future internet

MDPI

Article

Latent Autoregressive Student-*t* Prior Process Models to Assess Impact of Interventions in Time Series

Patrick Toman [1,2,*], Nalini Ravishanker [2], Nathan Lally [1] and Sanguthevar Rajasekaran [3]

[1] Hartford Steam Boiler, Hartford, CT 06106, USA; nathan_lally@hsb.com
[2] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA; nalini.ravishanker@uconn.edu
[3] Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA; sanguthevar.rajasekaran@uconn.edu
* Correspondence: patrick_toman3@hsb.com

Abstract: With the advent of the "Internet of Things" (IoT), insurers are increasingly leveraging remote sensor technology in the development of novel insurance products and risk management programs. For example, Hartford Steam Boiler's (HSB) IoT freeze loss program uses IoT temperature sensors to monitor indoor temperatures in locations at high risk of water-pipe burst (freeze loss) with the goal of reducing insurances losses via real-time monitoring of the temperature data streams. In the event these monitoring systems detect a potentially risky temperature environment, an alert is sent to the end-insured (business manager, tenant, maintenance staff, etc.), prompting them to take remedial action by raising temperatures. In the event that an alert is sent and freeze loss occurs, the firm is not liable for any damages incurred by the event. For the program to be effective, there must be a reliable method of verifying if customers took appropriate corrective action after receiving an alert. Due to the program's scale, direct follow up via text or phone calls is not possible for every alert event. In addition, direct feedback from customers is not necessarily reliable. In this paper, we propose the use of a non-linear, auto-regressive time series model, coupled with the time series intervention analysis method known as *causal impact*, to directly evaluate whether or not a customer took action directly from IoT temperature streams. Our method offers several distinct advantages over other methods as it is (a) readily scalable with continued program growth, (b) entirely automated, and (c) inherently less biased than human labelers or direct customer response. We demonstrate the efficacy of our method using a sample of actual freeze alert events from the freeze loss program.

Keywords: Gaussian process regression; Student-*t* process; impact of interventions; IoT sensor data; time series analysis; variational inference

Citation: Toman, P.; Ravishanker, N.; Lally, N.; Rajasekaran, S. Latent Autoregressive Student-*t* Prior Process Models to Assess Impact of Interventions in Time Series. *Future Internet* **2024**, *16*, 8. https://doi.org/10.3390/fi16010008

Academic Editors: Dionisis Kandris and Eleftherios Anastasiadis

Received: 10 November 2023
Revised: 19 December 2023
Accepted: 23 December 2023
Published: 28 December 2023

1. Introduction

With the advent of the "Internet of Things", it has become increasingly commonplace to develop real-time alerting and monitoring systems capable of mitigating the risk of mechanical failures via human intervention. Indeed, there are several challenges posed by such systems. First, the multiple time series generated in these scenarios often exhibit non-linear/non-Gaussian temporal dependencies. Next, in order for alerting mechanisms to be effective, there is a need to develop statistical methods capable of assessing the impact of exogenous interventions (alerts) in the form of spurring prompt human intervention. Finally, because there are likely thousands of such sensors involved in these types of IoT programs, any modeling paradigm must be readily scalable. In this paper, we propose a latent autoregressive Student-*t* process model to accomplish all three of these goals.

Intervention analysis is a well-established time series approach. An integral component of a successful intervention analysis is the use of a suitable time series model to learn the pre-intervention time series behavior. Linear models have traditionally been employed

for this purpose. These include Gaussian autoregressive moving average (ARIMA) models [1,2], Gaussian dynamic linear models (DLM) [3,4], or Bayesian structural time series (BSTS) models [5,6].

In the Bayesian framework, the pre-intervention model is used to derive the joint posterior predictive distribution of *post-intervention* observations. Samples from this posterior predictive distribution serve as *counterfactuals* to the post-intervention observations. By measuring the difference between these forecast values and the observed post-intervention data, a *semi-parametric* posterior estimate for the impact of the intervention is constructed. Due to its simplicity and versatility, this methodology described in [6] has been employed across a wide array of disciplines. For instance, Ref. [7] adapted the BSTS model to evaluate the impact of rebates for turf removal on water consumption across many households. In the public health context, Ref. [8] evaluated the impact of bariatric surgery (used for weight loss) on health care utilization in Germany. Another interesting example is given by [9], who used the impact framework in conjunction with a variety of climate time series to assess whether an anomalous climate change event can be credibly linked to the collapse of several Bronze age civilizations in the Mediterranean region.

For intervention impact analysis to be successful, it is critical that the underlying time series model adequately captures the pre-intervention time series dynamics. Traditional linear, Gaussian models can be inadequate for capturing the dynamics of time series that exhibit complex non-linear and/or long-term dependencies, and/or non-Gaussian behavior. As a consequence, the counterfactual forecasts may be inadequate to give a useful assessment of the impact. For example, multiple time series generated by "Internet of Things" (IoT) sensors often exhibit nonlinear temporal dependence that cannot be easily modeled by BSTS models. Successful intervention analysis of such time series requires sophisticated models of pre-intervention data such as those described in this paper.

For intervention impact analysis in multiple IoT time series, Ref. [10] proposed a Gaussian process (GP) prior regression model [11] with a covariance kernel tailored for these series as the underlying predictive model. This model is effective in that it can incorporate typical time series behavior such as seasonality and local linear trends but also non-linear time trends and dependencies between the target variable and exogenous predictors. While this model was demonstrated to be effective at capturing a wide array of time series dynamics, it does not directly incorporate information from past values of the time series. In addition, the GP prior can fail for time series that exhibit a heavy-tailed behavior.

With this in mind, we propose an extension to the latent Gaussian process time series model presented in [12] in which we replace the latent GP with that of a Student-t process (TP) [13,14] that we then use as the underlying model for a time series impact analysis using both simulated and real-world time series data from the IoT domain. In addition to going beyond the GP functional prior, our model had the added versatility in that it can accommodate arbitrary likelihoods, allowing for heavy-tailed observations to be modeled in a more robust way.

Note that because we require a model that allows for posterior sampling, mixture autoregressive models such as the one proposed in [15] are not suitable for solving our problem. In addition, the mixture autoregressive assumptions described in [15] may be unsuitable to describe the data generating process of IoT time series data.

The format of this paper is as follows: Section 2 describes the IoT temperature sensors and their associated data streams; Section 3 gives an overview of GP regression, including descriptions of existing GP regression models tailored for time series. Section 4 details the requisite background information regarding TP regression; we also introduce our autoregressive TP model. Section 5 compares the performance of our proposed model with existing methods on a time series intervention analysis problem. Section 6 summarizes our findings and discusses potential avenues for future research.

2. Background

One domain in which IoT sensor technology has been successfully deployed is the insurance context. For example, insurers have used IoT temperature sensors part of freeze loss prevention programs. The goal of these programs is to reduce insurance losses due to water-pipe burst (freeze loss) by providing temperature sensors to end users (insured property owners) to be installed in areas with a high risk of water pipe burst. In an ideal scenario, losses are prevented (or at least mitigated) by sending real-time alerts to customers to promptly take remedial action (i.e., raising temperatures to a safe level) in the event of dangerously low temperatures within the monitored space.

In this paper, we apply our methods to sensor temperature readings that are relayed in real time at a 15 min frequency. For each sensor stream, a decision rule algorithm combines information from (a) recent sensor readings and (b) outdoor temperatures from nearby weather stations to alert end users of potential imminent freeze loss. After receiving an alert, an end user is *expected* to take remedial action within 12 h of receiving the alert. Due to the program's scale, it is impossible to directly verify whether a customer took corrective action. Therefore, methods must be developed that can infer customer action only from the observed post-alert sensor streams themselves. To that end, we employ the *causal impact* methodology proposed by [6], which uses a counter-factual forecasting model to infer whether the alert system is effective in instigating customer action for a given alert event. More details, as well as an example of an alert event, can be found in Section 5.

3. Review of Gaussian Process Regression Models

We review the basic Gaussian process (GP) regression and its extensions for analyzing time series. Section 3.1 summarizes standard GP and regression techniques. Section 3.2 reviews the current literature on using nonlinear auto-regressions with exogenous predictors (NARX) in conjunction with GP regression (GP-NARX) models. In Section 3.3, we give a detailed review of the GP-RLARX model, a robust time series regression model that uses an auto-regressive latent state whose transition functions follow a GP prior, and the observations follow a normal distribution with time-varying scale.

