} \right),$$

where $A = \begin{bmatrix} \mathbf{I}_n & \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_m \end{bmatrix}$. By deriving the conditional distribution of \mathbf{v}^* given $\mathbf{y}, \mathfrak{X}, \mathfrak{X}^*$

and $R=r$, we arrive at the key predictive equations for EPR:

$$\mathbf{v}^*|\mathbf{y}, \mathfrak{X}, \mathfrak{X}^*, R=r \sim \mathcal{N}_m(\boldsymbol{\mu}^{**}, \boldsymbol{\Sigma}^{**}),$$

where $\boldsymbol{\mu}^{**} = r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}^*) + r\boldsymbol{\Sigma}_{21}(r\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{y} - r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}))$, and $\boldsymbol{\Sigma}^{**} = (r\boldsymbol{\Sigma}_{22} - r\boldsymbol{\Sigma}_{21}(r\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n)^{-1}r\boldsymbol{\Sigma}_{12})$.

Then the posterior predictive distribution of \mathbf{v}^* is:

$$p(\mathbf{v}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) = \int_0^\infty p(\mathbf{v}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r)p_R(r|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) dr, \quad (6)$$

where

$$p_R(r|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) = \frac{p(\mathbf{y}|\boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r)p_R(r)}{\int_0^\infty p(\mathbf{y}|\boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r)p_R(r) dr}, \quad (7)$$

and

$$p(\mathbf{y}|\boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r) \sim \mathcal{N}_n(r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}), r\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n). \quad (8)$$

Using (6), we find that $\widehat{\mathbf{v}}^* = \int_0^\infty E(\mathbf{v}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r)p_R(r|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) dr$, which can be approximated. An approximation, for large sample of size N via the MCMC methods of $p(\mathbf{v}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*)$, could be useful:

$$p(\mathbf{v}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) \approx \frac{1}{N} \sum_{i=1}^N p(\mathbf{v}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R^{(i)}),$$

where $R^{(i)}$ represents samples from the posterior $p_R(r|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*)$.

For predicting new observations \mathbf{y}^* at new inputs $\boldsymbol{\mathfrak{X}}^*$, we need to derive the conditional distribution of $\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*$. To this end, let $B = \begin{bmatrix} \mathbf{I}_n & \mathbf{I}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_m & \mathbf{I}_m \end{bmatrix}$. Then, the joint

vector of observations and predictions, $\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix}$, can be written as $B[\mathbf{v}^T, \boldsymbol{\epsilon}^T, \mathbf{v}^{*T}, \boldsymbol{\epsilon}^{*T}]^T$.

Since the prediction problem requires the knowledge of the conditional distribution of the unobserved data given the observed data, we proceed to derive the conditional distribution of \mathbf{y} given \mathbf{y}^* . To achieve this, we utilize the properties of the multivariate normal distribution. Therefore, the joint conditional distribution of \mathbf{y} and \mathbf{y}^* given $R=r$ is:

$$\begin{aligned}
\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} \Big|_{R=r} &= B \begin{bmatrix} \mathbf{v} \\ \boldsymbol{\epsilon} \\ \mathbf{v}^* \\ \boldsymbol{\epsilon}^* \end{bmatrix} \Big|_{R=r} \\
&\sim \mathcal{N}_{n+m} \left(B \begin{bmatrix} r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}) \\ \mathbf{0} \\ r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}^*) \\ \mathbf{0} \end{bmatrix}, B \begin{bmatrix} r\boldsymbol{\Sigma}_{11} & \mathbf{0} & r\boldsymbol{\Sigma}_{12} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_n & \mathbf{0} & \sigma^2\mathbf{I}_{nm} \\ r\boldsymbol{\Sigma}_{21} & \mathbf{0} & r\boldsymbol{\Sigma}_{22} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_{mn} & \mathbf{0} & \sigma^2\mathbf{I}_m \end{bmatrix} B^T \right), \\
&\sim \mathcal{N}_{n+m} \left(\begin{bmatrix} r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}) \\ r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}^*) \end{bmatrix}, \begin{bmatrix} r\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n & r\boldsymbol{\Sigma}_{12} \\ r\boldsymbol{\Sigma}_{21} & r\boldsymbol{\Sigma}_{22} + \sigma^2\mathbf{I}_m \end{bmatrix} \right).
\end{aligned}$$

Hence, the conditional distribution of $\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r$ is:

$$\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r \sim \mathcal{N}_m(\boldsymbol{\mu}^{***}, \boldsymbol{\Sigma}^{***}),$$

where $\boldsymbol{\mu}^{***} = r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}^*) + r\boldsymbol{\Sigma}_{21}(r\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n)^{-1}(\mathbf{y} - r\boldsymbol{\mu}(\boldsymbol{\mathfrak{X}}))$, and

$$\boldsymbol{\Sigma}^{***} = (r\boldsymbol{\Sigma}_{22} + \sigma^2\mathbf{I}_m) - r\boldsymbol{\Sigma}_{21}(r\boldsymbol{\Sigma}_{11} + \sigma^2\mathbf{I}_n)^{-1}r\boldsymbol{\Sigma}_{12}.$$

Then marginalizing $p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r)$ with respect to $p_R(r|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*)$ yields, the posterior predictive distribution of \mathbf{y}^* :

$$p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) = \int_0^\infty p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R=r) p_R(r|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) dr.$$

where equations (7) and (8) also apply here as well, and MCMC methods can be implemented

$$p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*) \approx \frac{1}{N} \sum_{i=1}^N p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\mathfrak{X}}^*, R^{(i)}).$$

5 Inference with Laplace Approximation and Learning Hyper-parameters

The Laplace approximation is a valuable method for simplifying complex probability distributions by approximating them with Gaussian distributions. It offers a simple method of approximating difficult posterior distributions by concentrating on the mode and curvature of the log-posterior distribution. It identifies the MAP estimate of the latent variables \mathbf{v} in a Bayesian model is the first step in the inference process. Next, a Gaussian distribution is used to approximate the posterior distribution $p(\mathbf{v}|\mathbf{y}, \boldsymbol{\mathfrak{X}}, \boldsymbol{\theta})$

around this MAP estimate, $q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta})$. Although the quality may deteriorate further beyond the MAP estimate, this Gaussian approximation is usually correct in the vicinity of this estimate. Crucially, a unimodal posterior distribution with well-behaved curvature close to the MAP estimate is assumed by the Laplace approximation (Rasmussen, 2006).

5.1 Computing the Posterior with Laplace Approximation

For models where the exact posterior distribution $p(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta})$ is intractable, approximate methods such as the Laplace approximation are essential. The posterior can be expressed as:

$$p(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{v}, \mathfrak{X}, \boldsymbol{\theta})p(\mathbf{v}|\mathfrak{X}, \boldsymbol{\theta})}{\int_{\mathbb{R}^n} p(\mathbf{y}|\mathbf{v}, \mathfrak{X}, \boldsymbol{\theta})p(\mathbf{v}|\mathfrak{X}, \boldsymbol{\theta}) d\mathbf{v}} = \frac{Q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta})}{p(\mathbf{y}|\mathfrak{X}, \boldsymbol{\theta})}. \quad (9)$$

Here, $Q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta})$ represents the unnormalized posterior, given by:

$$\begin{aligned} \log Q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) &= \log p(\mathbf{y}|\mathbf{v}, \mathfrak{X}, \boldsymbol{\theta}) + \log p(\mathbf{v}|\mathfrak{X}, \boldsymbol{\theta}), \\ &= \log p(\mathbf{y}|\mathbf{v}, \mathfrak{X}, \boldsymbol{\theta}) - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}_{11}| - \frac{1}{2} \mathbf{v}^T \boldsymbol{\Sigma}_{11}^{-1} \mathbf{v}. \end{aligned} \quad (10)$$

The Laplace approximation entails taking the second derivative of this log-posterior with respect to \mathbf{v} and evaluating it at the MAP estimate $\hat{\mathbf{v}}$:

$$-\nabla \nabla \log Q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) \Big|_{\mathbf{v}=\hat{\mathbf{v}}} = -\nabla \nabla \log p(\mathbf{y}|\mathbf{v}, \mathfrak{X}, \boldsymbol{\theta}) \Big|_{\mathbf{v}=\hat{\mathbf{v}}} + \boldsymbol{\Sigma}_{11}^{-1}.$$

Thus, we define

$$\mathbf{A} = \mathbf{W} + \boldsymbol{\Sigma}_{11}^{-1}, \quad (11)$$

where \mathbf{A} represents the Hessian of the negative log-posterior evaluated at $\hat{\mathbf{v}}$.

The MAP estimate, $\hat{\mathbf{v}}$, is found by maximizing the posterior distribution:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmax}} p(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}), \quad (12)$$

and Newton's method can be used to iteratively solve for $\hat{\mathbf{v}}$:

$$\mathbf{v}_{\text{new}} \leftarrow \mathbf{v}_{\text{old}} - \frac{\nabla \log Q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) \Big|_{\mathbf{v}=\mathbf{v}_{\text{old}}}}{\nabla \nabla \log Q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) \Big|_{\mathbf{v}=\mathbf{v}_{\text{old}}}}. \quad (13)$$

With a local expansion around $\hat{\mathbf{v}}$, the log-posterior is approximated as:

$$\log Q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) \approx \log Q(\hat{\mathbf{v}}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) - \frac{1}{2} (\mathbf{v} - \hat{\mathbf{v}})^T \mathbf{A} (\mathbf{v} - \hat{\mathbf{v}}).$$

The resulting Gaussian approximation around $\hat{\mathbf{v}}$ enables efficient computation of posterior summaries.

