ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-82

Student-t Process Regression with Dependent Student-t Noise

Qingtao Tang[1] and **Yisen Wang** and **Shu-Tao Xia**

Abstract. Gaussian Process Regression (GPR) is a powerful non-parametric method. However, GPR may perform poorly if the data are contaminated by outliers. To address the issue, we replace the Gaussian process with a Student-t process and introduce dependent Student-t noise in this paper, leading to a Student-t Process Regression with Dependent Student-t noise model (TPRD). Closed form expressions for the marginal likelihood and predictive distribution of TPRD are derived. Besides, TPRD gives a probabilistic interpretation to the Student-t Process Regression with the noise incorporated into its Kernel (TPRK), which is a common approach for the Student-t process regression. Moreover, we analyze the influence of different kernels. If the kernel meets a condition, called β-*property* here, the maximum marginal likelihood estimation of TPRD's hyperparameters is independent of the degrees of freedom ν of the Student-t process, which implies that GPR, TPRD and TPRK have exactly the same predictive mean. Empirically, the degrees of freedom ν could be regarded as a convergence accelerator, indicating that TPRD with a suitable ν performs faster than GPR. If the kernel does not have the β-property, TPRD has better performances than GPR, without additional computational cost. On benchmark datasets, the proposed results are verified.

1 INTRODUCTION

Gaussian processes are powerful Bayesian nonparametric methods with good interpretability and non-parametric flexibility. In addition, Gaussian processes have simple learning, exact inference and impressive empirical performances without manual parameter tuning [12].

In a regression problem, the basic model is $\mathbf{y} = f(X) + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is the target vector, X is the feature matrix and $\boldsymbol{\epsilon}$ is the noise. Gaussian Process Regression (GPR) assumes the latent function f is a Gaussian process and $\boldsymbol{\epsilon}$ is independent and identically distributed (i.i.d.) Gaussian noise. Based on these assumptions, exact inference can be performed by the Bayes' theorem [12]. However, GPR performs poorly on data sets contaminated by outliers because of the thin-tailed property of Gaussian distribution. To address the issue, heavy-tailed distributions, e.g., the Student-t distribution, have been introduced into GPR. Generally speaking, there are two ways. The first way assumes that the noise is from an i.i.d Student-t distribution. Then a Gaussian process with the i.i.d Student-t noise is obtained. Exact inference, however, is analytically intractable. Then one has to turn to approximate inference methods, such as MCMC (Markov Chain Monte Carlo, [10]), variational approximation [6] and Laplace approximation [16]. However, these methods require additional computational cost.

The second way assumes that the latent function f is a Student-t process, which leads to the Student-t Process Regression model (TPR) [12, 14]. The problem of this way is that the sum of two independent Student-t distributions or the sum of a Student-t and a Gaussian distribution is analytically intractable. In other words, the Student-t process regression with independent Gaussian or Student-t noise is analytically intractable. Thus, Rasmussen and Williams [12] said "Allowing for independent noise contributions removes analytic tractability, which may reduce the usefulness of the t process ". Later in [14, 15, 19], in order to increase the usefulness of TPR, the noise is incorporated to the kernel function, which leads to the Student-t process regression with the noise incorporated into its Kernel (TPRK). By this method, good empirical performances are achieved, however, probabilistic properties of the noise remain unknown.

In this paper, to obtain a model with robustness, reasonable computational cost and probabilistic interpretation, we propose a Student-t Process Regression with Dependent Student-t noise model (TPRD), which replaces the Gaussian process in GPR with a Student-t process and introduces dependent Student-t noise. The variance of the noise is dependent on how well the noise-free model fits the data. Owing to the novel noise, TPRD owns all the advantages of GPR, such as good interpretation, exact inference and simple hyperparameter learning. Besides, the marginal likelihood of TPRD is equivalent to that of TPRK, which indicates that TPRD gives a probabilistic interpretation to TPRK. Moreover, if the kernel has the β-property (defined later), we prove that the maximum marginal likelihood (ML) estimation of TPRD's hyperparameters is independent of the degrees of freedom ν, resulting in that TPRD, TPRK and GPR have the same predictive mean. And TPRD outperforms GPR without additional computational cost if the kernel does not have the β-property. Various experiments are conducted to evaluate the properties mentioned above.

In summary, the main contributions of this paper are as follows:

- We prove that GPR is a special case of TPRD. And closed form expressions for the marginal likelihood and predictive distribution of TPRD are derived.
- TPRD gives a probabilistic interpretation to the way of incorporating the noise into the kernel function, adopted by TPRK.
- If the kernel has the β-property, we prove that the ML estimation of TPRD's hyperparameters is independent of the degrees of freedom ν, and GPR, TPRD, TPRK have exactly the same predictive mean. But experiments show that TPRD with a suitable ν is faster than GPR.
- If the kernel does not have the β-property, empirically, TPRD obtains better performances at no additional computational cost over GPR .

[1] Tsinghua University, Beijing, China, email: tqt15@mails.tsinghua.edu.cn

The rest of the paper is organized as follows: Section 2 describes GPR and proposes TPRD. Section 3 analyzes the theoretical properties of TPRD and TPRK. Section 4 presents the experimental results. Section 5 concludes the work.