3.1. GP and Sparse GP Regression

Gaussian processes (GP) are a set of methods that generalize the multivariate normal to infinite dimensions. Not only do GPs have a flexible non-parametric form, GP methods are also attractive because they offer principled uncertainty quantification via a predictive distribution. For supervised learning problems, $\mathcal{GP}$ prior models have the distinct advantage of allowing the user to automatically learn the correct functional form linking the input space $\mathcal{X}$ to the output space $\mathcal{Y}$. This is achieved by specifying a prior over the distribution of functions, which then allows the derivation of the posterior distribution over these functions once data have been observed. Throughout this paper, we use the notation $f(\cdot) \sim \mathcal{GP}(\mu(\cdot), k(\cdot,\cdot))$ to denote a generic GP prior with mean function $\mu(\cdot)$ and covariance function $k(\cdot,\cdot)$. Let $y_i \in \mathbb{R}$ denote an observed response and $\mathbf{x}_i$ and $\mathbf{x}_j$ be two distinct input vectors in $\mathbb{R}^p$. The GP regression model is defined as

$$y_i = f(\mathbf{x}_i) + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(0, \sigma_y^2), \tag{1}$$

$$f(\cdot) \sim \mathcal{GP}(\mu(\cdot), k(\cdot,\cdot)), \tag{2}$$

where the mean function and the covariance kernel are, respectively,

$$\mu_{\mathbf{x}_i} = E[f(\mathbf{x}_i)], \tag{3}$$

$$k_{\mathbf{x}_i,\mathbf{x}_j} = E[(f(\mathbf{x}_i) - \mu(\mathbf{x}_i))(f(\mathbf{x}_j) - \mu(\mathbf{x}_j))], \tag{4}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are fixed predictors, and $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$ is an n-dimensional vector. Given observed responses $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$, it follows that the Gaussian

process in (2) has a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_{\mathbf{x}_1}, \ldots, \mu_{\mathbf{x}_n})'$ and variance–covariance matrix $\mathbf{K_{xx}} = \{k_{\mathbf{x}_i,\mathbf{x}_j}\} \in \mathbb{R}^{n\times n}$.

Standard GP regression provides convenient closed forms for posterior inference. The posterior distribution $p(f(\mathbf{x}_i)|\mathbf{X}, \mathbf{y})$ is Gaussian with mean and variance, respectively, given by

$$\begin{aligned} \mu_i &= \mu_{\mathbf{x}_i} + \mathbf{k}_i' \left[\mathbf{K_{xx}} + \sigma^2 \mathbf{I}_n\right]^{-1} (\mathbf{y} - \boldsymbol{\mu}), \\ k_i &= k_{\mathbf{x}_i,\mathbf{x}_i} - \mathbf{k}_i' \left[\mathbf{K_{xx}} + \sigma^2 \mathbf{I}_n\right]^{-1} \mathbf{k}_i, \end{aligned} \tag{5}$$

where $\mathbf{k}_i = (k_{\mathbf{x}_i,\mathbf{x}_j}, j = 1, \ldots, n)'$.

Given a new set of inputs $\mathbf{X}_* \in \mathbb{R}^{p\times m}$, the joint distribution of the observed response $\mathbf{y}$ and the GP prior $f(\cdot)$ and the posterior predictive evaluated at new input set $\mathbf{X}_*$ are

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}_{n+m}\left(\begin{bmatrix} \boldsymbol{\mu}_\mathbf{f} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \mathbf{K_{xx}} + \sigma^2 \mathbf{I}_n & \mathbf{K}_{\mathbf{x},*} \\ \mathbf{K}_{*,\mathbf{x}} & \mathbf{K}_{*,*} \end{bmatrix} \right), \tag{6}$$

Thus, we have the posterior prediction density for $\mathbf{f}_*$ as

$$p(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}_m\left(\boldsymbol{\mu}_{*|\mathbf{x}}, \boldsymbol{\Sigma}_{*|\mathbf{x}}\right), \tag{7}$$

where

$$\boldsymbol{\mu}_{*|\mathbf{x}} = \boldsymbol{\mu}_* + \mathbf{K}_{*\mathbf{x}} \left[\mathbf{K_{xx}} + \sigma^2 \mathbf{I}_n\right]^{-1} (\mathbf{y} - \boldsymbol{\mu}_\mathbf{f}), \tag{8}$$

$$\boldsymbol{\Sigma}_{*|\mathbf{x}} = \mathbf{K}_{**} - \mathbf{K}_{*\mathbf{x}} \left[\mathbf{K_{xx}} + \sigma^2 \mathbf{I}_n\right]^{-1} \mathbf{K}_{\mathbf{x}*}. \tag{9}$$

One notable drawback of the GP model is its difficulty in scaling to large datasets due to inversion of the kernel covariance matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ (which has $\mathcal{O}(n^3)$ time complexity). *Sparse GP* methods [16–19] remedy this issue and reduce the computational cost of fitting GP models to long time series.

For $m << n$, they approximate the GP posterior in (5) by learning *inducing inputs* $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\} \in \mathcal{X}$, which lead to a finite set of *inducing variables* $\mathbf{U} = \{u_1, \ldots, u_m\}$ with $u_i = f(\mathbf{z}_i)$, where $f(\cdot)$ was defined in (2). Let $\mathbf{u} = (f(\mathbf{z}_1), \ldots, f(\mathbf{z}_m))'$. Their joint distribution is

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{u} \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu}_\mathbf{x} \\ \boldsymbol{\mu}_\mathbf{z} \end{pmatrix}, \begin{pmatrix} \mathbf{K_{xx}} & \mathbf{K_{xz}} \\ \mathbf{K_{zx}} & \mathbf{K_{zz}} \end{pmatrix} \right), \tag{10}$$

and using properties of the multivariate normal distribution,

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}_n\left(\boldsymbol{\mu}_\mathbf{x} + \mathbf{K_{xz}}\mathbf{K_{zz}^{-1}}(\mathbf{u} - \boldsymbol{\mu}_\mathbf{z}), \mathbf{K_{xx}} - \mathbf{K_{xz}}\mathbf{K_{zz}^{-1}}\mathbf{K_{zx}}\right), \tag{11}$$

$$p(\mathbf{u}) = \mathcal{N}_m(\boldsymbol{\mu}_\mathbf{z}, \mathbf{K_{zz}}). \tag{12}$$

The conditional distribution in (11) now only requires inversion of the $m \times m$ matrix $\mathbf{K_{zz}}$ instead of the $n \times n$ matrix $\mathbf{K_{xx}}$. The target is the n-dimensional marginal distribution of $\mathbf{f}$ given by

$$p(\mathbf{f}) = \int p(\mathbf{f}|\mathbf{u}) p(\mathbf{u}) d\mathbf{u}. \tag{13}$$

To facilitate this computation, we replace $p(\mathbf{u})$ given in (12) by its variational approximation

$$q(\mathbf{u}) = \mathcal{N}_m(\mathbf{m}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{zz}), \tag{14}$$

which in turn leads to approximating $p(\mathbf{f})$ by $q(\mathbf{f})$. Again, using properties of the multivariate normal distribution, $q(\mathbf{f})$ is given by

$$\begin{aligned} q(\mathbf{f}) &= \int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})d\mathbf{u}. \\ &= \mathcal{N}_n\Big(\boldsymbol{\mu}_\mathbf{x} + \mathbf{K}_{\mathbf{xz}}\mathbf{K}_{\mathbf{zz}}^{-1}(\mathbf{m}_\mathbf{z} - \boldsymbol{\mu}_\mathbf{z}), \mathbf{K}_{\mathbf{xx}} + \mathbf{K}_{\mathbf{xz}}\mathbf{K}_{\mathbf{zz}}^{-1}(\boldsymbol{\Sigma}_{\mathbf{zz}} - \mathbf{K}_{\mathbf{zz}})\mathbf{K}_{\mathbf{zz}}^{-1}\mathbf{K}_{\mathbf{zx}}\Big). \end{aligned} \tag{15}$$

Furthermore, given a new set of test inputs $\mathbf{X}_*$, the *approximate posterior predictive* density for $\mathbf{f}_*$ has form

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{y}) &\approx \int p(\mathbf{f}_*|\mathbf{f},\mathbf{u})p(\mathbf{f},\mathbf{u}|\mathbf{y})q(\mathbf{u})d\mathbf{f}d\mathbf{u} \\ &= \int p(\mathbf{f}_*|\mathbf{u})q(\mathbf{u})d\mathbf{u}. \end{aligned} \tag{16}$$

The integral in (16) is tractable and takes a form analogous to that in (15).

Given observed data $\mathbf{y} \in \mathbb{R}^n$, the variational inference approach for approximating the exact posterior of $\mathbf{f}$ in sparse GP regression reduces to minimizing the *evidence lower bound* (ELBO) [20]

$$\begin{aligned} \log p(\mathbf{y}) &\geq E_{q(\mathbf{u})}\Big[E_{p(\mathbf{f}|\mathbf{u})}\log p(\mathbf{y}|\mathbf{f})\Big] - \mathrm{KL}[q(\mathbf{u}) \parallel p(\mathbf{u})] \\ &= E_{q(\mathbf{f})}[\log p(\mathbf{y}|\mathbf{f})] - \mathrm{KL}[q(\mathbf{u}) \parallel p(\mathbf{u})], \end{aligned} \tag{17}$$

where $q(\mathbf{f})$ is defined as in (15). For more details on the ELBO optimization procedure, refer to Section 4 of [19].

3.2. GP-NARX Models

Quite often, we seek to model time series data as a function of exogenous inputs and an autoregressive function of past observations. A class of GP models incorporating both non-linear autoregressive structure and exogenous predictors (typically abbreviated as GP-NARX) offer a principled way to propagate uncertainty when forecasting.