5.2 Hyperparameters Learning

To optimize the hyperparameters $\boldsymbol{\theta}$, we maximize the marginal likelihood $p(\mathbf{y}|\mathfrak{X}, \boldsymbol{\theta})$. Given the Gaussian approximation for $\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}$, we approximate the marginal likelihood as:

$$\begin{aligned} p(\mathbf{y}|\mathfrak{X}, \boldsymbol{\theta}) &= \int_{\mathbb{R}^n} p(\mathbf{y}|\mathbf{v}, \mathfrak{X}, \boldsymbol{\theta}) p(\mathbf{v}|\mathfrak{X}, \boldsymbol{\theta}) d\mathbf{v}, \\ &\approx Q(\hat{\mathbf{v}}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) (2\pi)^{\frac{n}{2}} |\mathbf{A}^{-1}|^{\frac{1}{2}}. \end{aligned} \quad (14)$$

Thus, by maximizing (14), we estimate $\boldsymbol{\theta}$, which can then be used in the Gaussian approximation $q(\mathbf{v}|\mathbf{y}, \mathfrak{X}, \boldsymbol{\theta}) \sim \mathcal{N}_n(\hat{\mathbf{v}}, \mathbf{A}^{-1})$. For details on the full procedure, see sources such as (Kuss and Rasmussen, 2006), (Bishop, 2006) and (Rasmussen, 2006).

6 Elliptical Processes in the Weight Space

Due to the computational complexity of a fully Bayesian EPR, representing the model in the weight space provides a feasible alternative. In this approach, an EP is represented as a linear combination of basis functions with elliptically distributed weights, allowing flexibility in modeling non-Gaussian data characteristics. In the weight-space formulation, an EP can be expressed as:

$$v(\mathbf{x}) = \sum_{j=0}^{\infty} \beta_j \phi_j(\mathbf{x}) \approx \sum_{j=0}^J \beta_j \phi_j(\mathbf{x}), \quad (15)$$

where $\{\beta_j\}_{j=0}^J$ are elliptically distributed prior weights, $\phi_j(\mathbf{x})$ are basis functions, and J is suitable number of basis functions. Specifically, if $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_J)^\top$ follows an elliptical distribution $\mathcal{E}_{J+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, then $v(\mathbf{x})$ inherits the elliptical properties, allowing it to model a variety of tail behaviors depending on the choice of the density generator $g(\cdot)$. The observed output array \mathbf{y} can then be represented as:

$$\mathbf{y} = \phi(\mathfrak{X})\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\phi(\mathfrak{X})$ is the design matrix of basis functions derived from the full dataset \mathfrak{X} , $\boldsymbol{\beta}$ represents the vector of elliptically distributed weights, and $\boldsymbol{\epsilon}$ is Gaussian noise. Given $\phi(\mathfrak{X})$, the model becomes:

$$\mathbf{y}|\mathfrak{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_n(\Phi\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \quad (16)$$

where $\Phi \in \mathbb{R}^{n \times (J+1)}$ is the design matrix of basis functions, with $\Phi_{ij} = \phi_j(\mathbf{x}_i)$. Good candidates for such basis functions are the probabilistic Hermite polynomials $H_j(x)$,

which are defined recursively as:

$$H_0(x) = 1, \quad H_1(x) = 2x, \quad H_{j+1}(x) = 2xH_j(x) - 2jH_{j-1}(x).$$

Using Hermite polynomials as basis functions in an EP enables the capture of complex, nonlinear relationships in data. Then (15) can be written as:

$$v(\mathbf{x}) = \sum_{j=0}^J \beta_j H_j(\mathbf{x}),$$

where $\{\beta_j\}_{j=0}^J$ are the elliptically distributed weights, and $H_j(\mathbf{x})$ represents the j -th Hermite polynomial evaluated at \mathbf{x} . For more information, see (Szegő, 1939) and (Rahman, 2017).

In practice, our approach is designed with several choices for the prior distribution of β , outlined as follows:

- Student-t distributions: $\beta \sim \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \text{df})$, where $\boldsymbol{\mu}$ is the location vector, $\boldsymbol{\Sigma}$ is the scale matrix, and df (degrees of freedom) controls the heaviness of the tails, with values $\text{df} = 5, 10, 20, 50$.
- Asymmetric Laplace distribution: $\beta \sim \text{AL}(\boldsymbol{\mu}, b, k)$, where $\boldsymbol{\mu}$ is the location parameter, b is the scale parameter, and k is the asymmetry parameter.
- Cauchy distribution: $\beta \sim \mathcal{C}(\boldsymbol{\mu}, \gamma)$, where $\boldsymbol{\mu}$ is the location parameter and γ is the scale parameter.
- Laplace distribution: $\beta \sim \text{Laplace}(\boldsymbol{\mu}, b)$, with $\boldsymbol{\mu}$ as the location parameter and b as the scale parameter.
- Power Exponential distribution: $\beta \sim \text{PE}(\boldsymbol{\mu}, b, s)$, where $\boldsymbol{\mu}$ is the location, b is the scale, and s is the shape parameter, which controls tail behavior.
- Gaussian distribution: $\beta \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}$ as the mean vector and $\boldsymbol{\Sigma}$ as the covariance matrix.

For more explanation, see (Fang et al., 2018) and (Gelman et al., 2004).

The posterior distribution of β and σ^2 (parameterized as $\Theta = \{\beta, \log(\sigma^2)\}$) can be efficiently approximated using the Laplace approximation in the framework of Bayesian inference for the model given in Equation (16). The Maximum A Posteriori (MAP)

estimate, $\hat{\Theta} = \{\hat{\beta}, \log(\hat{\sigma}^2)\}$, represents the center of a Gaussian distribution that approximates the joint posterior distribution $p(\Theta|\mathbf{y}, \mathcal{X}, \Phi)$. Focusing on β , the marginal posterior $p(\beta|\mathbf{y}, \mathcal{X}, \Phi, \sigma^2)$ is central to inference. The posterior for β , given the observed data $\mathcal{D} = \{(\mathcal{X}, \mathbf{y})\}$ is obtained by combining the likelihood $p(\mathbf{y}|\mathcal{X}, \Phi, \Theta)$ and a prior on β , $p(\beta)$. The log-posterior can be approximated as a quadratic function around $\hat{\beta}$, the MAP estimate.

The Laplace approximation entails evaluating the Hessian matrix of the negative log-posterior with respect to Θ at $\hat{\Theta}$:

$$\mathcal{H} = -\nabla^2 \log p(\Theta|\mathbf{y}, \mathcal{X}, \Phi) \Big|_{\Theta=\hat{\Theta}}.$$

Then, the posterior distribution $p(\Theta|\mathbf{y}, \mathcal{X}, \Phi)$ is approximated by:

$$p(\Theta|\mathbf{y}, \mathcal{X}, \Phi) \approx \mathcal{N}(\hat{\Theta}, \mathcal{H}^{-1}).$$

By including $\log(\sigma^2)$ in Θ , we account for uncertainty in the noise scale. A prior on $\log(\sigma^2)$, typically chosen as $\log(\sigma^2) \sim \mathcal{N}(\mu_{\log(\sigma^2)}, \Sigma_{\log(\sigma^2)})$, we used $\mu_{\log(\sigma^2)} = 0$ and $\Sigma_{\log(\sigma^2)} = 1$. Taking the logarithm of σ^2 also helps avoid optimization algorithms getting stuck at the boundary of $(0, \infty)$, thereby improving convergence and ensuring stability in the estimation process.

The posterior distribution of Θ can be expressed as:

$$p(\Theta|\mathbf{y}, \mathcal{X}, \Phi) \propto p(\mathbf{y}|\mathcal{X}, \Phi, \Theta)p(\Theta),$$

where the prior $p(\Theta)$ incorporates priors on β and $\log(\sigma^2)$. The Laplace approximation extends to the entire parameter set Θ , enabling joint optimization of β and $\log(\sigma^2)$. This Gaussian approximation simplifies inference on Θ , and posterior summaries and predictive distributions can be efficiently computed. The approach used here aligns with the techniques for deriving the marginal likelihood as described in (Bishop, 2006) and (Rasmussen, 2006).

6.1 Application to Simulated Data

This simulation study compare the performance of EPR models (non-Gaussian models) with the most common used model (the Gaussian model) by introducing outliers at different levels and comparing the results using Mean Squared Error (MSE) for both

training and test datasets. This approach allows us to examine how different models handle data deviations, such as heavy tails and skewness, that often arise in real-world applications.

Example 1: The first simulation experiment uses the output function:

$$f(x) = e^{-\frac{x^2}{10}} + x \sin(x),$$

with x values evenly spaced over $[-5, 5]$. For this setup, we generated 100 observations with added Gaussian noise for training and 15 test observations. The models tested include Student-t distributions with different degrees of freedom 5, 10, 20, 50, denoted respectively by (T5, T10, T20, T50), Asymmetric Laplace (AL), Cauchy (C), Laplace (L), Power Exponential (PE), and GP.

Outliers were introduced into the training data at 5%, 10%, and 20% levels. The number of outliers was determined based on the desired percentage (5%, 10%, and 20% of the training data) and randomly selected positions in the response variable. Their values were generated from a normal distribution with zero mean and large variance ($\text{Var}(\epsilon_{\text{outliers}}) = 12^2 = 144$) to ensure extremity. Negative outliers were set near the minimum value of the response, while positive outliers were set near the maximum. These modified values replaced the original values at the selected positions, creating distinct anomalies in the simulated dataset.

Table 2 provides descriptive statistics, highlighting the growing spread as the outlier percentage increases. The MSE results for training and test data are shown in Tables 3 and 4, respectively.