2 TPR WITH DEPENDENT STUDENT-T NOISE

In this section, we will give a brief review to GPR and propose TPRD. Besides, we also discuss the relationships between TPRD, GPR and TPRK.

2.1 Review of GPR

In a regression problem, we have a training set $\mathcal{D}$ of n instances, $\mathcal{D} = \{X, \mathbf{y}\}$, where $X = \{\mathbf{x}_i\}_{i=1}^n$ is the $n \times D$ design matrix with D being the dimension of attributes, and $\mathbf{y} = \{y_i\}_{i=1}^n$ denotes the output or target vector of dimension n. In GPR, the basic model is

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

where ϵ_i is the i.i.d Gaussian noise. The latent function f is given a Gaussian process prior. In practice, as n is finite, $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^n$ has a multivariate Gaussian distribution as

$$p(\mathbf{f}|X, K_g) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_g, K_g), \tag{2}$$

where $\boldsymbol{\mu}_g$ is the mean. The subscript g indicates that the hyperparameters are of GPR. Usually, for notational simplicity, we assume $\boldsymbol{\mu}_g = \mathbf{0}$. And K_g is the covariance matrix. $(K_g)_{i,j} = cov(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$, where k is a kernel function, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_l)$ is the parameters of the kernel and l is the number of parameters. As ϵ_i is i.i.d Gaussian, the likelihood is

$$p(\mathbf{y}|\mathbf{f}, \sigma_g) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_g^2 I), \tag{3}$$

where σ_g^2 is the variance of the noise and I denotes the identity matrix. By the Bayes' theorem and integrating out $\mathbf{f}$, we can get the marginal likelihood

$$p(\mathbf{y}|X, \sigma_g, K_g) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \Sigma_g), \tag{4}$$

where $\Sigma_g = K_g + \sigma_g^2 I$. Then, to learn the hyperparameters σ_g and $\boldsymbol{\theta}$, the maximum marginal likelihood can be used, which is equivalent to minimizing the negative logarithm marginal likelihood denoted by

$$-\ln p(\mathbf{y}|X, \sigma_g, K_g) = \frac{1}{2}\mathbf{y}^T \Sigma_g^{-1} \mathbf{y} + \frac{1}{2}\ln|\Sigma_g| + \frac{n}{2}\ln 2\pi. \tag{5}$$

The three terms of the negative marginal likelihood in Eq. (5) are explained in [12]: the only term involving the observed targets is the data-fit term $\frac{1}{2}\mathbf{y}^T\Sigma_g^{-1}\mathbf{y}$; $\ln|\Sigma_g|$ is the complexity penalty depending only on the kernel function and the inputs; and $\frac{n}{2}\ln 2\pi$ is a normalization constant.

After learning the hyperparameters σ_g and $\boldsymbol{\theta}$, for a known input $\mathbf{x}_* \in R^D$, the predictive distribution is given as follows [12]

$$\begin{aligned} p(y_*|\boldsymbol{y}) &= \mathcal{N}(y_*|\mu_*, \sigma_*), & (6)\\ \mu_* &= \mathbf{k}_*^T \Sigma_g^{-1} \mathbf{y}, & (7)\\ \sigma_*^2 &= k(\mathbf{x}_*, \mathbf{x}_*; \boldsymbol{\theta}) - \mathbf{k}_*^T \Sigma_g^{-1} \mathbf{k}_*, & (8)\\ \mathbf{k}_* &= \{k(\mathbf{x}_i, \mathbf{x}_*; \boldsymbol{\theta})\}_{i=1}^n. & (9) \end{aligned}$$

Clearly, the predictive mean of GPR is a linear combination of $y_i (i = 1, 2, \ldots, n)$ and the predictive variance does not depend on $\mathbf{y}$.

2.2 Derivation of TPRD

The definition of a multivariate Student-t distribution is as follows.

Definition 1. An n-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is said to have the n-variate Student-t distribution with degrees of freedom ν, mean vector $\boldsymbol{\mu}$, and correlation matrix R if its joint probability density function (PDF) is given by

$$\begin{aligned} St(\mathbf{x}|\nu, \boldsymbol{\mu}, R) = & \frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|R|^{1/2}} \\ & \cdot \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} R^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\nu+n)/2}. \end{aligned} \tag{10}$$

We adopt the most common definition of Student-t distribution here, which could be found in [5, 8].

Now we introduce the Student-t process regression with dependent Student-t noise. In the regression problem Eq. (1), the latent function f is given a Student-t process prior, i.e., $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^n$ has the PDF

$$\begin{aligned} p(\mathbf{f}|X, \boldsymbol{\theta})) =& St(\mathbf{f}|\nu, \mathbf{0}, K_t) \\ =& \frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|K_t|^{1/2}} \left[1 + \frac{1}{\nu}\mathbf{f}^{\mathrm{T}} {K_t}^{-1} \mathbf{f}\right]^{-(\nu+n)/2}. \end{aligned} \tag{11}$$

Just as in GPR,

$$(K_t)_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}), \tag{12}$$

and we also assume the mean vector is $\mathbf{0}$ for simplicity.