An early example comes from [21], who proposed a GP-NARX model in which the inputs at time t consist of the past L lags of the response time series y_t, as well as available exogenous inputs $\mathbf{c}_t \in \mathbb{R}^{n_c}$. The input vector at time t in the GP-NARX model is the tuple $(\mathbf{x}_t', \mathbf{c}_t')'$, where $\mathbf{x}_t = (y_{t-1}, \ldots, y_{t-L})'$. Since $y_{t-1}, \ldots, y_{t-L}$ are known during the training phase of model fitting, estimation is performed using maximum likelihood or *maximum a posteriori* methods.

Although training the GP-NARX model is similar to training the GP regression model and is straightforward, predicting future values is more challenging. Suppose our goal is to generate *k-step ahead* forecasts for future responses $y_{T+1}, \ldots, y_{T+k}$ given training data $\mathbf{y} = (y_1, \ldots, y_T)' \in \mathbb{R}^T$. Because all or part of $\mathbf{x}_t$ is unobserved in the holdout period (since it involves $y_{T+j-1}, \ldots, y_{T+j-L}$, $j = 1, \ldots, k$) and is an uncertain input during forecasting, direct application of (8) would fail to take into account this inherent uncertainty.

Ref. [21] deals with the uncertain inputs issue by assuming that for each $j = 1, \ldots, k$, $\mathbf{x}_{T+j} \sim \mathcal{N}_L\Big(\boldsymbol{\mu}_{\mathbf{x}_{T+j}}, \boldsymbol{\Sigma}_{\mathbf{x}_{T+j}}\Big)$. Then, given the training data $\mathcal{D} = \{\mathbf{c}_t, \mathbf{x}_t, y_t\}_{t=L_x+1}^T$, and a set of exogenous inputs $\mathbf{c}_{T+j}$, $j = 1, \ldots, k$, the posterior predictive distribution for y_{T+j} is

$$p(f_{T+j}) = \int p(f_{T+j}|\mathbf{x}_{T+j}, \mathbf{w}_{T+j}, \mathcal{D})p(\mathbf{x}_{T+j})d\mathbf{x}_{T+j}. \tag{18}$$

Although there is no closed form for (18), the moments of the posterior predictive distribution can be obtained via Monte Carlo sampling or one of several different approximation methods [22].

Recently, these ideas have been extended to sparse GP models. Ref. [23] developed an approximate uncertainty propagation approach to be used alongside the sparse pseudo-

input GP regression method, known as the Fully Independent Sparse Training Conditional (FITC) model [16]. Ref. [24] derived uncertainty propagation methods for a wide variety of competing sparse GP methods, and [25] extended sparse GP-NARX time series modeling to an online setting.

3.3. GP-RLARX Models

Ref. [12] proposed an alternative to the GP-NARX model. Their GP-RLARX model assumes a *latent* autoregressive structure for the lagged inputs, leading to the description below:

$$y_t = x_t + \varepsilon_t^{(y)} \tag{19a}$$
$$x_t = f(x_{t-1}, \ldots, x_{t-L_x}, c_{t-1}, \ldots, c_{t-L_c}) + \varepsilon_t^{(x)} \tag{19b}$$
$$f(\cdot) = \mathcal{GP}(\mu(\cdot), k(\cdot,\cdot)) \tag{19c}$$
$$\varepsilon_t^{(x)} \sim \mathcal{N}(\varepsilon_t^{(x)}|0, \sigma_x^2) \tag{19d}$$
$$\varepsilon_t^{(y)} \sim \mathcal{N}\left(\varepsilon_t^{(y)}|0, \tau_t\right) \tag{19e}$$
$$\tau_t \sim \mathcal{IG}(\tau_t|\alpha, \beta). \tag{19f}$$

where $c_{t-1}, \ldots, c_{t-L_c}$ are lagged exogenous inputs with maximum lag L_c, and $x_{t-1}, \ldots, x_{t-L_x}$ are the lagged latent states with maximal lag L_x.

This framework is reminiscent of a state-space model in which (19a) denotes the observation equation at time t and (19b) is the corresponding state equation, where x_t is an autoregressive function of the preceding L_x lags of the latent state.

To facilitate inference in the GP-RLARX model, Ref. [12] used a sparse variational approximation similar to that described in Section 3.1, where $\mathbf{u} \in \mathbb{R}^m$ are inducing points generated by evaluating the GP prior over *pseudo-inputs* $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$, $\mathbf{z}_i \in \mathbb{R}^{L_x+L_c}, i = 1, \ldots, m$. It follows that $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K_{zz}})$, where $\mathbf{K_{zz}}$ denotes the kernel covariance matrix evaluated over the *pseudo-inputs* $\mathbf{Z}$. Then, the GP-RLARX hierarchical model takes the form

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K_{zz}}) \tag{20a}$$
$$p(f_t|\mathbf{u}, \mathbf{x}) = \mathcal{N}(f_t|[\mathbf{a_x}]_t, [\mathbf{\Sigma_{xx}}]_{tt}) \tag{20b}$$
$$p(x_t) = \mathcal{N}(x_t|\mu_t, \lambda_t),\ \forall\, t \in \{1, \ldots, L_x\} \text{ (initial state)} \tag{20c}$$
$$p(x_t|f_t) = \mathcal{N}\left(x_t|f_t, \sigma_x^2\right),\ \forall\, t \in \{L_x+1, \ldots, T\} \tag{20d}$$
$$p(\tau_t) = \mathcal{IG}(\tau_t|\alpha, \beta),\ \forall\, t \in \{L_x+1, \ldots, T\} \tag{20e}$$
$$p(y_t|x_t, \tau_t) = \mathcal{N}(y_t|x_t, \tau_t),\ \forall\, t \in \{L_x+1, \ldots, T\}, \tag{20f}$$

where $\mathbf{a_x} = \mathbf{K_{xz}}\mathbf{K_{zz}}^{-1}\mathbf{u}$, $\mathbf{\Sigma_{xx}} = \mathbf{K_{xx}} - \mathbf{K_{xz}}\mathbf{K_{zz}}^{-1}\mathbf{K_{zx}}$, and $f_t = f(x_{t-1}, \ldots, x_{t-L_x}, w_{t-1}, \ldots, w_{t-L_w})$, with $f(\cdot) \sim \mathcal{GP}(0, k(\cdot,\cdot))$. For brevity, we denote $\tilde{\mathbf{x}}_t = (x_{t-1}, \ldots, x_{t-L_x}, w_{t-1}, \ldots, w_{t-L_w})'$ in the remainder of the paper. The joint distribution is succinctly expressed as

$$p(\mathbf{y}, \mathbf{x}, \mathbf{f}, \mathbf{u}, \boldsymbol{\tau}) = \left(\prod_{t=L_x+1}^{T} p(y_t|x_t, \tau_t)p(x_t|f_t)p(\tau_t)p(f_t|\mathbf{u}, \tilde{\mathbf{x}}_t)p(\mathbf{u})\right)\prod_{t=1}^{L_x} p(x_t). \tag{21}$$

Ref. [12] used a variational inference approach [20] to estimate the latent variables, adopting the variational approximation

$$q(\mathbf{x}, \boldsymbol{\tau}, \mathbf{f}, \mathbf{u}) = \left[\prod_{t=1}^{T} q(x_t)\right]\left[\prod_{t=L_x+1}^{T} q(\tau_t)\right]\left[\prod_{t=L_x+1}^{T} p(f_t|\mathbf{u}, \tilde{\mathbf{x}}_t)\right]q(\mathbf{u}), \tag{22}$$

where $q(x_t) = \mathcal{N}(x_t|\mu_t^{(x)}, \lambda_t^{(x)})$, $q(\tau_t) = \mathcal{IG}(\tau_t|a_t, b_t)$, $p(f_t|\mathbf{u}, \tilde{\mathbf{x}}_t) = \mathcal{N}(\mathbf{f}|[\mathbf{a}_\mathbf{x}]_t, [\mathbf{\Sigma}_{\mathbf{xx}}]_{tt})$, and $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}_\mathbf{z}, \mathbf{\Sigma}_{\mathbf{zz}})$. In this framework, $\mu_t^{(x)}, \lambda_t^{(x)}, \mathbf{m}_\mathbf{z}, \mathbf{\Sigma}_{\mathbf{zz}}, a_t, b_t$ are variational parameters that are optimized according to a variational inference strategy, similar to that found in [26]. We refer readers to Section 5.1 of [12] for more details, including the exact expression of the ELBO.

Table 1 summarizes the basic characteristics of the GP-NARX and GP-RLARX models as well as their main pros and cons.

Table 1. Comparison of GP-NARX and GP-RLARX methods.