Table 2: Descriptive Statistics for Simulated Data for Example 1 with Different Outlier Levels

Statistic	No Outliers	5% Outliers	10% Outliers	20% Outliers
Minimum	-4.9755	-18.6051	-37.1447	-141.7652
Median	1.1467	1.2151	1.0995	0.6314
Maximum	3.3925	37.0767	61.7211	80.1186
Mean	0.0687	0.5191	1.2727	-9.0135
Standard Deviation	2.5016	5.9947	12.9809	37.2353
Skewness	-0.7639	2.7046	2.4245	-1.7871
num. of Outliers	0	5	10	20

Table 3: MSE Results for Simulated Data for Example 1 (Training Data)

Model	No Outliers	5% Outliers	10% Outliers	20% Outliers
T5	0.2126	27.1231	151.0524	1337.0620
T10	0.2133	27.1063	151.0267	1334.7000
T20	0.2129	27.0697	151.3325	1334.8590
T50	0.2128	27.1095	151.2625	1336.7570
AL	0.2150	27.8473	157.6951	1391.1710
C	0.2124	27.1030	151.2303	1337.1940
L	0.2125	27.1111	150.9032	1337.1560
PE	0.2143	27.1417	151.1884	1336.7970
GP	0.2127	27.1334	151.3482	1335.0250

It can be notice from Table 4 That the AL model consistently shows robustness against outliers, achieving the lowest MSE values across all levels of outliers, highlighting its potential in scenarios with heavy-tailed noise.

Table 4: MSE Results for Simulated Data for Example 1 (Test Data)

Model	No Outliers	5% Outliers	10% Outliers	20% Outliers
T5	0.2414	4.0065	5.1191	108.7690
T10	0.2391	3.9416	5.1718	109.2632
T20	0.2376	3.7520	4.9718	110.5813
T50	0.2327	4.0548	5.0328	111.0066
AL	0.2256	2.0913	2.6810	17.1864
C	0.2374	4.2459	5.4981	109.5441
L	0.2353	3.9784	5.4553	111.6354
PE	0.2285	3.8981	5.5296	110.4472
GP	0.2358	3.9873	4.6253	113.5658

The results demonstrate that, while all models perform comparably in the absence of outliers, their sensitivity to outliers varies significantly. Non-Gaussian models performed better in the presence of outliers. the AL model demonstrates robust performance across various outlier levels, consistently maintaining low MSE values in test datasets. This finding underscores the effectiveness of the AL model in capturing heavy-tailed noise and providing outlier-resistant modeling, emphasizing the importance of expanding research into skewed Elliptical Processes (SEP).

These findings are further supported by visualizations of the fitted models especially Figure 2 across different levels of outliers. Particularly, the AL model’s performance is visually closer to the underlying data, even with injected outliers, highlighting its robustness relative to other elliptical models.

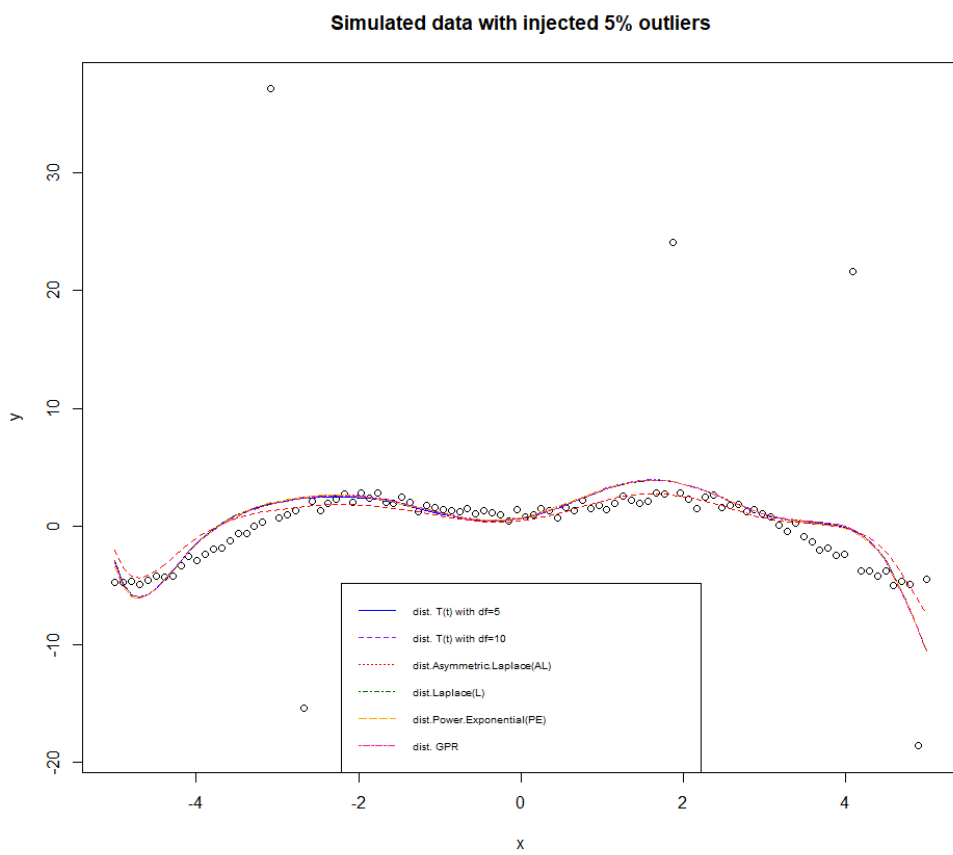
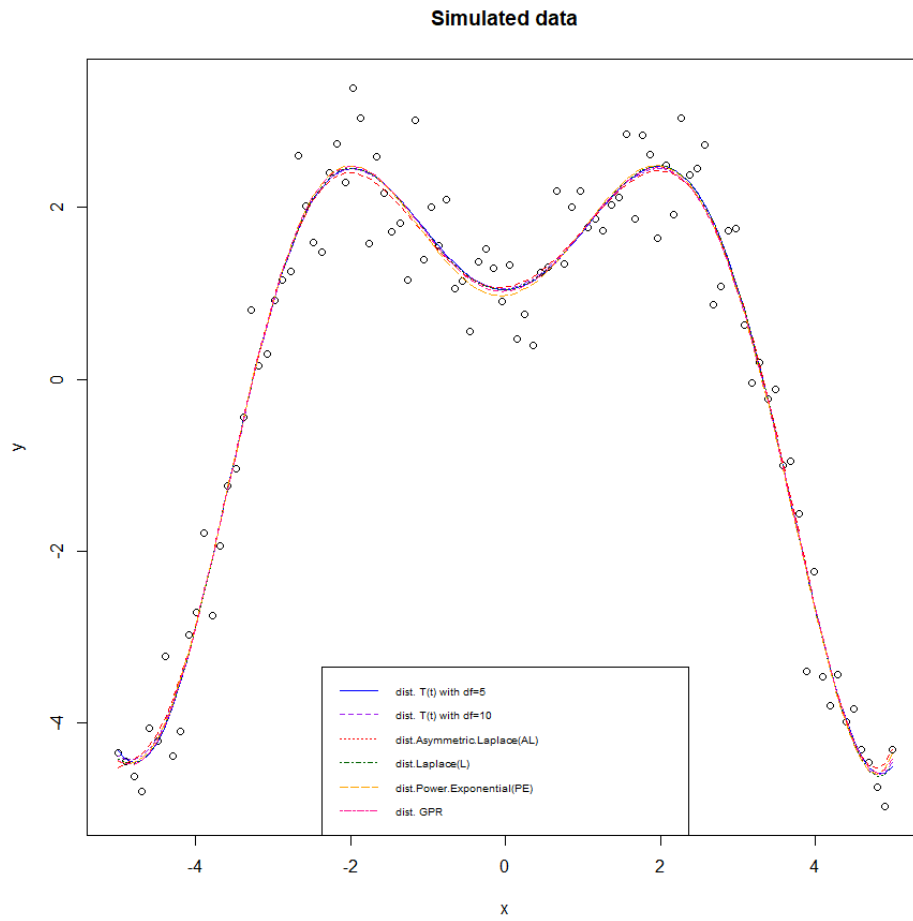
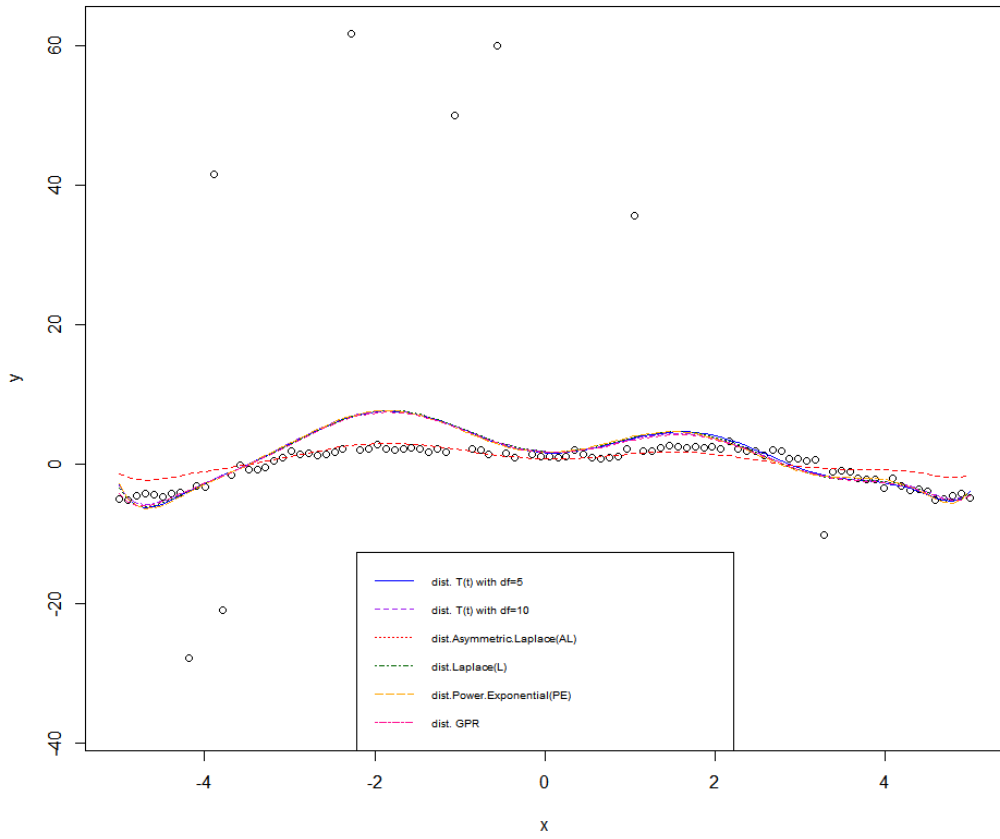


Figure 1: Simulated Data Analysis for Example 1: Model predictions with no outliers and 5% outliers.