Assume the noise $\boldsymbol{\epsilon}$ is an n-dimensional Student-t distribution dependent on $p(\mathbf{f}|X, \boldsymbol{\theta})$ with the following form

$$p(\boldsymbol{\epsilon}|\beta) = St\left(\boldsymbol{\epsilon}\middle|\nu+n, \mathbf{0}, \left(1 + \frac{1}{\nu}\mathbf{f}^T K_t^{-1}\mathbf{f}\right)\frac{1}{\beta}I\right). \tag{13}$$

For a given ν, $\frac{1}{\nu}\mathbf{f}^T K_t^{-1}\mathbf{f}$ is the data-fit term without considering noise by Eq. (1) and the explanation of the first term of Eq. (5). The variance of the noise depends on how well the noise-free model fits the data. To be specific, if the noise-free model fits the data well, the negative logarithm marginal likelihood is small, which implies the variance in the Eq. (13) is small. Otherwise, if the noise-free model does not fit the model well, the variance in the Eq. (13) is relatively large. And the degrees of freedom of $\boldsymbol{\epsilon}$ is $n + \nu$, larger than the degrees of freedom of $\mathbf{f}$ by n. As in practice, the number of instances n is relatively large, $\boldsymbol{\epsilon}$ approximates to a multivariate Gaussian distribution with a diagonal covariance matrix, which implies that the noise $\epsilon_i (i = 1, 2, \ldots, n)$ approximate to i.i.d Gaussian distributions. In summary, TPRD is in effect a Student-t process with approximate i.i.d Gaussian noise whose variance is adjusted to the data-fit term of noise-free model. Since the Student-t process is more robust than the Gaussian process, it is expected that TPRD performs better than GPR in some cases.

With the assumptions of Eq. (11) and Eq. (13), exact inference is achieved as follows,

$$\begin{aligned} p(\mathbf{y}|\mathbf{f}, \beta) =& St\left(\mathbf{y}\middle|\nu+n, \mathbf{f}, \left(1 + \frac{1}{\nu}\mathbf{f}^T {K_t}^{-1}\mathbf{f}\right)\frac{1}{\beta}I\right) \\ =& \frac{\Gamma[(\nu+2n)/2]\beta^{n/2}}{\Gamma[(\nu+n)/2](\nu+n)^{n/2}\pi^{n/2}\left(1 + \frac{1}{\nu}\mathbf{f}^T {K_t}^{-1}\mathbf{f}\right)^{n/2}} \\ & \cdot \left[1 + \frac{\beta}{(\nu+n)}\frac{(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f})}{1 + \frac{1}{\nu}\mathbf{f}^T {K_t}^{-1}\mathbf{f}}\right]^{-(\nu+2n)/2}. \end{aligned} \tag{14}$$

Multiplying Eq. (11) by Eq. (14), the joint distribution of $\mathbf{y}$ and $\mathbf{f}$ is

$$\begin{aligned} p\left(\mathbf{y},\mathbf{f}|X,\boldsymbol{\theta},\beta\right) \propto & \left[1+\frac{1}{\nu}\mathbf{f}^T {K_t}^{-1}\mathbf{f}\right. \\ & \left.+\frac{\beta}{(\nu+n)}(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f})\right]^{-(\nu+2n)/2} \\ \propto & \left[1+\frac{\beta}{\nu+n}\mathbf{y}^T\left(I-\frac{\beta}{\nu+n}A^{-1}\right)\mathbf{y}\right. \\ & \left.+(\mathbf{f}-\bar{\mathbf{f}})^T A(\mathbf{f}-\bar{\mathbf{f}})\right]^{-(\nu+2n)/2}, \end{aligned} \tag{15}$$

where

$$\begin{aligned} A &= \frac{1}{\nu}{K_t}^{-1}+\frac{\beta}{\nu+n}I, \\ \bar{\mathbf{f}} &= \frac{\beta}{\nu+n}A^{-1}\mathbf{y}. \end{aligned} \tag{16}$$

By integrating out $\mathbf{f}$ in Eq. (15), we get the marginal likelihood

$$p(\mathbf{y}) \propto \left[1+\frac{\beta}{\nu+n}\mathbf{y}^T\left(I-\frac{\beta}{\nu+n}A^{-1}\right)\mathbf{y}\right]^{-(\nu+n)/2}, \tag{17}$$

$$p(\mathbf{y}) = \frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|\Sigma_t|^{1/2}}\left[1+\frac{1}{\nu}\mathbf{y}^T{\Sigma_t}^{-1}\mathbf{y}\right]^{-(\nu+n)/2}, \tag{18}$$

where Σ_t is the correlation matrix of $\mathbf{y}$ and $\Sigma_t^{-1} = \frac{\nu\beta}{\nu+n}(I - \frac{\beta}{\nu+n}A^{-1})$. It's not difficult to show that $\Sigma_t = K_t + \frac{\nu+n}{\nu}\frac{1}{\beta}I$, i.e.,

$$\Sigma_t = K_t + \sigma_t^2 I, \quad \sigma_t^2 = \frac{\nu+n}{\nu}\cdot\frac{1}{\beta}. \tag{19}$$

The negative logarithm marginal likelihood of TPRD is

$$\begin{aligned} -\ln p(\mathbf{y}) = & \frac{\nu+n}{2}\ln\left[1+\frac{1}{\nu}\mathbf{y}^T{\Sigma_t}^{-1}\mathbf{y}\right]+\frac{1}{2}\ln|\Sigma_t| \\ & -\ln\left(\frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}}\right). \end{aligned} \tag{20}$$