Model	Characteristics	Pros	Cons
GP-NARX	1. Target is a non-linear, autoregressive function of observed past values and exogenous predictors. 2. Trained via Type II MLE 3. Forecasts attained via simple Monte Carlo sampling	1. Fast to train 2. Non-parametric GP prior 3. Predictive uncertainty	1. Incapable of handling heavy-tailed noise outliers 2. Assumes a Gaussian likelihood
GP-RLARX	1. Target variable is assumed to equal a latent state plus noise 2. Autoregressive behavior captured through latent state dynamics 3. Exogenous predictors can be placed at observed and latent level 4. Trained using variational Bayesian and sparse GP methods 5. Forecasts by sampling from approximate posterior	1. Robust to heavy-tailed noise and outliers 2. Non-parametric sparse GP prior 3. Predictive uncertainty 4. Arbitrary likelihoods	1. Slower to train than GP-NARX 2. Somewhat more challenging to train

4. Proposed Methods: Autoregressive TP Models

Recently, there has been growing interest in extending Gaussian process models to other types of elliptical process models, with particular emphasis on Student-t process models (TP) [13,14]. In this section, we present extensions to both the GP-NARX and GP-RLARX models by replacing the GP functional prior by a Student-t process prior. Section 4.1 gives an overview of the Student-t process as well as a recently developed method for sparse Student-t processes. Next, Section 4.2 describes the TP-NARX model as an extension of the GP-NARX model. Finally, Section 4.3 gives details of the proposed extension of the GP-RLARX model to the TP-RLARX model. To the authors' best knowledge, there has been no research on the development or implementation of a NARX model or RLARX model using TP priors. These are useful additions to the literature, and they are discussed in the following sections.

4.1. Review and Notation for Student-t Processes

We say that $\mathbf{f} \in \mathbb{R}^n$ follows a *multivariate Student's-t* distribution with degrees of freedom $v \in \mathbb{R}^+$, location $\boldsymbol{\mu} \in \mathbb{R}^n$, and positive definite scale matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ if and only if it has the following density:

$$p(\mathbf{f}) = \frac{\Gamma((v+n)/2)}{(v\pi)^{n/2}\Gamma(v/2)|\mathbf{K}|^{1/2}} \left(1 + \frac{1}{v}(\mathbf{f}-\boldsymbol{\mu})'\mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\mu})\right)^{-\frac{v+n}{2}}, \tag{23}$$

which can be written succinctly as $\mathbf{f} \sim \mathcal{T}_n(v, \boldsymbol{\mu}, \mathbf{K})$. Now, suppose that we have $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{f}_* \in \mathbb{R}^m$ with joint density

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{T}_{m+n}\left(v, \begin{bmatrix} \boldsymbol{\mu}_\mathbf{f} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mathbf{ff}} & \mathbf{K}_{\mathbf{f}*} \\ \mathbf{K}_{*\mathbf{f}} & \mathbf{K}_{**} \end{bmatrix}\right). \tag{24}$$

By properties of the multivariate Student-t distribution, we have

$$\mathbf{f} \sim \mathcal{T}_n(v, \boldsymbol{\mu}_\mathbf{f}, \mathbf{K}_{\mathbf{ff}}) \text{ and } \mathbf{f}_*|\mathbf{f} \sim \mathcal{T}_m\left(v+n, \tilde{\boldsymbol{\mu}}_*, \frac{v+\beta-2}{v+n-2}\tilde{\mathbf{K}}_{**}\right) \tag{25}$$

where

$$\tilde{\boldsymbol{\mu}}_* = \boldsymbol{\mu}_* + \mathbf{K}_{*\mathbf{f}}\mathbf{K}_{\mathbf{ff}}^{-1}(\mathbf{f}-\boldsymbol{\mu}_\mathbf{f}),\ \beta = (\mathbf{f}-\boldsymbol{\mu}_\mathbf{f})'\mathbf{K}_{\mathbf{ff}}^{-1}(\mathbf{f}-\boldsymbol{\mu}_\mathbf{f}), \tilde{\mathbf{K}}_{**} = \mathbf{K}_{**} - \mathbf{K}_{*\mathbf{f}}\mathbf{K}_{\mathbf{ff}}^{-1}\mathbf{K}_{\mathbf{f}*}. \tag{26}$$

Finally, we say that $f(\cdot)$ follows a Student-t process on $\mathcal{X}$, denoted $\mathcal{TP}(v, \mu(\cdot), k(\cdot,\cdot))$, where $v > 2$ denotes the degrees of freedom, $\mu(\cdot) \in \mathbb{R}$ denotes the mean function, and $k(\cdot,\cdot) \in \mathbb{R}$ is the covariance function, if for any finite collection of function values, we have $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))' \sim \mathcal{T}_n(v, \boldsymbol{\mu}, \mathbf{K})$.

While less popular than GP models, Student-t processes have still been employed in a number of contexts. For instance, Ref. [27] proposed an online time series anomaly detection algorithm that employs TP regression to simultaneously learn time series dynamics in the presence of heavy-tailed noise and identify anomalous events. Another example comes from [28], in which the authors proposed a Student-t process latent variable model with the goal of identifying a low-dimensional set of latent factors capable of explaining variation among non-Gaussian financial time series. Ref. [29] employed Student-t processes in the development of degradation models used to analyze the lifetime reliability of manufactured products.

Recently, Ref. [30] proposed a variational inference approach for sparse Student-t processes, similar to the sparse GP methods described in Section 3.1. Suppose that we have $r \sim \mathcal{IG}(\alpha, \beta)$. Now, if we let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_m] \in \mathcal{X}$ with $m << n$ denote a set of *inducing inputs*, then we can define a corresponding *inducing variables* $\mathbf{u}|r = [f(\mathbf{z}_1), \ldots, f(\mathbf{z}_m)] \sim \mathcal{N}_m(\mathbf{0}, r\mathbf{K}_{\mathbf{zz}})$. It follows that the joint density of $\mathbf{f}, \mathbf{u}, r$ is

$$\begin{aligned} p(\mathbf{f}, \mathbf{f}_\mathbf{z}, r) &= p(\mathbf{f}|\mathbf{u}, r)p(\mathbf{u}|r)p(r) \\ &= \mathcal{N}_n\left(\mathbf{K}_{\mathbf{xz}}\mathbf{K}_{\mathbf{zz}}^{-1}\mathbf{u}, r(\mathbf{K}_{\mathbf{xx}} - \mathbf{K}_{xz}\mathbf{K}_{\mathbf{zz}}^{-1}\mathbf{K}_{\mathbf{zx}})\right)\mathcal{N}_m(\mathbf{0}, r\mathbf{K}_{\mathbf{zz}})\mathcal{IG}(\alpha, \beta). \end{aligned} \tag{27}$$

The goal is to develop an approximate distribution $q(\mathbf{f}, \mathbf{u}, r)$ capable of accurately approximating $p(\mathbf{f}, \mathbf{u}, r)$. Ref. [30] proposed the following variational distribution:

$$\begin{aligned} q(\mathbf{f}, \mathbf{u}, r) &= p(\mathbf{f}|\mathbf{u}, r)q(\mathbf{u}|r)q(r) \\ &= \mathcal{N}_n\left(\mathbf{K}_{\mathbf{xz}}\mathbf{K}_{\mathbf{zz}}^{-1}\mathbf{u}, r(\mathbf{K}_{\mathbf{xx}} - \mathbf{K}_{\mathbf{xz}}\mathbf{K}_{\mathbf{zz}}^{-1}\mathbf{K}_{\mathbf{zx}})\right)\mathcal{N}_m(\mathbf{m}_\mathbf{z}, r\boldsymbol{\Sigma}_{\mathbf{zz}})\mathcal{IG}(a, b). \end{aligned} \tag{28}$$

It follows that the *evidence lower bound* (ELBO) is

$$\log p(\mathbf{y}) \geq E_{q(\mathbf{f},\mathbf{u},r)}[\log p(\mathbf{y}|\mathbf{f}, \mathbf{u}, r)] - \mathrm{KL}[q(\mathbf{f}, \mathbf{u}, r) \parallel p(\mathbf{f}, \mathbf{u}, r)] \tag{29}$$

where $\mathrm{KL}(\cdot)$ denotes the KL divergence between the respective joint densities. The KL term can be re-expressed as

$$\mathrm{KL}(q(\mathbf{f}, \mathbf{u}, r) \parallel p(\mathbf{f}, \mathbf{u}, r)) = \int\int q(\mathbf{u}, r) \log\left[\frac{q(\mathbf{u}, r)}{p(\mathbf{u}, r)}\right] d\mathbf{u}dr \tag{30}$$

since the $p(\mathbf{f}, \mathbf{u}, r)$ terms are canceled out. Furthermore, we can evaluate the likelihood component as

$$\begin{aligned} E_{q(\mathbf{f},\mathbf{u},r)}[\log p(\mathbf{y}|\mathbf{f},\mathbf{u},r)] &= \int\int\int q(\mathbf{f},\mathbf{u},r)\log p(\mathbf{y}|\mathbf{f},\mathbf{u},r)d\mathbf{f}d\mathbf{u}dr \\ &= \int q(\mathbf{f})\log p(\mathbf{y}|\mathbf{f})d\mathbf{f} \end{aligned} \tag{31}$$

where

$$q(\mathbf{f}) = \int\int p(\mathbf{f}|\mathbf{u},r)q(\mathbf{u}|r)q(r)d\mathbf{u}dr$$

which can be expressed as

$$q(\mathbf{f}) = \mathcal{T}_n\left(2a, \mathbf{K_{xz}}\mathbf{K_{zz}^{-1}}\mathbf{m_z}, \frac{b}{a}(\mathbf{K_{xx}} - \mathbf{K_{xz}}\mathbf{K_{zz}^{-1}}\mathbf{K_{zx}} + \mathbf{K_{xz}}\mathbf{K_{zz}^{-1}}\mathbf{\Sigma}\mathbf{K_{zz}^{-1}}\mathbf{K_{zx}})\right). \tag{32}$$