Simulated data with injected 10% outliers



Simulated data with injected 20% outliers

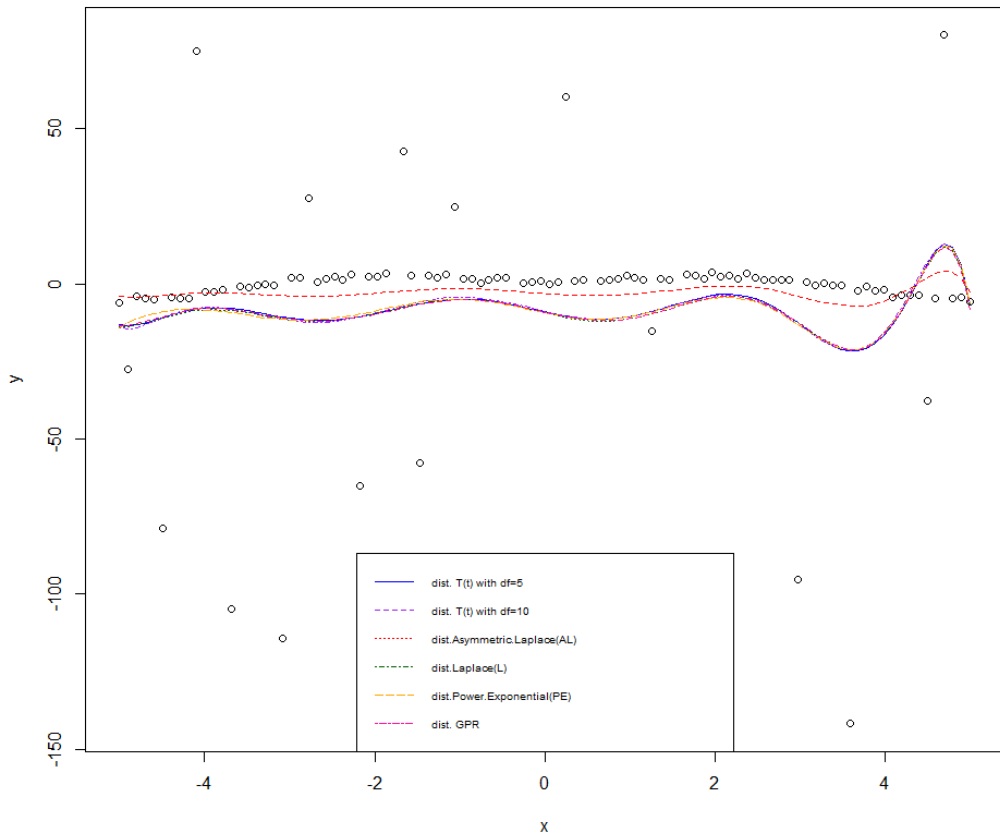


Figure 2: Simulated Data Analysis for Example 1: Model predictions with 10% outliers and 20% outliers.

Example 2: The second simulation experiment uses a different function to provide further insights:

$$f(x) = 2 \sin(1.5x) \exp\left(-\frac{x^2}{8}\right),$$

with x values in the range $[-4, 4]$. Similar to the first example, this setup includes 100 training observations and 15 test points. Models tested remain the same as in Example 1.

Descriptive statistics for this example across various outlier levels are shown in Table 5, reflecting increased variability with added outliers. Tables 6 and 7 summarize the MSE results for training and test data. Consistent with the findings in Example 1, the AL model outperforms others in terms of MSE across higher outlier levels, underscoring its robustness under heavy-tailed noise conditions.

Table 5: Descriptive Statistics for Example 2 with Incremental Outliers

Statistic	No Outliers	5% Outliers	10% Outliers	20% Outliers
Minimum	-2.524	-12.995	-45.071	-100.100
Median	0.0003	-0.167	-0.025	0.096
Maximum	3.173	52.998	51.423	91.571
Mean	0.0214	0.672	0.378	1.750
Standard Deviation	1.128	6.379	11.172	23.849
Skewness	0.138	6.083	0.188	-0.059
num. of Outliers	0	5	10	20

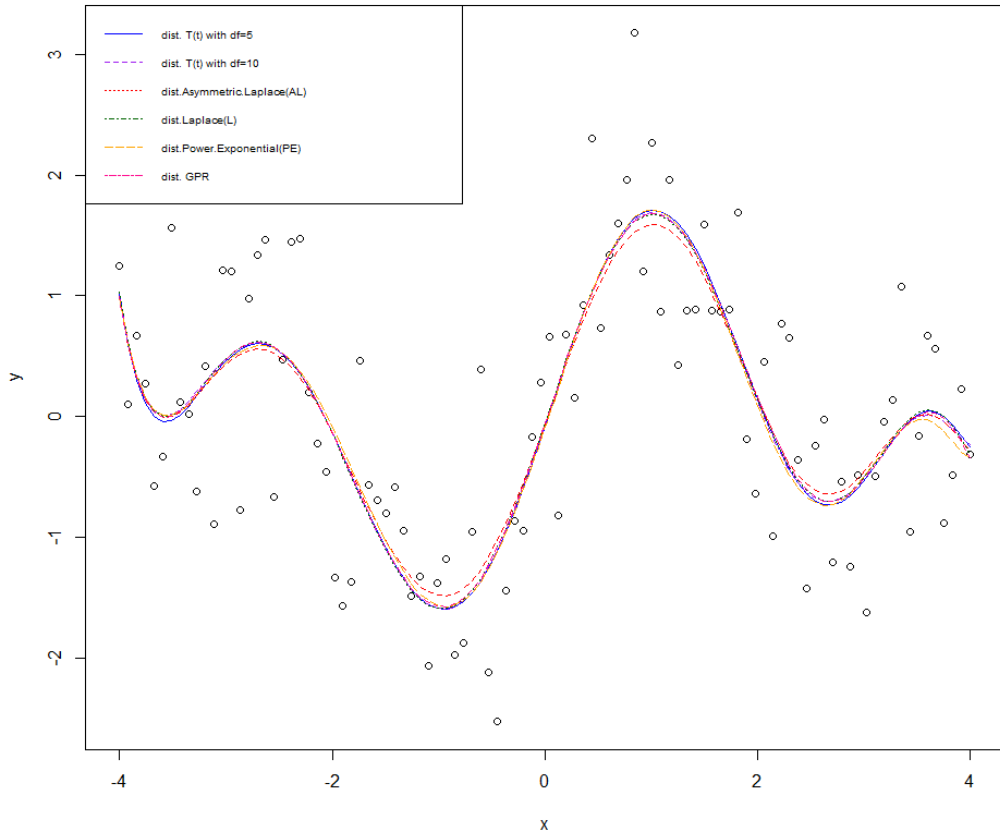
Table 6: MSE Results for Example 2 (Training Data)

Model	No Outliers	5% Outliers	10% Outliers	20% Outliers
T5	0.4981	30.6792	109.9236	516.2366
T10	0.4980	30.6365	109.9261	516.0781
T20	0.4975	30.6559	109.8089	515.8187
T50	0.4979	30.6401	109.7542	514.9786
AL	0.5006	31.4041	114.4363	537.7384
C	0.4976	30.6351	109.7840	515.5270
L	0.4976	30.5914	109.7198	515.1100
PE	0.4991	30.6589	109.9562	515.5150
GPR	0.4968	30.6126	109.9069	515.9277

Table 7: MSE Results for Example 2 (Test Data)

Model	No Outliers	5% Outliers	10% Outliers	20% Outliers
T5	0.5423	24.7433	9.6571	47.0600
T10	0.5459	24.0543	9.6175	47.7361
T20	0.5421	24.4608	9.9843	47.7402
T50	0.5367	24.6002	9.5744	47.0242
AL	0.5371	13.1293	1.6139	6.7137
C	0.5284	25.0833	9.4549	46.4362
L	0.5370	24.3839	9.4875	47.6020
PE	0.5306	24.1452	10.0239	47.4196
GPR	0.5258	24.7649	9.8065	46.7006