Similar to the three terms of GPR's negative logarithm marginal likelihood in Eq. (5), the three terms in Eq. (20) has readily interpretable roles: the second term $\frac{1}{2}\ln|\Sigma_t|$ is the complexity penalty depending only on the kernel function and the inputs; the last term $\ln(\frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}})$ is for normalization; the first term is related to the data-fit term $\mathbf{y}^T{\Sigma_t}^{-1}\mathbf{y}$. Comparing the first term in Eq. (20) with the one in Eq. (5), the main difference is that the first term in Eq. (20) is a logarithm function of $\mathbf{y}^T{\Sigma_t}^{-1}\mathbf{y}$ while the one in Eq. (5) is a linear function of $\mathbf{y}^T{\Sigma_g}^{-1}\mathbf{y}$. That implies if there are outliers in $\mathbf{y}$, the negative logarithm marginal likelihood of TPRD would be much less disturbed than that of GPR. So, from the view of the negative logarithm marginal likelihood, TPRD should be more robust than GPR.

After deriving the negative logarithm marginal likelihood of TPRD Eq. (20), we need to estimate the hyperparameters $\sigma_t, \boldsymbol{\theta}$ and ν. It's not difficult to derive the partial derivatives of the marginal likelihood of TPRD for each hyperparameter, which implies that the hyperparameters $\sigma_t, \boldsymbol{\theta}$ and ν can be learned by gradient-based optimization methods.

After learning the hyperparameters, we can make predictions by the following result [14].

Lemma 1 *Suppose* $\mathbf{y} \sim St(\nu, \boldsymbol{\mu}, K)$. $\mathbf{y}_1 \in R^{n_1}$ *and* $\mathbf{y}_2 \in R^{n_2}$ *represent the first* n_1 *and remaining* n_2 *entries of* $\mathbf{y}$ *respectively. Then* $\mathbf{y}_1|\mathbf{y}_2 \sim St(\mathbf{y}_1|\boldsymbol{\mu}_{1|2}, K_{1|2}, \nu+n_1)$, *where* $\boldsymbol{\mu}_{1|2} = K_{12}K_{22}^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2)+\boldsymbol{\mu}_1$, $K_{1|2} = \frac{\nu+\alpha_1}{\nu+n_1}(K_{22}-K_{21}K_{11}^{-1}K_{12})$, $\alpha_1 = (\mathbf{y}_1-\mu_{\mathbf{1}})^T K_{11}^{-1}(\mathbf{y}_1-\mu_{\mathbf{1}})$.

For a known input $x_* \in R^D$, we need to give the prediction y_*. As $\{\mathbf{y}, y_*\}$ is a multivariate Student-t distribution, by Lemma 1,

$$\begin{aligned} p(y_*|\mathbf{y}) &= St\left(y_*\middle|\mathbf{k}_*^T\Sigma_t^{-1}\mathbf{y}, \frac{\nu+\alpha_t}{\nu+n}\left(k(x_*,x_*;\boldsymbol{\theta})-\mathbf{k}_*^T\Sigma_t^{-1}\mathbf{k}_*\right)\right) \\ \mathbf{k}_* &= \{k(\mathbf{x}_i,\mathbf{x}_*;\boldsymbol{\theta})\}_{i=1}^n, \quad \alpha_t = \mathbf{y}^T\Sigma_t^{-1}\mathbf{y}. \end{aligned} \tag{21}$$

Comparing Eq. (21) with Eq. (6), we see that the form of the predictive mean of TPRD is identical to that of GPR, which implies if the kernel function and hyperparameters are the same, the predictive mean of TPRD is also the same to that of GPR. As mentioned in Section 2.1, the predictive covariance of GPR does not depend on the target vector. In contrast, from Eq. (21), we see that the predictive covariance of TPRD explicitly depends on the target vector, which implies the uncertainties are better accounted for.

2.3 Relation to GPR

Theorem 1 *TPRD* $\rightarrow$ *GPR when* $\nu \rightarrow +\infty$.

As $\nu \rightarrow +\infty$, it's well known that a Student-t distribution converges to a Gaussian distribution (refer to, e.g., [9]), hence,

$$\begin{aligned} p\left(\mathbf{f}|X\right) &= St\left(\mathbf{f}\middle|\nu,\mathbf{0},K\right) \rightarrow \mathcal{N}\left(\mathbf{f}|\mathbf{0},K\right), \\ p(\boldsymbol{\epsilon}) &= St\left(\boldsymbol{\epsilon}\middle|\nu+n,\mathbf{0},\left(1+\frac{1}{\nu}\mathbf{f}^T K^{-1}\mathbf{f}\right)\frac{1}{\beta}I\right) \\ &\rightarrow \mathcal{N}\left(\boldsymbol{\epsilon}\middle|\mathbf{0},\frac{1}{\beta}I\right), \end{aligned} \tag{22}$$

i.e., TPRD $\rightarrow$ GPR when $\nu \rightarrow +\infty$, which implies that if we choose ν by cross-validation, we can guarantee the performances of TPRD, if not better, are at least as good as that of GPR. Of course, cross-validation for hyperparameter selection requires more computational cost. But gradient-based optimization methods can roughly achieve the same performances, as showed in the experiments of Section 4.