Assuming that we have a set of test inputs $\mathbf{X}_*$, we can attain the *approximate predictive distribution* for $\mathbf{f}_*$ as

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{y}) &= \int\int p(\mathbf{f}_*|\mathbf{u},r)q(\mathbf{u}|r)q(r)d\mathbf{u}dr \\ &= \int\int \mathcal{N}_n\left(\mathbf{K_{*z}}\mathbf{K_{zz}^{-1}}\mathbf{u}, r(\mathbf{K_{**}} - \mathbf{K_{*z}}\mathbf{K_{zz}^{-1}}\mathbf{K_{z*}})\right)\mathcal{N}_m(\mathbf{m_z}, r\mathbf{\Sigma_{zz}})\mathcal{IG}(a,b)d\mathbf{u}dr \\ &= \mathcal{T}_n\left(2a, \mathbf{K_{*z}}\mathbf{K_{zz}^{-1}}\mathbf{m_z}, \frac{b}{a}(\mathbf{K_{**}} - \mathbf{K_{*z}}\mathbf{K_{zz}^{-1}}\mathbf{K_{z*}} + \mathbf{K_{*z}}\mathbf{K_{zz}^{-1}}\mathbf{\Sigma_{zz}}\mathbf{K_{zz}^{-1}}\mathbf{K_{z*}})\right) \end{aligned}$$

which, as we can see, is structurally quite similar to its sparse GP counterpart described in (16).

4.2. TP-NARX Model

Our first proposed model is the TP-NARX model, which is a straightforward extension of the GP-NARX model (see Section 3.2) obtained by replacing the GP functional prior with that of a *t*-process prior defined in Section 4.1. Further, during the forecasting phase, rather than assuming $\mathbf{x}_{T+j}$ from (18) follows an approximately multivariate normal distribution, we assume instead that $\mathbf{x}_{T+j} \overset{approx.}{\sim} \mathcal{T}_L\left(v, \boldsymbol{\mu}_{\mathbf{x}_{T+j}}, \mathbf{\Sigma}_{\mathbf{x}_{T+j}}\right)$, where $v > 2$ denotes the degrees of freedom for the multivariate Student-*t* distribution. A Monte Carlo sampling approach is used to approximate the integral in (18).

4.3. TP-RLARX Model

For our second proposed model, we extend the GP-RLARX model by replacing the Gaussian process prior with a Student-*t* process prior. Similar to the GP-RLARX model's sparse approximation approach, we employ a sparse variational Student-*t* process (SVTP) framework presented in [30] to act as the functional prior over our state transition. Therefore, the TP-RLARX generative model is

$$p(r) = \mathcal{IG}(r|\alpha,\beta) \tag{33a}$$
$$p(\mathbf{u}|r) = \mathcal{N}_m(\mathbf{u}|\mathbf{0}, r\mathbf{K_{zz}}) \tag{33b}$$
$$p(f_t|\mathbf{u},\mathbf{x},r) = \mathcal{N}(f_t|[\mathbf{a_x}]_t, [r\mathbf{\Sigma_{xx}}]_{tt}) \tag{33c}$$
$$p(x_t) = \mathcal{N}(x_t|\mu_t,\lambda_t), \forall\, t \in \{1,\ldots,L_x\} \tag{33d}$$
$$p(x_t|f_t) = \mathcal{N}(x_t|f_t,\sigma_x^2), \forall\, t \in \{L_x+1,\ldots,T\} \tag{33e}$$
$$p(\tau_t) = \mathcal{IG}(\tau_t|\kappa,\theta), \forall\, t \in \{L_x+1,\ldots,T\} \tag{33f}$$
$$p(y_t|x_t,\tau) = \mathcal{N}(y_t|x_t,\tau_t), \tag{33g}$$

where $\mathbf{a_x}$ and $\mathbf{\Sigma_{xx}}$ are the same as in Section 3.3, whereas $f(\cdot)$ is now marginally distributed as $\mathcal{TP}\left(2\alpha, 0, \frac{\beta}{\alpha}k(\cdot,\cdot)\right)$. We employ a variational inference approach to approximate the generative model described in (33). The variational distribution has form

$$q(\mathbf{f}, \mathbf{u}, \mathbf{x}, \boldsymbol{\tau}, r) = q(\mathbf{x})q(\boldsymbol{\tau}) \prod_{t=L_x+1}^{T} p(f_t|\mathbf{u}, r, \tilde{\mathbf{x}}_t)q(\mathbf{u}|r)q(r), \tag{34}$$

where each term is identical to that found in Section 3.3 with the exception of $q(\tau_t) = \mathcal{IG}(a_t, b_t)$, $q(\mathbf{u}|r) = \mathcal{N}_m(\mathbf{m_z}, r\mathbf{\Sigma_{zz}})$, $\prod_{t=L_x+1}^{T} p(f_t|\mathbf{u}, r, \tilde{\mathbf{x}}_t) = \prod_{t=L_x+1}^{T} \mathcal{N}(f_t|[\mathbf{a_x}]_t, r[\mathbf{\Sigma_{xx}}]_{tt})$, and the additional variational distribution $q(r) = \mathcal{IG}(r|\gamma, \sigma)$. The *evidence lower bound* (ELBO) for this model takes form

$$\begin{aligned}
\log p(\mathbf{y}) &\geq E_{q(\tau,\mathbf{x},\mathbf{f},\mathbf{u},r)}\left[\log \frac{p(\mathbf{y}, \boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)}{q(\tau, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)}\right] \\
&= E_{q(\boldsymbol{\tau},\mathbf{x},\mathbf{f},\mathbf{u},r)}[\log p(\mathbf{y}|\boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)] \\
&+ E_{q(\boldsymbol{\tau},\mathbf{x},\mathbf{f},\mathbf{u},r)}\left[\log \frac{p(\boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)}{q(\boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)}\right] \\
&= E_{q(\boldsymbol{\tau},\mathbf{x},\mathbf{f},\mathbf{u},r)}[\log p(\mathbf{y}|\boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)] \\
&- \mathrm{KL}[q(\boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r) \parallel p(\boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)] \\
&= E_{q(\boldsymbol{\tau},\mathbf{x},\mathbf{f},\mathbf{u},r)}[\log p(\mathbf{y}|\boldsymbol{\tau}, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)] \\
&- \mathrm{KL}[q(\boldsymbol{\tau}) \parallel p(\boldsymbol{\tau})] \\
&- \mathrm{KL}[q(\mathbf{x}) \parallel p(\mathbf{x}|\mathbf{f})] \\
&- \mathrm{KL}[q(\mathbf{u}, r) \parallel p(\mathbf{u}, r)].
\end{aligned} \tag{35}$$

With the exception of the additional scale parameter for the Student-t process, the derivation of the ELBO terms follows similarly to [12]. For the likelihood term, we have

$$\begin{aligned}
E_{q(\cdot)}[\log p(\mathbf{y}|\tau, \mathbf{x}, \mathbf{f}, \mathbf{u}, r)] &= \sum_{t=L_x+1}^{T} E_{q(x_t,\tau)}[\log p(y_t|x_t, \tau)] \\
&= \sum_{t=L_x+1}^{T} E_{q(x_t,\tau)}\left[\frac{1}{2}\left(-\log 2\pi - \log \tau - \frac{(y_t - x_t)^2}{\tau}\right)\right] \\
&= \sum_{t=L_x+1}^{T} E_{q(x_t)}\left[\frac{1}{2}\left(-\log 2\pi + E_{q(\tau)}[-\log \tau] - E_{q(\tau)}\left[\frac{(y_t - x_t)^2}{2\tau}\right]\right)\right] \\
&\propto \sum_{t=L_x+1}^{T}\left[a_t + \log(b_t\Gamma(a_t)) - (a_t + 1)\psi(a_t) - \frac{a_t}{b_t}E_{q(x_t)}\left[(y_t - x_t)^2\right]\right] \\
&\propto \sum_{t=L_x+1}^{T}\left[a_t + \log(b_t\Gamma(a_t)) - (a_t + 1)\psi(a_t) - \frac{a_t}{b_t}\left(y_t^2 - 2\mu_t^{(x)}y_t + \lambda_t^{(x)} + (\mu_t^{(x)})^2\right)\right]
\end{aligned} \tag{36}$$

where $\psi(\cdot)$ denotes the digamma function. Next, for the KL divergence between $q(\boldsymbol{\tau})$ and $p(\boldsymbol{\tau})$, we have

$$\mathrm{KL}[q(\boldsymbol{\tau}) \parallel p(\boldsymbol{\tau})] = \sum_{t=1}^{T}\left[(a_t - \kappa)\psi(a_t) - \log\Gamma(a_t) + \log\Gamma(\kappa) + \kappa\log\frac{b_t}{\theta} - \log\theta) + a_t\left(\frac{\theta - b_t}{b_t}\right)\right]. \tag{37}$$