Simulated data



Simulated data with injected 5% outliers

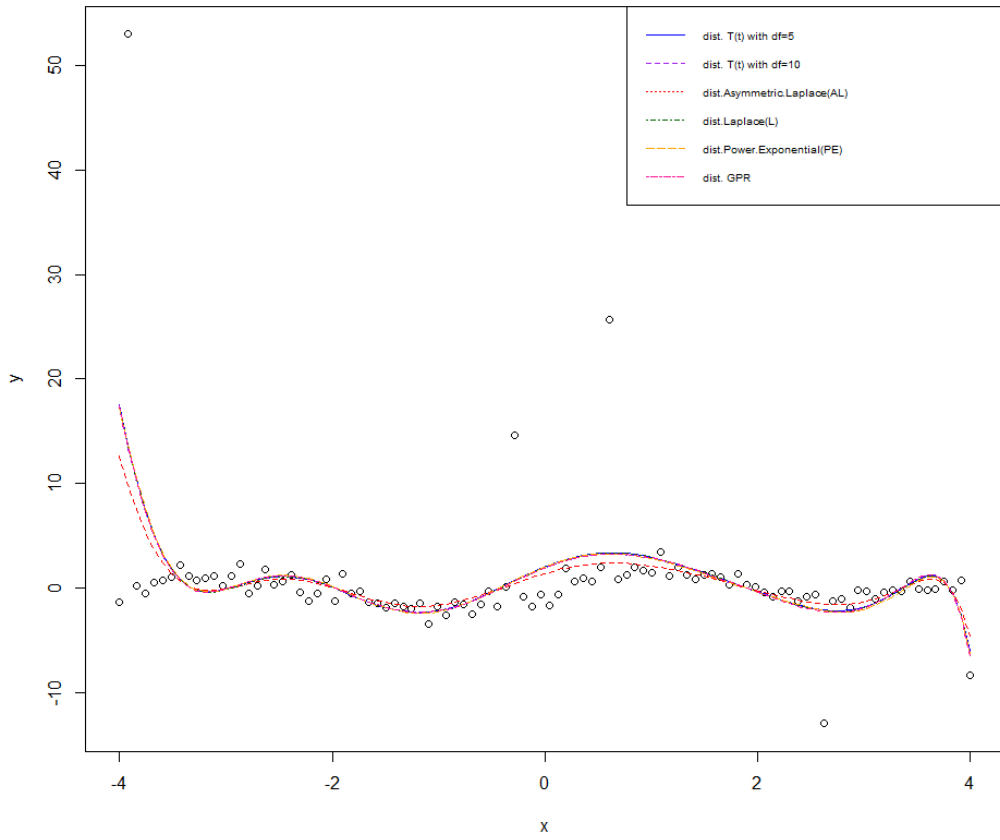
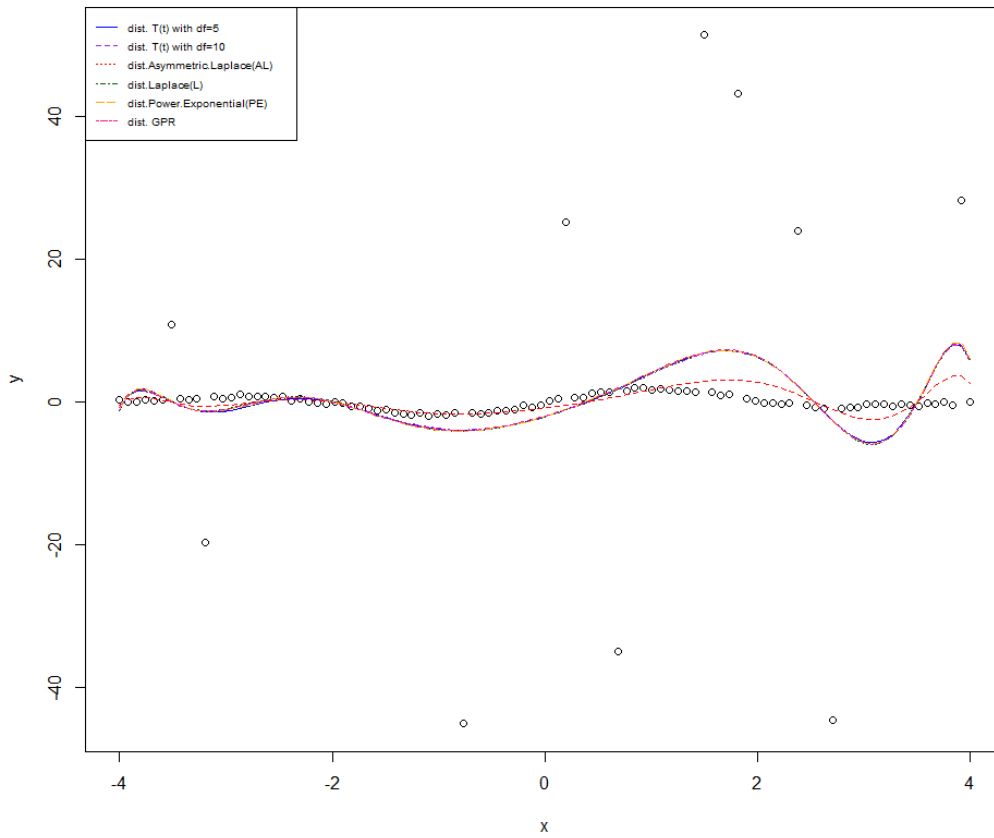


Figure 3: Simulated Data Analysis for Example 2: Model predictions with no outliers and 5% outliers.

Simulated data with injected 10% outliers



Simulated data with injected 20% outliers

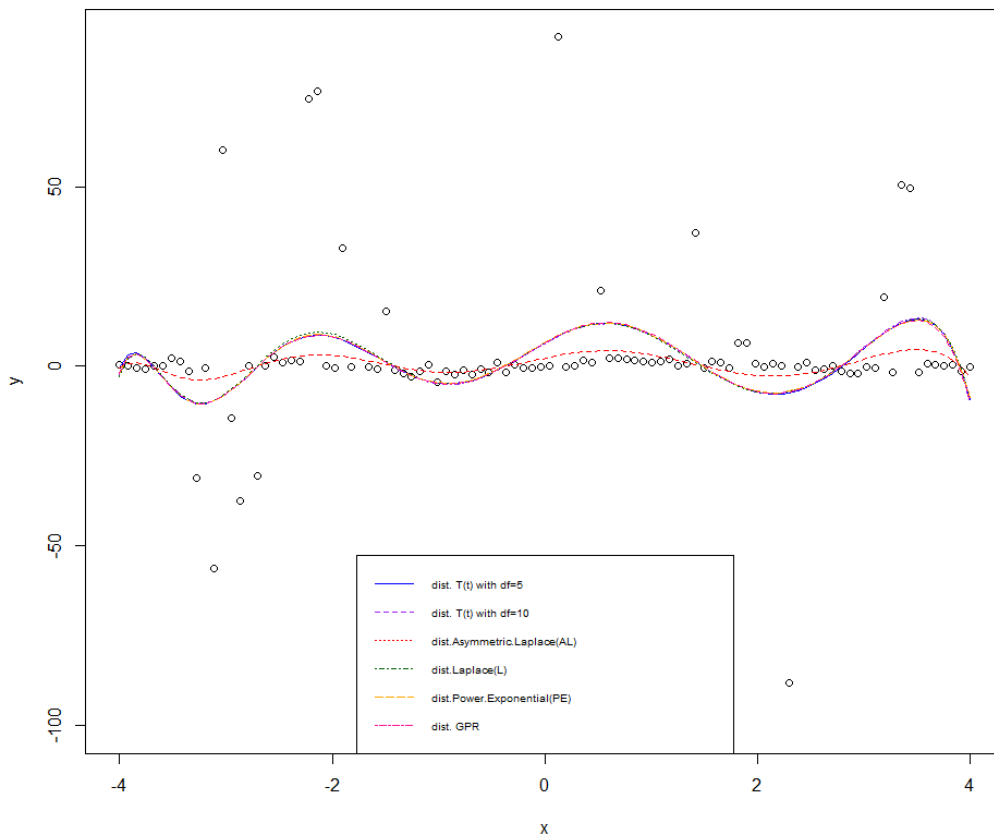


Figure 4: Simulated Data Analysis for Example 2: Model predictions with 10% outliers and 20% outliers.

Across both examples, results demonstrate that while all models perform comparably with no outliers, sensitivity to outliers differs considerably as their levels increase. The AL model exhibits the most consistent performance, particularly in high outlier scenarios, confirming its capability in managing heavy-tailed noise. These findings are clearly visualized in Figure 2 and Figure 4, which illustrate model predictions at high (10% and 20%) outlier levels for both examples. The AL model likely performed well in test datasets with outliers due to its ability to handle asymmetric, heavy-tailed distributions. Unlike symmetric models like the Gaussian, the AL model adapts to data with skewness or one-sided outliers, making it robust in cases where one tail is longer or heavier. By down-weighting the influence of extreme values on one side of the distribution through the L1 norm, the AL model maintains low MSE even with irregular data points. This flexibility makes the AL model effective in non-Gaussian settings, providing reliable predictions where other models might be overly influenced by extreme values. More about AL model in (Yu, 2005).

6.2 Application to Real Data

In this section, we applied our theoretical findings to the datasets described below:

Natural Gas Production Data Set: Monthly data from NCSI Oman (Jan 2014 - Dec 2021) with 96 observations, detailing associated and non-associated gas production in standard cubic feet.

MSX30 Data Set: Monthly closing values of the MSX30 Index in Oman (July 2009 - Oct 2023, 173 observations) from the Muscat Stock Exchange website, reflecting market trends in Oman's top 30 companies.

Crude Oil Prices Data Set: Monthly crude oil prices in Oman from NCSI (Jan 1994 - Jan 2023), across 349 observations, capturing long-term trends in the oil sector.

Bike Sharing Data Set: Daily bike rental counts in Washington D.C. over 731 days (2011 - 2012), including both casual and registered users, available on Kaggle.

Table 8: Descriptive Statistics for Real Data Sets

Statistic	Gas		MSX30	Oil	Bike-Count
	Non-Associated	Associated			
Minimum	-2.3506	-1.0523	-0.8747	-0.6909	-1.1569
Median	-0.0791	-0.0975	0.0807	-0.0530	0.0151
Maximum	1.3801	1.2028	1.0196	1.2426	1.0865
Mean	0.0166	-0.0095	0.0114	0.0015	0.0005
Standard Deviation	0.5166	0.4758	0.4977	0.5062	0.5020
Skewness	-0.3980	0.3216	-0.0321	0.4733	0.0623
Kurtosis	8.3061	2.6792	1.7364	2.0799	2.1965
num. of Outliers	6	0	0	0	0

In Table 8, the descriptive statistics indicate that most of the used real datasets have few or no outliers, as identified by the IQR method. Notably, the Gas (Non-Associated) dataset has 6 detected outliers, while the other datasets have none. Further, the violin plot in Figure 5 visualizes the distribution, spread, and shape of the transformed datasets used, side-by-side, clearly illustrating their tail behavior, outliers, and skewness.