2.4 Relation to TPRK

The Student-t process has been studied for a long time [11]. However, as the sum of two independent student-t distributions or the sum of a student-t distribution and a Gaussian distribution is analytically intractable, it is difficult to use Student-t process with noise.

In [14, 15, 19], the noise is incorporated into the kernel function, specifically, by adding a diagonal matrix to the kernel matrix. In this way, the covariance matrix of marginal likelihood equals the kernel matrix plus a diagonal matrix. In practice, this way achieved good performances. However, probabilistic properties of the noise are unknown. In [14], the negative marginal likelihood of TPRK has the form

$$\begin{aligned} -\ln p\left(\mathbf{y}|\nu,\sigma_a,\boldsymbol{\theta}\right) = & \frac{\nu+n}{2}\log\left(1+\frac{1}{\nu-2}\mathbf{y}^T\Sigma_a^{-1}\mathbf{y}\right) \\ & +\frac{1}{2}\log\left(|\Sigma_a|\right)-\ln\left(\frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)(\nu-2)^{n/2}\pi^{n/2}}\right), \end{aligned} \tag{23}$$

where $\Sigma_a = K_a + \sigma_a^2 I$, σ_a^2 denotes the noise incorporated to the kernel . Comparing Eq. (23) with Eq. (20), it is easy to see that if $\Sigma_a = \frac{\nu-2}{\nu}\Sigma_t$, Eq. (23) and Eq. (20) would have exactly the same form. The slight difference is caused by the fact that [14] adopts a slightly different definition of Student-t distribution, where the definition is

$$\begin{aligned} St(\nu, \boldsymbol{\mu}, R) =& \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\nu-2)^{p/2}\pi^{p/2}|R|^{1/2}} \\ & \cdot \left[1 + \frac{1}{\nu-2}(\mathbf{y}-\boldsymbol{\mu})^T R^{-1}(\mathbf{y}-\mu)\right]^{-\frac{\nu+p}{2}}. \end{aligned} \tag{24}$$

It is not difficult to prove that if we applied this definition of Student-t distribution to Eq. (11) and Eq. (13), the marginal likelihood of TPRD would have the same form as that of TPRK. Besides, Σ_t is equivalent to Σ_a. So, TPRD is equivalent to TPRK. In fact, TPRD gives a probabilistic interpretation to TPRK. By incorporating the noise into the kernel, in fact, the noise approximates to the i.i.d Gaussian noise with variance adjusted to the data-fit term.

3 THEORETICAL ANALYSIS

GPR, TPRD and TPRK are all kernel methods. With different kernels, their performances change a lot. In this section, we will give the definition of the β-property. Then we will prove that if the kernel has the β-property, the predictive mean of TPRD has the same predictive mean as GPR does. Moreover, TPRK also has identical predictive mean as GPR does.

3.1 Maximum likelihood estimation of hyperparameters independent of ν

In this section, we will prove the maximum marginal likelihood estimation of hyperparameters $\boldsymbol{\theta}$ and σ_t is independent of ν if the kernel function has the β-property, which is defined as follows.

Definition 2. For a kernel function $k(\mathbf{x}_1, \mathbf{x}_2; \theta_1, \theta_2, \ldots, \theta_l)$, where $\theta_1, \theta_2, \ldots, \theta_l$ is the parameters of the kernel, if $k(\mathbf{x}_1, \mathbf{x}_2; \theta_1, \theta_2, \ldots, \theta_l) = g(\theta_1)k(\mathbf{x}_1, \mathbf{x}_2; 1, \theta_2', \ldots, \theta_l')$, where $\theta_i'(i = 2, 3, \ldots, l)$ corresponds to θ_i one to one for given θ_1, and $g(\theta_1)$ is an injective function of θ_1 with the range $(0, +\infty)$, then the kernel $k(\mathbf{x}_1, \mathbf{x}_2; \theta_1, \theta_2, \ldots, \theta_l)$ is called to have the β-property.

There are several common kernels with the β-property, e.g., a diagonal squared exponential kernel [13] has the form

$$\begin{aligned} k(\mathbf{x}_1, \mathbf{x}_2; \sigma_f, \ell) &= \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}\mathbf{x}_1^T\mathbf{x}_2\right) \\ &= \sigma_f^2 k(\mathbf{x}_1, \mathbf{x}_2; 1, \ell), \quad \sigma_f > 0, \end{aligned} \tag{25}$$

and a linear with diagonal weighting kernel [13]

$$\begin{aligned} k(\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\lambda}) =& \mathbf{x}_1^T\Lambda^{-2}\mathbf{x}_2, \quad \Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_D) \\ =& \frac{\mathbf{x}_1^T\Lambda'^{-2}\mathbf{x}_2}{\lambda_1^2}, \ \Lambda' = diag\left(1, \frac{\lambda_2}{\lambda_1}, \ldots, \frac{\lambda_D}{\lambda_1}\right) \\ =& \frac{1}{\lambda_1^2} k\left(\mathbf{x}_1, \mathbf{x}_2; 1, \frac{\lambda_2}{\lambda_1}, \ldots, \frac{\lambda_D}{\lambda_1}\right), \quad \lambda_1 > 0, \end{aligned} \tag{26}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_D)$.