For the KL divergence of the latent states, we have

$$\begin{aligned}
\mathrm{KL}[q(\mathbf{x}) \parallel p(\mathbf{x}|\mathbf{f}))] &= \sum_{t=1}^{L_x} \mathrm{KL}[q(x_t) \parallel p(x_t))] + \sum_{t=L_x+1}^{T} \mathrm{KL}[q(x_t) \parallel p(x_t|f_t)] \\
&= \frac{1}{2} \sum_{t=1}^{L_x} \left[\frac{\lambda_t^{(x)} + (\mu_t - \mu_t^{(x)})^2}{\lambda_t} - \log \frac{\lambda_t^{(x)}}{\lambda_t} - 1 \right] + \sum_{t=L_x+1}^{T} E_{q(x_t,\tau_t,f_t,\mathbf{u},r)} \left[\log \frac{q(x_t)}{p(x_t|ft)} \right] \\
&= \frac{1}{2} \sum_{t=1}^{L_x} \left[\frac{\lambda_t^{(x)} + (\mu_t - \mu_t^{(x)})^2}{\lambda_t} - \log \frac{\lambda_t^{(x)}}{\lambda_t} - 1 \right] \\
&+ \sum_{t=L_x+1}^{T} \frac{1}{2} \left[\log 2\pi\lambda_t^{(x)} + 1 \right] \\
&- \sum_{t=L_x+1}^{T} \int q(\mathbf{x}) q(\mathbf{u}, r) p(f_t|\tilde{\mathbf{x}}_t, \mathbf{u}, r) \log p(x_t|f_t) d\mathbf{x} d\mathbf{f} d\mathbf{u} dr. \qquad (38)
\end{aligned}$$

Closed forms are not available for the third term in $\mathrm{KL}[q(\mathbf{x}) \parallel p(\mathbf{x}|\mathbf{f})]$ for most kernel configurations; therefore, we apply employ a *black box variational inference* procedure as described in [31].

Finally, from [30], we have

$$\begin{aligned}
\mathrm{KL}[q(\mathbf{u}, r) \parallel p(\mathbf{u}, r)] &= \int \int q(\mathbf{u}, r) \log \left[\frac{q(\mathbf{u}, r)}{p(\mathbf{u}, r)} \right] d\mathbf{u} dr \\
&= \left(\frac{\gamma}{2\sigma} \right) \mathbf{m}_{\mathbf{z}}' \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}_{\mathbf{z}} + \frac{1}{2} Tr\left(\mathbf{K}_{\mathbf{zz}}^{-1} \boldsymbol{\Sigma}_{\mathbf{zz}} \right) + \frac{1}{2} \log \frac{|\mathbf{K}_{\mathbf{zz}}|}{|\boldsymbol{\Sigma}_{\mathbf{zz}}|} - \frac{m}{2} + \alpha \log \frac{\sigma}{\beta} \\
&- \log \frac{\Gamma(\gamma)}{\Gamma(\alpha)} + (\gamma - \alpha)\psi(\gamma) + (\beta - \sigma)\frac{\gamma}{\sigma}. \qquad (39)
\end{aligned}$$

where $Tr(\cdot)$ denotes the matrix trace. Model fitting is done using the Python library **Pyro** [32], which is dedicated to probabilistic programming with a particular emphasis on BBVI and SVI methods.

Table 2 summarizes our contributions discussed in this section as well as our findings that are more thoroughly presented in Sections 5.2 and 5.3.

Table 2. Summary of contributions and findings.

Proposed Contributions	Summary	Findings
TP-NARX	Extends the GP-NARX model to the Student-*t* likelihood in order to accommodate heavy-tailed noise and outliers	1. Gain robustness of a heavy-tailed likelihood without increasing computational speed relative to GP-NARX
TP-RLARX	Extends the GP-RLARX by substituting the latent GP prior with a Student-*t* process prior. Proposed method is now robust to heavy-tailed noise at the observational and latent levels. Derived the ELBO for this proposed model	1. Gain robustness of a heavy-tailed latent state with minor increase to computational speed relative to GP-RLARX 2. TP-RLARX has performance at least as good as GP-RLARX on intervention analysis task.

5. Application: IoT Temperature Time Series

We apply the TP-NARX and TP-RLARX models to perform impact analysis on temperature time series. First, Section 5.1 describes the IoT sensor data and the objective of the intervention impact analysis. Section 5.2 shows the performance for each model

on a number of different forecasting metrics and shows some example forecasts. Finally, Section 5.3 presents detailed results from the impact analysis and a thorough interpretation of the findings. We compare the TP-NARX and RLARX approaches with their Gaussian process counterparts.

5.1. Data Description

We analyze data spanning the time period from 1 October 2020 to 25 February 2021 on a sample of $N = 50$ sensors that are distributed across the contiguous US, with a concentration in the Upper Midwest, Southeast, and the East coast. A sensor measures the internal room temperature at 15 min intervals. Figure 1 shows an example of a sensor's temperature stream and when an alert was sent (the alert time is denoted by the vertical black line). We see that, just prior to the alert, internal temperatures plunge, while the external temperature is at a low level.

In order for this program to be effective, it is imperative that there is accurate information on whether a customer actually takes meaningful action in a timely manner after receiving an alert in order to avoid freeze loss. Although ideally, this information would be directly obtained from the customer, this is rarely possible in practice. As such, the effectiveness of the alert must be ascertained purely based on the observed pre-alert and post-alert time series. We refer to this analysis as intervention impact analysis. Essentially, a customer action has likely occurred if there is a large increase in post-alert temperatures that are incongruous with forecasts generated by a suitable time series model trained on pre-alert temperatures.

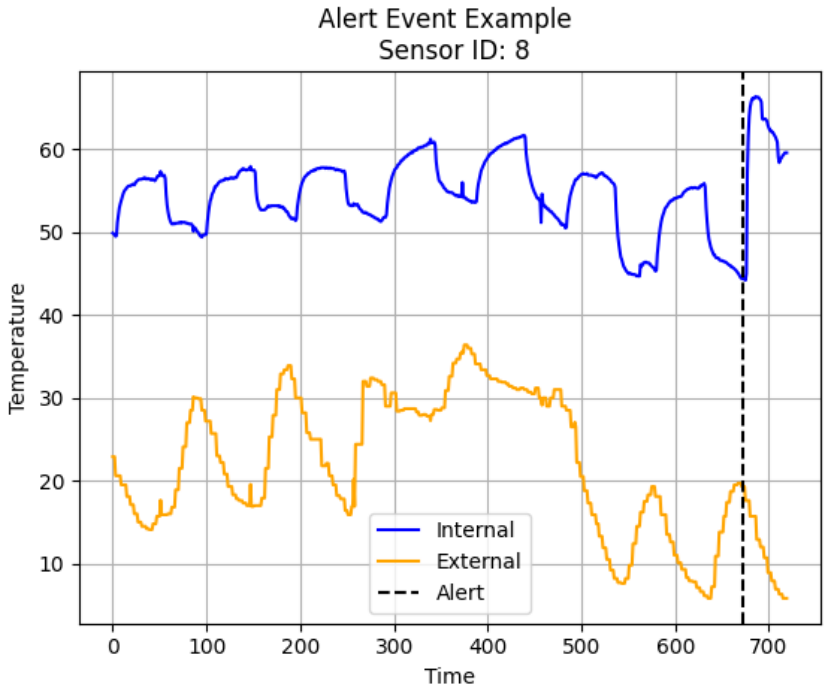

Figure 1. Example of an alert event. The black dotted line denotes the time of an alert.

Since many of the IoT temperature time series exhibit non-linear behavior, methods such as Bayesian structural time series are inadequate for intervention impact analysis. Alternatively, the NARX and RLARX models are capable of learning non-linear behavior directly from the data, thus, we will use them in substitution of BSTS models for our impact analysis. Performance will be compared to both GP-NARX and GP-RLARX models.

5.2. Results

Once again, we apply the TP-NARX and TP-RLARX models to each alert event in our dataset and then compare the results to the GP-NARX and GP-RLARX models. For each model, we use one of four covariance kernels: radial basis function (RBF), Matérn

3/2, Matérn 5/2, or Ornstein–Uhlenbeck (OU). *Outdoor (external) temperature* is the only exogenous predictor used in the experiment.

Figures 2 and 3 depict the forecasts for the GP-RLARX and TP-RLARX models with RBF kernel on the same alert event. As expected, the point estimates (red lines) are similar for both models; however, the predictive interval (red shaded areas) for TP-RLARX is slightly wider. For the results depicted in Figures 2 and 3, the average difference between the 0.025 quantile and 0.975 quantile for GPRLARX is ≈ 11.76, whereas the average difference between the 0.025 quantile and 0.975 quantile for TPRLARX is ≈ 15.703. The wider predictive interval of TP-RLARX means that our decision on whether a customer has taken appreciable action will be more conservative. Firms wishing to be more conservative in assessing customer behavior or those with noisier time series data might prefer the TP-RLARX model.