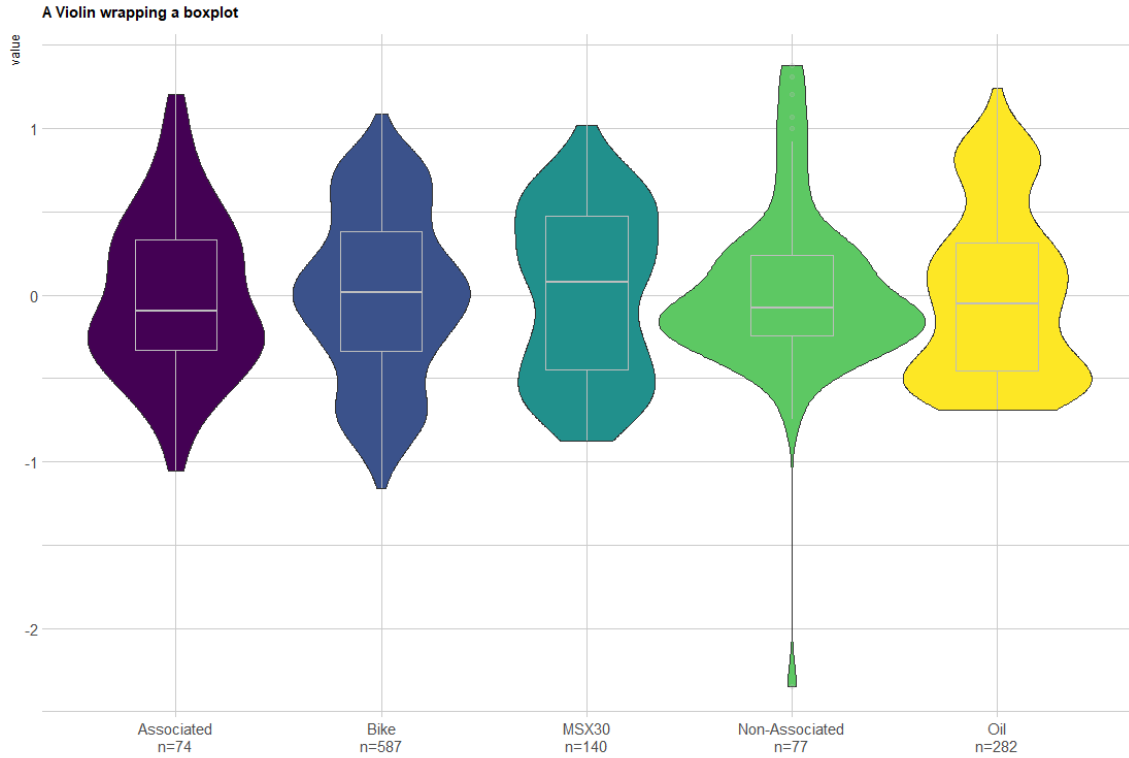


Figure 5: Violin-boxplot visualization of the 5 real datasets, showing the distribution, density, and summary statistics (e.g., median, quartiles, and outliers) for each transformed dataset.

Tables 9 and 10 present the MSE results for both training and test points across the real datasets for the models T5, T10, T20, T50, AL, C, L, PE, and GP. While the Student’s-t process has been extensively studied and compared with the GP (Tracey and Wolpert, 2018), for its robustness to outliers and adaptive variance capabilities, no prior research has examined or compared the GP with alternative processes, such as the C, L, PE, and AL processes. This study addresses this critical gap by introducing these processes and systematically evaluating their performance, offering novel insights into their utility and applications.

Table 9: MSE for Training Data Points

Model	Gas-Non-Associated	Gas-Associated	MSX30	Oil	Bike-Count	
T5	0.0812	0.0656	0.0140	0.0494	0.0627	
T10	0.0808	0.0658	0.0140	0.0492	0.0627	
T20	0.0804	0.0659	0.0140	0.0493	0.0627	
T50	0.0802	0.0656	0.0140	0.0493	0.0626	
{	AL	0.0812	0.0661	0.0140	0.0495	0.0627
	C	0.4561	0.0839	0.0623	0.0496	0.0628
	L	0.0814	0.0661	0.0143	0.0502	0.0628
	PE	0.0800	0.0660	0.0138	0.0489	0.0627
GP	0.0803	0.0657	0.0139	0.0492	0.0627	

Table 10: MSE for Test Data Points

Model	Gas-Non-Associated	Gas-Associated	MSX30	Oil	Bike-Count	
T5	0.0689	0.0639	0.0211	0.0354	0.0741	
T10	0.0695	0.0677	0.0207	0.0354	0.0741	
T20	0.0684	0.0642	0.0207	0.0352	0.0739	
T50	0.0666	0.0653	0.0208	0.0352	0.0741	
{	AL	0.0664	0.0671	0.0207	0.0355	0.0742
	C	0.5508	0.3164	0.0517	0.0358	0.0744
	L	0.0698	0.0733	0.0213	0.0361	0.0746
	PE	0.0675	0.0681	0.0203	0.0346	0.0738
GP	0.0669	0.0680	0.0207	0.0352	0.0740	

In general, for datasets without significant outliers (Gas-Associated, MSX30, Oil, and Bike-Count), the PE and T models often achieve the lowest MSE across both training

(Table 9) and test data points (Table 10). For a dataset containing outliers (Gas-Non-Associated), the PE model demonstrates the best performance on the training dataset, achieving the lowest MSE values. Meanwhile, the AL model performs robustly on the test dataset, effectively managing the presence of outliers. For the Gas-Non-Associated, Gas-Associated, and MSX30 datasets, the C model consistently exhibits the highest MSE.

Overall, PE emerges as the most effective model for real datasets without excessively extreme outliers, consistently yielding lower MSE values. The PE model’s success may be attributed to its adaptability in capturing diverse data characteristics. With a parameter that adjusts tail “thickness,” the PE model can flexibly respond to variations in data structure, providing robustness against moderate outliers and data variability. This flexibility enables the PE model to balance sensitivity to central trends while resisting the undue influence of outliers. Unlike the Gaussian model, which has a fixed tail weight, the PE model accommodates data with heavier tails, effectively capturing both core trends and minor deviations. This adaptability makes it particularly suitable for non-Gaussian distributions with moderate data deviations, resulting in low MSE values and strong performance in both training and test datasets. For more on the Power Exponential distribution properties and its handling of moderate outliers in non-Gaussian data, see (Gómez, 1998).

To explore the 5 real-world datasets and compare the Elliptical models used, refer to Figures 6, ??, and ?. For a summary of the results, refer to Figure 9.

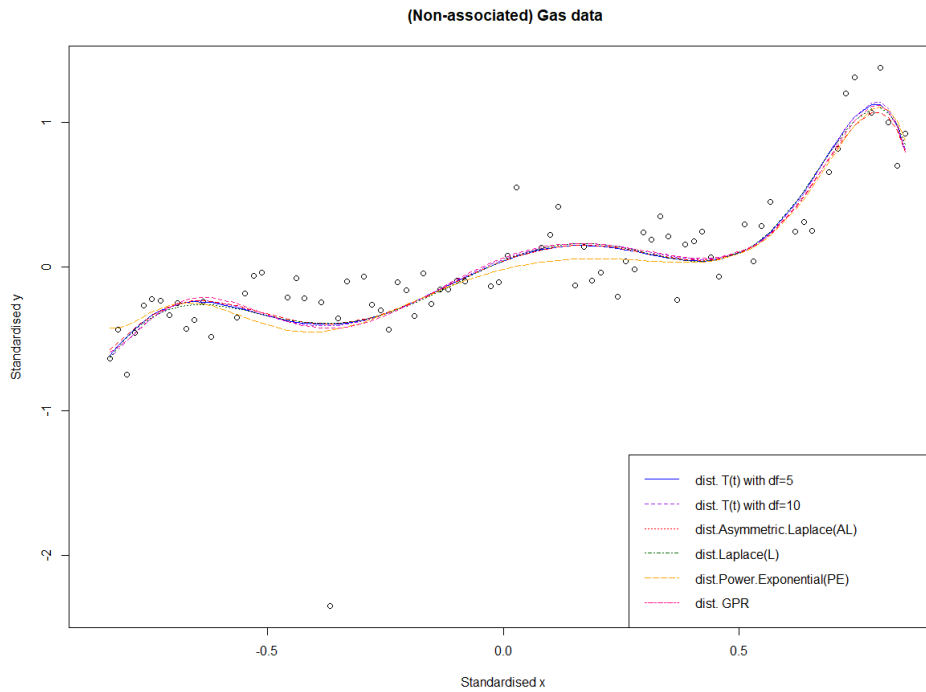
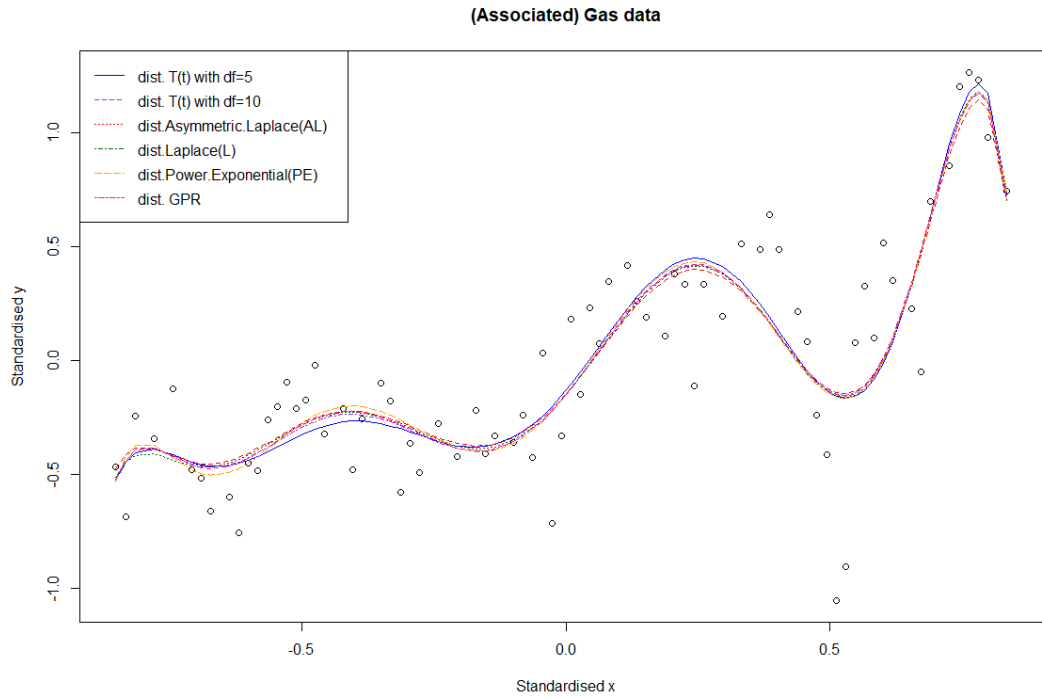


Figure 6: Comparison of Different Elliptical Models on Gas-Associated Dataset and Gas-Non-Associated Dataset.