There are also common kernels without the β-property, e.g., a squared exponential kernel [13] has the form

$$k(\mathbf{x}_1, \mathbf{x}_2; \ell) = \exp\left(-\frac{1}{2\ell^2}\mathbf{x}_1^T\mathbf{x}_2\right). \tag{27}$$

By Eq. (19), σ_t seems to be dependent on ν, however, we have the following surprising results.

Theorem 2 *If the kernel in TPRD has the β-property, then the maximum likelihood estimation of hyperparameters $\boldsymbol{\theta}$ and σ_t is independent of ν. Furthermore, the predictive mean of TPRD is the same as that of GPR.*

Proof. The marginal likelihood of TPRD has the form

$$p(\mathbf{y}|\boldsymbol{\theta}, \sigma_t) = \frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|\Sigma_t|^{1/2}}\left[1 + \frac{1}{\nu}\mathbf{y}^T{\Sigma_t}^{-1}\mathbf{y}\right]^{\frac{(\nu+n)}{-2}}, \tag{28}$$

where $\Sigma_t = K_t + \sigma_t^2 I$. As the kernel has the β-property, we have

$$\begin{aligned} \Sigma_t &= K_t + \sigma_t^2 I, \qquad (K_t)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j; \theta_1, \theta_2, \ldots, \theta_l) \\ &= g(\theta_1)K_t' + \sigma_t^2 I, \ \ (K_t')_{ij} = k(\mathbf{x}_i, \mathbf{x}_j; 1, \theta_2', \ldots, \theta_l') \\ &= \sigma_t^2\Sigma_t', \end{aligned} \tag{29}$$

where

$$\Sigma_t' = \lambda K_t' + I, \quad \lambda = \frac{g(\theta_1)}{\sigma_t^2}, \tag{30}$$

and $\theta_i'(i = 2, 3, \ldots, l)$ depends on θ_1 and θ_i.

As $g(\theta_1)$ is an injective function, it is not difficult to show that $\sigma_t, \lambda, \theta_2', \ldots, \theta_l'$ have one-to-one mapping to $\sigma_t, \theta_1, \theta_2, \ldots, \theta_l$. By Eq. (28), for a given ν, we have

$$\begin{aligned} & \max_{\sigma_t, \boldsymbol{\theta}} \ p(\mathbf{y}|\sigma_t, \boldsymbol{\theta}) \\ \Leftrightarrow & \max_{\sigma_t, \lambda, \theta_2', \ldots, \theta_l'} p(\mathbf{y}|\sigma_t, \lambda, \theta_2', \ldots, \theta_l') \\ \Leftrightarrow & \min_{\sigma_t, \lambda, \theta_2', \ldots, \theta_l'} (p(\mathbf{y}|\sigma_t, \lambda, \theta_2', \ldots, \theta_l'))^{\frac{-2}{\nu+n}} \\ \Leftrightarrow & \min_{\sigma_t, \lambda, \theta_2', \ldots, \theta_l'} h(\mathbf{y}|\sigma_t, \lambda, \theta_2', \ldots, \theta_l'), \end{aligned} \tag{31}$$

where

$$\begin{aligned} & h(\mathbf{y}|\sigma_t, \lambda, \theta_2', \ldots, \theta_l') \\ = & \left[\frac{1}{|\Sigma_t|^{1/2}}\left(1 + \frac{1}{\nu}\mathbf{y}^T{\Sigma_t}^{-1}\mathbf{y}\right)^{\frac{\nu+n}{-2}}\right]^{\frac{-2}{\nu+n}} \\ = & |\Sigma_t|^{\frac{1}{\nu+n}}(1 + \frac{1}{\nu}\mathbf{y}^T{\Sigma_t}^{-1}\mathbf{y}) \\ = & \sigma_t^{\frac{2n}{\nu+n}}|\Sigma_t'|^{\frac{1}{\nu+n}}(1 + \frac{1}{\nu\sigma_t^2}\mathbf{y}^T{\Sigma_t'}^{-1}\mathbf{y}). \end{aligned} \tag{32}$$

Recall that σ_t and λ have one-to-one mapping to σ_t and θ_1. Then deriving $h(\mathbf{y}|\sigma_t, \lambda, \theta_2', \ldots, \theta_l')$ with respect to σ_t,

$$\frac{\partial h(\mathbf{y})}{\partial \sigma_t} = \frac{2n}{\nu+n}\sigma_t^{\frac{2n}{\nu+n}-1}|\Sigma_t'|^{\frac{1}{\nu+n}} - \frac{2\sigma_t^{\frac{2n}{\nu+n}-3}|\Sigma_t'|^{\frac{1}{\nu+n}}}{\nu+n}\mathbf{y}^T{\Sigma_t'}^{-1}\mathbf{y}. \tag{33}$$

Letting Eq. (33) be zero, we get the maximum marginal likelihood estimation of σ_t

$$\hat{\sigma}_t = \sqrt{\frac{\mathbf{y}^T{\Sigma_t'}^{-1}\mathbf{y}}{n}}. \tag{34}$$