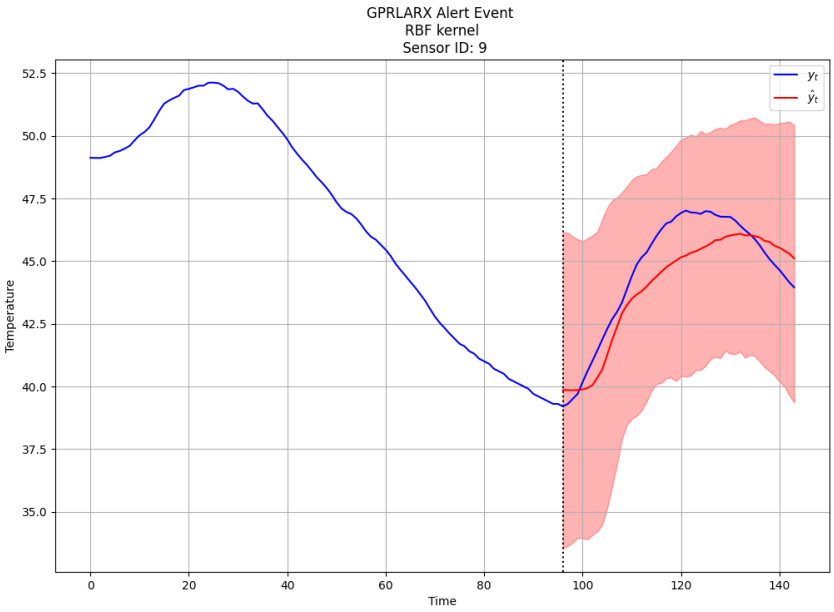

Figure 2. Forecasts for GP-RLARX.

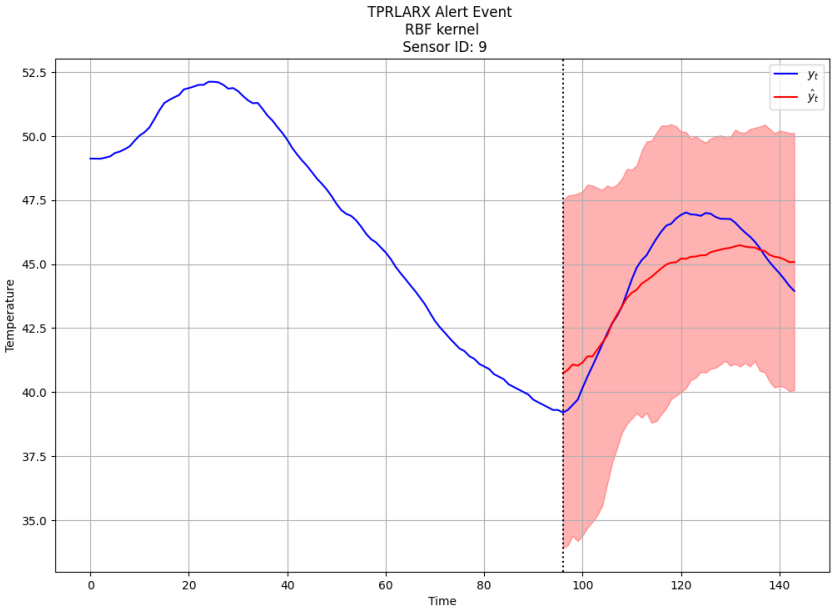

Figure 3. Forecasts for TP-RLARX.

Furthermore, Table 3 shows the root mean squared error (RMSE), symmetric mean absolute percentage error (sMAPE), and continuous ranked probability score (CRPS) [33]

for each combination of model and kernel. Each metric is calculated by averaging the metrics for each alert event within each model combination. In addition, Table 3 also gives CPU times for each model. Overall, we find that the TP-RLARX model using the Ornstein–Uhlenbeck kernel provides the best average RMSE, followed by the TP-NARX using the Matérn 5/2 kernel. Furthermore, we find that the GP-RLARX models give considerably worse performance than both TP-NARX and TP-RLARX models, regardless of kernel.

Table 3. Forecast metrics and CPU times.

Model	RMSE	sMAPE	CRPS	CPU Time
GP-NARX RBF	13.456	0.046	10.705	33.69
GP-NARX Matérn 3/2	13.605	0.046	10.882	34.88
GP-NARX Matérn 5/2	13.669	0.047	10.875	34.63
GP-NARX OU	13.385	0.046	10.717	32.51
GP-RLARX RBF	15.748	0.051	12.253	741.50
GP-RLARX Matérn 3/2	15.610	0.051	12.429	889.78
GP-RLARX Matérn 5/2	15.453	0.051	12.309	954.87
GP-RLARX OU	14.831	0.049	11.798	907.22
TP-NARX RBF	13.110	0.044	10.340	31.95
TP-NARX Matérn 3/2	13.234	0.046	10.361	30.72
TP-NARX Matérn 5/2	13.073	0.046	10.361	31.27
TP-NARX OU	13.728	0.047	10.707	28.82
TP-RLARX RBF	13.666	0.046	10.886	967.20
TP-RLARX Matérn 3/2	13.149	0.046	10.481	1131.45
TP-RLARX Matérn 5/2	13.312	0.046	10.574	1129.89
TP-RLARX OU	13.003	0.045	10.394	1104.85

5.3. Intervention Impact Analysis and Interpretation

For each alert event in this experiment, we are given a label indicating whether a panel of domain experts believed the customer took corrective action based on visual inspection of time series plots similar to Figure 1. If a majority of experts thought that action had been taken, an alert was labeled as "Action", indicating that appreciable customer action likely occurred; otherwise, it was labeled as "No Action". In the absence of observed labels, this is the closest approximation to the ground truth that we have and constitutes the benchmark we will compare against.

Due to this inherently biased labeling scheme, the labels we have are more aptly described as "pseudo-labels" in that they do not represent an objective truth. For alerts that experts labeled as "No Action", we find that there is a high degree of correspondence between the model and expert labels, as indicated in Table 4. This result is unsurprising, as it is quite obvious to both experts and the models when no action has been taken because the observed internal temperature will remain flat or even decrease after the alert. Conversely, for instances labeled "Action" by experts, every model is likely to disagree, as shown in Table 4. This is attributable to the fact that whenever post-alert temperatures experience a sharp positive increase, human labelers are biased towards labeling it an action, regardless of the historical time series behavior or its correlation with the exogenous predictor. For example, Figure 4 shows an alert event labeled as "Action" by domain experts; however, there is clearly a strong, positive correlation between the internal and external temperatures that appears to instigate the increase in post-alert internal temperature. Furthermore, the post-alert increase in the response variable is quite modest and is congruent with pre-intervention temperature levels. Unsurprisingly, every possible combination of model and kernel tested returns a decision of "No Action" for this alert.

Indeed, the results of the impact analysis are congruent with our goals in that the models are far more conservative in assessing customer intervention. Expert opinion is that customers typically do not take action, so it is desirable to have models that require a large shift in post-alert behavior in order to declare an alert event as having been addressed. To

that end, the RLARX models yield the best results in that they are both highly unwilling to assume an intervention has been successful without significant evidence.

Table 4. Confusion matrices for each model (RBF kernel).

(**a**) GP-NARX			(**b**) GP-RLARX		
	Human Labels			Human Labels	
Predicted Labels	No Action	Action	Predicted Labels	No Action	Action
No Action	33	7	No Action	38	10
Action	7	3	Action	2	0
(**c**) TP-NARX			(**d**) TP-RLARX		
	Human Labels			Human Labels	
Predicted Labels	No Action	Action	Predicted Labels	No Action	Action
No Action	24	8	No Action	39	9
Action	16	2	Action	1	1

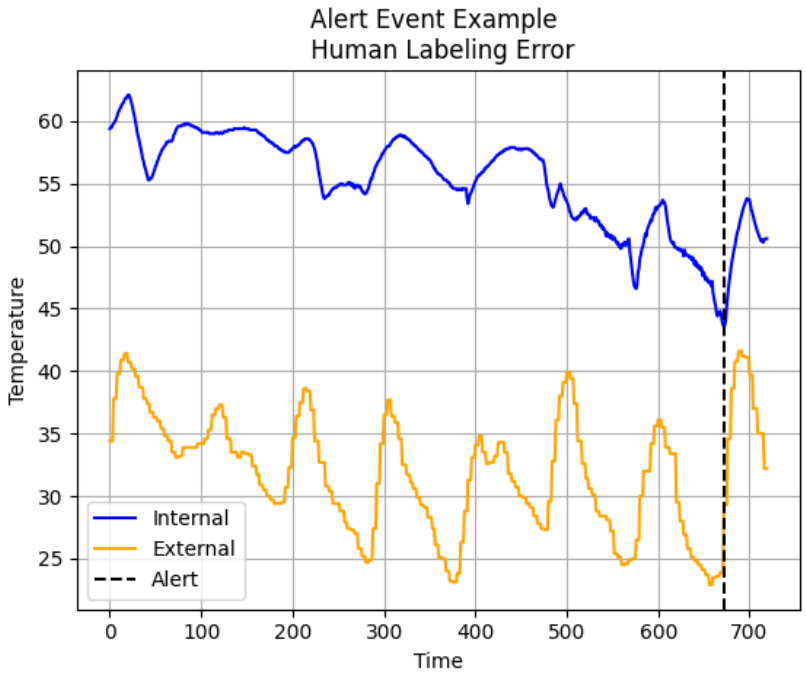

Figure 4. Alert event mislabeled by human labelers as "Action".