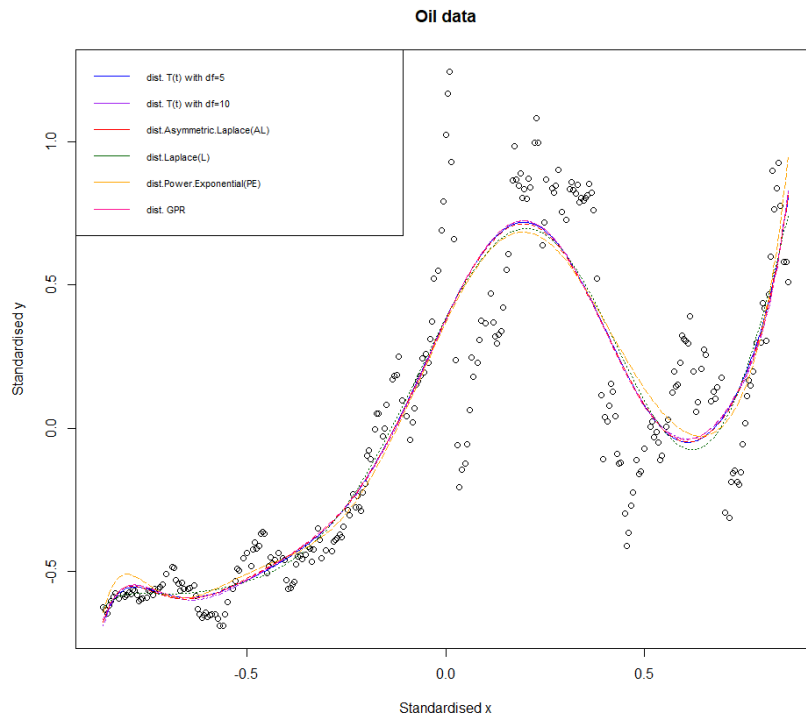
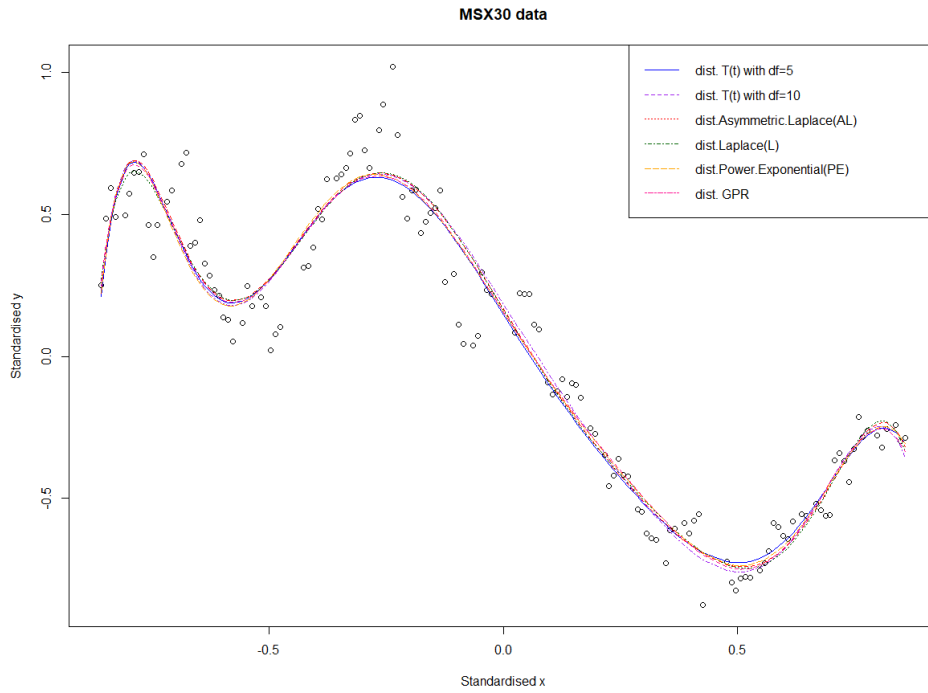


Figure 7: Comparison of Different Elliptical Models on MSX30 Dataset and Oil Dataset.

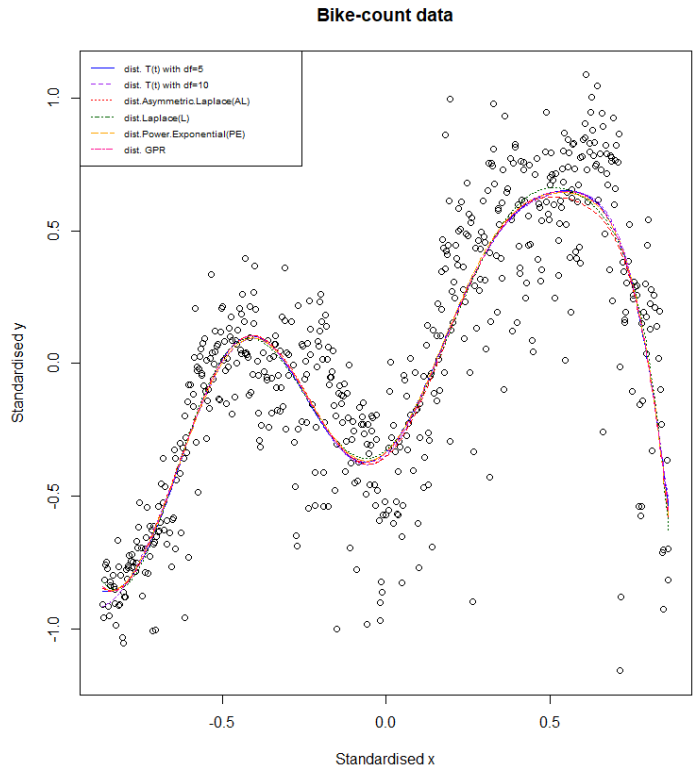


Figure 8: Comparison of Different Elliptical Models on Bike-Count Dataset.

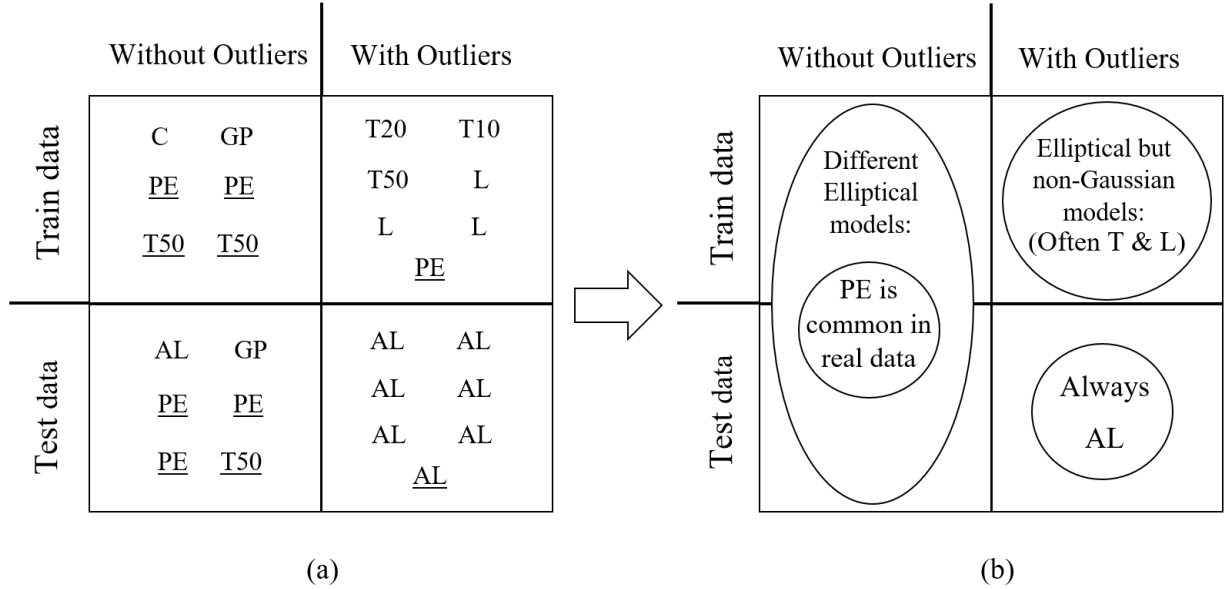


Figure 9: Comparing the performance of EPR models, including non-Gaussian and Gaussian models. The study introduces outliers and evaluates MSE on both training and test datasets. (a) The best model for each situation of the simulated datasets and real-world datasets, which are underlined. (b) Summary of findings, highlighting the robustness of non-Gaussian models under conditions with data deviations.

7 Conclusion and Directions of Future Work

This paper introduces a fully Bayesian framework for non-parametric regression based on EPs, advancing beyond traditional Gaussian models to incorporate flexible, heavy-tailed distributions. The EPR model leverages Bayesian inference to integrate uncertainty in both model parameters and latent variables, accommodating various elliptical distributions such as the AL, Student-t, Cauchy, and PE processes. This adaptability renders EPR particularly valuable for datasets characterized by non-Gaussian noise and heavy tails.