From Eq. (34) and Eq. (30), we can see that $\hat{\sigma}_t$ is determined by training data $\mathcal{D}$ and hyperparameters $\lambda, \theta_2', \theta_3', \ldots, \theta_l'$, which implies that $\hat{\sigma}_t$ is independent of ν. From Eq. (34) and Eq. (32), we obtain

$$h(\mathbf{y}) = \left(\frac{\mathbf{y}^T{\Sigma_t'}^{-1}\mathbf{y}}{n}\right)^{\frac{n}{\nu+n}}|\Sigma_t'|^{\frac{1}{\nu+n}}\left(1 + \frac{n}{\nu}\right). \tag{35}$$

Let $h'(\mathbf{y}) = \left(\frac{\mathbf{y}^T\Sigma_t'^{-1}\mathbf{y}}{n}\right)^n |\Sigma_t'|$ and $\phi = \lambda, \theta_2', \ldots, \theta_l'$, similar to Eq. (31) we have

$$\begin{aligned} &\min_{\phi} h(\mathbf{y}) \\ \Leftrightarrow &\min_{\phi} \left(\frac{\mathbf{y}^T\Sigma_t'^{-1}\mathbf{y}}{n}\right)^{\frac{n}{\nu+n}} |\Sigma_t'|^{\frac{1}{\nu+n}} \\ \Leftrightarrow &\min_{\phi} h'(\mathbf{y}). \end{aligned} \tag{36}$$

Since $h'(\mathbf{y})$ is independent of ν, the solution of $\partial h'(\mathbf{y})/\partial\phi = 0$ is independent of ν.

Now, we have proved that the maximum likelihood estimation of $\sigma_t, \lambda, \theta_2', \ldots, \theta_l'$ is independent of ν. As there is a one-to-one mapping between $\boldsymbol{\theta}, \sigma_t$ and $\sigma_t, \lambda, \theta_2', \ldots, \theta_l'$, the maximum likelihood estimation of $\boldsymbol{\theta}, \sigma_t$ is independent of ν.

From Lemma 1 in Section 2.2, we know the predictive mean of TPRD has no relationship with ν, it is only determined by the training data $\mathcal{D}$ and hyperparameters $\boldsymbol{\theta}, \sigma_t$. If the kernel has the β-property, the hyperparameters $\boldsymbol{\theta}, \sigma_t$ are shown to be independent of ν, which implies that the predictive mean of TPRD is independent of ν. In other words, no matter what ν is, the prediction of TPRD does not change. As GPR is a special case of TPRD with $\nu \to +\infty$, the predictive mean of TPRD is the same as that of GPR. □

This result might be interesting. Since in Section 2.2, we state TPRD should be more robust than GPR since the difference of their negative logarithm marginal likelihoods. Now, we prove that they have the same predictive mean with the β-property kernel. The underlying reason is that the β-property kernel has a free factor $g(\theta_1)$, which compromises their difference between the negative logarithm marginal likelihoods.

As the degrees of freedom ν does not affect the predictive mean, we can choose a suitable ν for speedup.

3.2 Predictive mean comparison

In TPRK, the marginal likelihood is different from the marginal likelihood of GPR. So, Shah, Wilson and Zoubin [14] expect that the predictive mean of TPRK would differ from that of GPR. However, we prove that when we use the maximum marginal likelihood to estimate the hyperparameters, the predictive mean of TPRK is identical to that of GPR if the kernel has the β-property.

Theorem 3 *If the maximum marginal likelihood is used to estimate the hyperparameters and the kernel has the β-property, for a given ν, the predictive mean of TPRK is the same as that of TPRD.*

Proof. Firstly, as the kernel has the β-property, we have

$$k(\mathbf{x}_1, \mathbf{x}_2; \theta_1, \theta_2, \ldots, \theta_l) = g(\theta_1)k(\mathbf{x}_1, \mathbf{x}_2; 1, \theta_2', \ldots, \theta_l'), \tag{37}$$

where the range of $g(\theta_1)$ is $(0, +\infty)$. We assume the maximum marginal likelihood estimation solution of TPRD's hyperparameters is $(\sigma_t^0, \theta_1^0, \theta_2'^0, \ldots, \theta_l'^0)$. As $g(\theta_1)$ is an injective function with the range $(0, +\infty)$ and in TPRK $\nu > 2$, there is a $\theta_1'^0$ satisfying that $\frac{\nu-2}{\nu}g(\theta_1'^0) = g(\theta_1^0)$. Then at $\left(\sqrt{\frac{\nu}{\nu-2}}\sigma_t^0, \theta_1'^0, \theta_2'^0, \ldots, \theta_l'^0\right)$, we have

$$k\left(\mathbf{x}_1, \mathbf{x}_2; \theta_1'^0, \theta_2'^0, \ldots, \theta_l'^0\right) = \frac{\nu}{\nu-2}k\left(\mathbf{x}_1, \mathbf{x}_2; \theta_1^0, \theta_2'^0, \ldots, \theta_l'^0\right) \tag{38}$$

$$\Sigma_a = K_a + \sigma_a^2 I = \frac{\nu}{\nu-2}K_t + \frac{\nu}{\nu-2}\sigma_t^2 I = \frac{\nu}{\nu-2}\Sigma_t, \tag{39}$$

which implies the marginal likelihood value of TPRK equals that of TPRD. In fact, at $\left(\sqrt{\frac{\nu}{\nu-2}}\sigma_t^0, \theta_1'^0, \theta_2'^0, \ldots, \theta_l'^0\right)$, the marginal likelihood value of TPRK reaches the maximum. Otherwise, if there is another set of hyperparameters at which the marginal likelihood value of TPRK is larger, there is a corresponding set of hyperparameters at which the marginal likelihood of TPRD is larger, contradictory to the fact that the maximum marginal likelihood solution of TPRD is $(\sigma_t^0, \theta_1^0, \theta_2'^0, \ldots, \theta_l'^0)$.