6. Conclusions

In this paper, we have proposed extensions to both the GP-NARX and GP-RLARX models by replacing the GP functional prior with a Student-*t* process prior. The goal is to use these models as underlying forecasting models for intervention impact analysis of IoT temperature data streams. We have demonstrated that the TP-NARX and TP-RLARX models provide improved forecasting accuracy relative to the GP-NARX and GP-RLARX models. Furthermore, we have shown that the TP-RLARX model has the desirable trait of being more conservative, relative to both the GP models and human labelers, in declaring that an intervention was effective in instigating appreciable customer action in Section 5. As such, the TP-RLARX model is preferable in impact analyses where the ground truth is not necessarily known and there is a high cost associated with false positives.

The analysis performed here opens several avenues for future research. First, it would be interesting to apply the same Student-*t* process extension to Gaussian process state space models, such as those presented in [34–36], and compare their performance with models

presented in this research. Furthermore, a comparison of our model with *parametric* non-linear time series models, such as the *deep state space* framework proposed in [37], would also be a worthwhile endeavor. The authors intend to explore these ideas in future research.

Author Contributions: Conceptualization, P.T., N.R. and N.L.; methodology, P.T., N.R. and N.L.; writing, reviewing and editing, P.T., N.R. and S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors gratefully acknolwedge the support provided by Hartford Steam Boiler in providing the IoT sensor data used in this paper.

Conflicts of Interest: Patrick Toman and Nathan Lally were employed by the company Hartford Steam Boiler. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Brockwell, P.J.; Davis, R.A. *Introduction to Time Series and Forecasting*; Springer: Berlin/Heidelberg, Germany, 2016.
2. Abraham, B. Intervention analysis and multiple time series. *Biometrika* **1980**, *67*, 73–78. [CrossRef]
3. Shumway, R.H.; Stoffer, D.S.; Stoffer, D.S. *Time Series Analysis and its Applications*; Springer: Berlin/Heidelberg, Germany, 2000; Volume 3.
4. Van den Brakel, J.; Roels, J. Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *Ann. Appl. Stat.* **2010**, *4*, 1105–1138. [CrossRef]
5. Scott, S.L.; Varian, H.R. Predicting the present with Bayesian structural time series. *Int. J. Math. Model. Numer. Optim.* **2014**, *5*, 4–23. [CrossRef]
6. Brodersen, K.H.; Gallusser, F.; Koehler, J.; Remy, N.; Scott, S.L. Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* **2015**, *9*, 247–274. [CrossRef]
7. Schmitt, E.; Tull, C.; Atwater, P. Extending Bayesian structural time-series estimates of causal impact to many-household conservation initiatives. *Ann. Appl. Stat.* **2018**, *12*, 2517–2539. [CrossRef]
8. Kurz, C.F.; Rehm, M.; Holle, R.; Teuner, C.; Laxy, M.; Schwarzkopf, L. The effect of bariatric surgery on health care costs: A synthetic control approach using Bayesian structural time series. *Health Econ.* **2019**, *28*, 1293–1307. [CrossRef]
9. Ön, Z.B.; Greaves, A.; Akçer-Ön, S.; Özeren, M.S. A Bayesian test for the 4.2 ka BP abrupt climatic change event in southeast Europe and southwest Asia using structural time series analysis of paleoclimate data. *Clim. Chang.* **2021**, *165*, 1–19. [CrossRef]
10. Toman, P.; Soliman, A.; Ravishanker, N.; Rajasekaran, S.; Lally, N.; D'Addeo, H. Understanding insured behavior through causal analysis of IoT streams. In Proceedings of the 2023 6th International Conference on Data Mining and Knowledge Discovery (DMKD 2023), Chongqing, China, 24–26 June 2023.
11. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2.
12. Mattos, C.L.C.; Damianou, A.; Barreto, G.A.; Lawrence, N.D. Latent autoregressive Gaussian processes models for robust system identification. *IFAC-PapersOnLine* **2016**, *49*, 1121–1126. [CrossRef]
13. Shah, A.; Wilson, A.; Ghahramani, Z. Student-t processes as alternatives to Gaussian processes. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Reykjavik, Iceland, 22–25 April 2014; pp. 877–885.
14. Solin, A.; Särkkä, S. State space methods for efficient inference in Student-t process regression. In Proceedings of the Artificial Intelligence and Statistics, PMLR, San Diego, CA, USA, 9–12 May 2015; pp. 885–893.
15. Meitz, M.; Preve, D.; Saikkonen, P. A mixture autoregressive model based on Student'st—Distribution. *Commun.-Stat.-Theory Methods* **2023**, *52*, 499–515. [CrossRef]
16. Snelson, E.; Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005; Volume 18.
17. Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Clearwater Beach, FL, USA, 16–18 April 2009; pp. 567–574.
18. Hensman, J.; Fusi, N.; Lawrence, N.D. Gaussian processes for big data. *arXiv* **2013**, arXiv:1309.6835.
19. Hensman, J.; Matthews, A.; Ghahramani, Z. Scalable variational Gaussian process classification. In Proceedings of the Artificial Intelligence and Statistics, PMLR, San Diego, CA, USA, 9–12 May 2015; pp. 351–360.
20. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
21. Girard, A.; Rasmussen, C.E.; Quinonero-Candela, J.; Murray-Smith, R.; Winther, O.; Larsen, J. Multiple-step ahead prediction for non linear dynamic systems–a Gaussian process treatment with propagation of the uncertainty. *Adv. Neural Inf. Process. Syst.* **2002**, *15*, 529–536.

22. Girard, A. *Approximate Methods for Propagation of Uncertainty with Gaussian Process Models*; University of Glasgow (United Kingdom): Glasgow, UK, 2004.
23. Groot, P.; Lucas, P.; Bosch, P. Multiple-step time series forecasting with sparse gaussian processes. In Proceedings of the 23rd Benelux Conference on Artificial Intelligence, Gent, Belgium, 3–4 November 2011.
24. Gutjahr, T.; Ulmer, H.; Ament, C. Sparse Gaussian processes with uncertain inputs for multi-step ahead prediction. *IFAC Proc. Vol.* **2012**, *45*, 107–112. [CrossRef]
25. Bijl, H.; Schön, T.B.; van Wingerden, J.W.; Verhaegen, M. System identification through online sparse Gaussian process regression with input noise. *IFAC J. Syst. Control* **2017**, *2*, 1–11. [CrossRef]
26. Titsias, M.; Lawrence, N.D. Bayesian Gaussian process latent variable model. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; JMLR Workshop and Conference Proceedings; pp. 844–851.
27. Xu, Z.; Kersting, K.; Von Ritter, L. Stochastic Online Anomaly Analysis for Streaming Time Series. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; pp. 3189–3195.
28. Uchiyama, Y.; Nakagawa, K. TPLVM: Portfolio Construction by Student'st-Process Latent Variable Model. *Mathematics* **2020**, *8*, 449. [CrossRef]
29. Peng, C.Y.; Cheng, Y.S. Student-t processes for degradation analysis. *Technometrics* **2020**, *62*, 223–235. [CrossRef]
30. Lee, H.; Yun, E.; Yang, H.; Lee, J. Scale mixtures of neural network Gaussian processes. *arXiv* **2021**, arXiv:2107.01408.
31. Ranganath, R.; Gerrish, S.; Blei, D. Black box variational inference. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Reykjavik, Iceland, 22–25 April 2014; pp. 814–822.
32. Bingham, E.; Chen, J.P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; Goodman, N.D. Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.* **2018**, *20*, 973–978.
33. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
34. Frigola, R.; Chen, Y.; Rasmussen, C.E. Variational Gaussian process state-space models. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27.
35. Doerr, A.; Daniel, C.; Schiegg, M.; Duy, N.T.; Schaal, S.; Toussaint, M.; Sebastian, T. Probabilistic recurrent state-space models. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 1280–1289.
36. Curi, S.; Melchior, S.; Berkenkamp, F.; Krause, A. Structured variational inference in partially observable unstable Gaussian process state space models. In Proceedings of the Learning for Dynamics and Control, PMLR, Berkeley, CA, USA, 11–12 June 2020; pp. 147–157.
37. Krishnan, R.; Shalit, U.; Sontag, D. Structured inference networks for nonlinear state space models. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.