Because EPR is fully Bayesian, uncertainty is thoroughly handled, producing reliable predictions and believable intervals that adapt to the complexity of the data. Simulation tests and real-world applications show how resilient the model is against outliers, non-normality, and irregular data, making EPR a strong contender to replace Gpr in situations where Gaussian assumptions would not hold true.

The ability of EPR to handle heavy-tailed distributions with flexibility highlights its

potential in domains that need complicated modeling of real-world complications. Although GPR provides computational simplicity, EPR's Bayesian framework produces accurate predictions and nuanced insights, particularly when outliers are present. The Bayesian method developed by EPR is well-suited for a variety of applications that demand accurate predictions and strong data interpretation as computing power increases. The theoretical advantages of EPR are empirically validated, demonstrating its suitability for intricate regression tasks and its potential for widespread use in data-intensive domains.

The research demonstrates the clear superiority of certain non-Gaussian models. As illustrated in Figure 10, the more comprehensive the model, the better it captures diversity beyond the limited framework of special cases like Gaussian or Cauchy, which apply only specific scenarios.

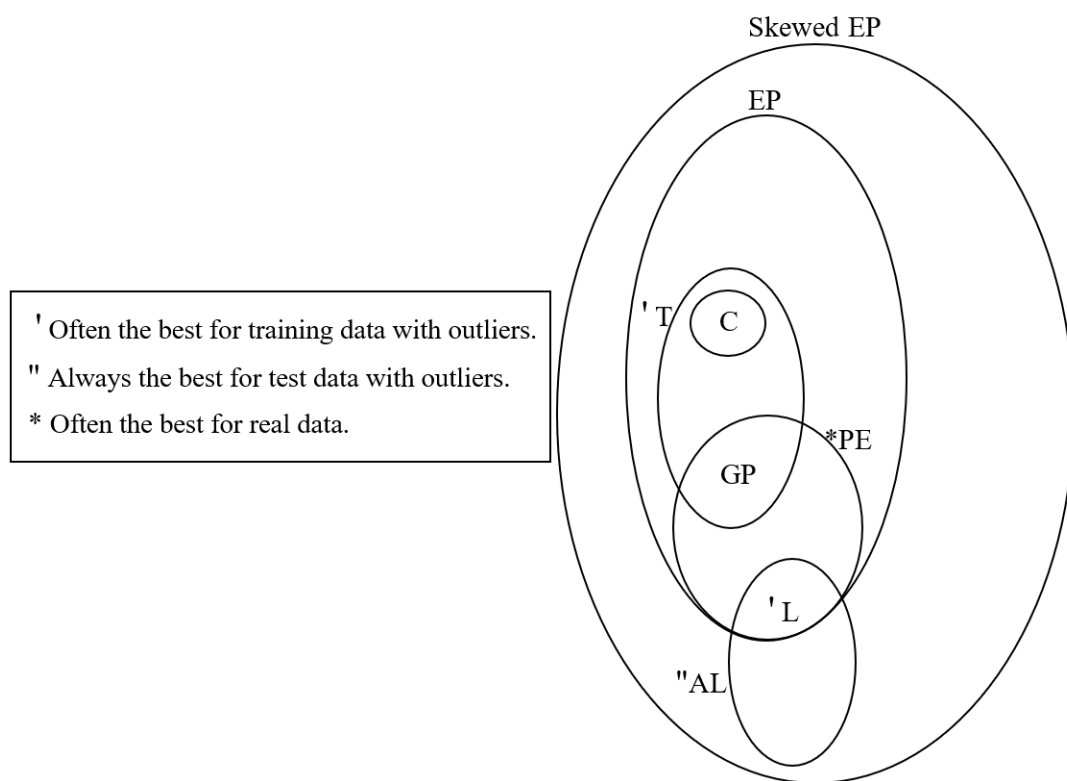


Figure 10: Diagram illustrating the relationships between various models and a general summary of the research findings. The figure highlights the clear superiority of certain non-Gaussian models in capturing diverse data patterns, extending beyond the limited framework of special cases like Gaussian or Cauchy distributions. More comprehensive models demonstrate greater adaptability to data diversity and robustness.

Future research could improve the EPR model's capacity to manage data abnormalities by substituting robust error distributions for conventional white noise. For example, including error distributions like Student-t or Laplace distributions, which have thicker tails or more flexibility, could improve resilience against moderate and extreme outliers. This modification would enhance forecast accuracy and model robustness in difficult real-world circumstances by better accommodating datasets with aberrant noise patterns.

It would also be beneficial to investigate skewness within elliptical processes in order to expand the existing fully Bayesian EPR model. Asymmetry in the data, which is frequently present in a variety of real-world applications, could be captured by the model by including skewed elliptical distributions, such as the skew-t or skew-PE processes. The impact of skewness on model performance, especially in a fully Bayesian framework, could be better understood by comparing regular EPR and skewed-EPR. By improving the model's adaptability and forecast precision in asymmetric data situations, this update may provide practitioners working with non-Gaussian, skewed data distributions with a more sophisticated tool.

References

- [1] O'Hagan, Anthony. 1978. "Curve Fitting and Optimal Design for Prediction." *Journal of the Royal Statistical Society: Series B (Methodological)* 40, no. 1: 1–24.
- [2] Neal, Radford M. 1995. *Bayesian Learning for Neural Networks*. PhD diss., Graduate Department of Computer Science, University of Toronto.
- [3] Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press. ISBN: 026218253X.
- [4] Neal, Radford M. 1998. "Regression and Classification Using Gaussian Process Priors." In *Bayesian Statistics 6*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 000–000. Oxford: Oxford University Press.
- [5] Benavoli, Alessio, Dario Azzimonti, and Dario Piga. 2020. "Skew Gaussian Processes for Classification." *Machine Learning* 109: 1877–1902.

- [6] Kuss, M., and C. E. Rasmussen. 2006. "Assessing Approximate Inference for Binary Gaussian Process Classification." *Journal of Machine Learning Research* 6 (Oct): 1679–1704.
- [7] Alodat, M. T., and K. M. Aludaat. 2008. "The Generalized Hyperbolic Process." *Brazilian Journal of Probability and Statistics* 22 (1): 1–8.
- [8] Vanhatalo, Jarno, Pasi Jylänki, and Aki Vehtari. 2009. "Gaussian Process Regression with Student-t Likelihood." *Advances in Neural Information Processing Systems* 22.
- [9] Alodat, M. T., and E. Y. Al-Momani. 2014. "Skew Gaussian Process for Nonlinear Regression." *Communications in Statistics - Theory and Methods* 43 (23): 4936–4961.
- [10] Alodat, M. T., and Mohammed K. Shakhathreh. 2020. "Gaussian Process Regression with Skewed Errors." *Journal of Computational and Applied Mathematics* 370: 112665.
- [11] Alodat, M. T., and M. Y. Al-Rawwash. 2014. "The Extended Skew Gaussian Process for Regression." *Metron* 72: 317–330.
- [12] Riihimäki, Jaakko, and Aki Vehtari. 2014. "Laplace Approximation for Logistic Gaussian Process Density Estimation and Regression." *Bayesian Analysis* 9 (2): 425–448.
- [13] Tang, Qingtao, Li Niu, Yisen Wang, Tao Dai, Wangpeng An, Jianfei Cai, and Shu-Tao Xia. 2017. "Student-t Process Regression with Student-t Likelihood." In *International Joint Conference on Artificial Intelligence 2017*, 2822–2828. Association for the Advancement of Artificial Intelligence (AAAI).
- [14] Tracey, Brendan D., and David Wolpert. 2018. "Upgrading from Gaussian Processes to Student's t Processes." In *2018 AIAA Non-Deterministic Approaches Conference*, 1659. American Institute of Aeronautics and Astronautics (AIAA).
- [15] Bänkestad, M., Sjölund, J., Taghia, J., and Schön, T. 2020. "The Elliptical Processes: A New Family of Flexible Stochastic Processes." *arXiv preprint arXiv:2001.XXXX*. Available at: https://www.researchgate.net/profile/Maria-Bankestad/publication/339972475_The_Elliptical_Processes_a_New_Family_of_

[Flexible_Stochastic_Processes/links/5e7b17b74585152fc0ec8c9e/
The-Elliptical-Processes-a-New-Family-of-Flexible-Stochastic-Processes.
pdf](https://flexible-stochastic-processes.com/links/5e7b17b74585152fc0ec8c9e/The-Elliptical-Processes-a-New-Family-of-Flexible-Stochastic-Processes.pdf)

- [16] Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.
- [17] Pu, Tong, Yiyang Zhang, and Chuancun Yin. 2024. "Generalized Location-Scale Mixtures of Elliptical Distributions: Definitions and Stochastic Comparisons." *Communications in Statistics - Theory and Methods* 53 (11): 3851–3875.
- [18] Fang, Kai Wang, Samuel Kotz, and Kai-Tai Ng. 2018. *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC.
- [19] Gómez Eusebio, Miguel A. Gómez-Villegas, and J. Miguel Marín. 2003. "A Survey on Continuous Elliptical Vector Distributions." *Revista Matemática Complutense* 16 (1): 345–361.
- [20] Szegő, Gábor. 1939. *Orthogonal Polynomials*. American Mathematical Society Colloquium Publications, Vol. 23. Revised 1959, 3rd edition 1967, 4th edition 1975.
- [21] Rahman, Sharif. 2017. "Wiener–Hermite Polynomial Expansion for Multivariate Gaussian Probability Measures." *Journal of Mathematical Analysis and Applications* 454 (1): 303–334.
- [22] Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd edition. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, London, New York, Washington, D.C.
- [23] Yu, Keming, and Jin Zhang. 2005. "A Three-Parameter Asymmetric Laplace Distribution and Its Extension." *Communications in Statistics - Theory and Methods* 34 (9–10): 1867–1879.
- [24] Gómez, Eusebio, M. A. Gómez-Villegas, and J. Miguel Marín. 1998. "A Multivariate Generalization of the Power Exponential Family of Distributions." *Communications in Statistics - Theory and Methods* 27 (3): 589–600.