For a given known $x_* \in R^D$, the predictive mean of TPRK is

$$\begin{aligned} &k(x_*, X; \theta_1'^0, \theta_2'^0, \ldots, \theta_l'^0)^T\Sigma_a^{-1}\mathbf{y} \\ &= \frac{\nu}{\nu-2}k(x_*, X; \theta_1^0, \theta_2'^0, \ldots, \theta_l'^0)^T\frac{\nu-2}{\nu}\Sigma_t^{-1}\mathbf{y} \\ &= k(x_*, X; \theta_1^0, \theta_2'^0, \ldots, \theta_l'^0)^T\Sigma_t^{-1}\mathbf{y}. \end{aligned} \tag{40}$$

The RHS is the predictive mean of TPRD. □

Corollary 1 *If the maximum marginal likelihood is used to estimate the hyperparameters and the kernel has the β-property, the predictive mean of TPRK is the same as the prediction of GPR.*

Proof. By Theorem 3, we know that if the maximum marginal likelihood is used to estimate the hyperparameters and the kernel has the β-property, the predictive mean of TPRK has the same predictive mean as that of TPRD. And in that case, by Theorem 2, TPRD and GRP also have the same predictive mean. We conclude that in that case, TPRK has the same predictive mean as that of GPR. □

This result is interesting. As the negative logarithm marginal likelihood of TPRK is different from the one of GPR. The predictive mean of TPRK is expected to be different from that of GPR after learning the hyperparameters in [14]. However, by Corollary 1, TPRK and GPR have identical predictive mean if the maximum marginal likelihood is used and the kernel has the β-property. The underlying reason is that the kernel with the β-property has a free factor $g(\theta_1)$, which compromises the difference between their marginal likelihood.

4 EXPERIMENTS

Our experiments are designed to verify the following three propositions:

- If the kernel has the β-property, the ML estimation of TPRD's hyperparameters is independent of ν and the predictive mean of TPRD is the same as that of GPR. ν influences the convergence rate.
- If the kernel has the β-property and ML estimation is used, TPRK, TPRD and GPR have the same predictive mean.
- If the kernel does not have the β-property, TPRD performs better than GPR.

4.1 Data sets

We use the following seven data sets to carry out the experiments. All data are from the UCI data sets [7].

- **Servo Data.** This data set contains 4 attributes and 167 instances from a simulation of a servo system. The attributes include motor, screw, pgain and vgain. The output variable is class from 0.13 to 7.10.

- **Stock Data.** This data set includes returns of Istanbul Stock Exchange with seven other international index. There are 536 instances and 8 attributes. We randomly choose a subset of 400 instances for the experiments.
- **Wine Data [2].** This data set contains 12 attributes and 1599 instances associated with red wine. The attributes include acidity, residual sugar, pH and so on. The output variable is quality (score between 0 and 10). We randomly choose a subset of 400 instances for the experiments.
- **Airfoil Data.** Airfoil data comprises different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack. It contains 1503 instances and 6 attributes. We randomly select a subset of 400 instances for the experiments.
- **Yacht Data.** This data set is used to predict the hydodynamic performances of sailing yachts from dimensions and velocity. It contains 308 instances and 7 attributes.
- **Space Data.** This data set is to predict the number of O-rings that will experience thermal distress for a given flight when the launch temperature is below freezing point. It contains 23 instances and 4 attributes.
- **Concrete Data [18].** This data set is about the slump flow of concrete. It contains 1030 instances and 9 attributes. We randomly choose a subset of 400 instances for the experiments.

4.2 Experimental setup

We use the gradient descent algorithm [1] to get the minimum of the negative logarithm marginal likelihood. The maximum iteration number is 5000 and the stop criterion is that the absolute value of each hyperparameter's derivative is less than 10^{-8}. The initial value for step length is 0.001. And the initial value for the hyperparameters σ and $\theta_i (i = 1, 2, \cdots, l)$ are 1. All the data are standardized.

4.3 ν independent of the β-property kernel

We verify the ν independent property on two kernels and two data sets. The kernels are the diagonal squared exponential kernel Eq. (25) and the linear kernel with isotropic weighting [13], which has the form $k(\mathbf{x}_1, \mathbf{x}_2; \ell) = \frac{\mathbf{x}_1^T \mathbf{x}_2}{\ell^2}$. As shown in Eq. (25) and Eq. (26), both of the kernels have the β-property. The data sets are the Servo and Stock data. We compare TPRD($\nu = 3, 10, 1000, 100000$) with GPR.

Table 1. The ML estimation results on the Servo data set with the linear kernel

Model	$\ln \ell$	$\ln \sigma$	MSE	iteration
GPR	0.4880	-0.3706	0.7947	486
TPRD(ν=3)	0.4880	-0.3706	0.7947	581
TPRD(ν=10)	0.4880	-0.3706	0.7947	339
TPRD(ν=1000)	0.4880	-0.3706	0.7947	384
TPRD(ν=100000)	0.4880	-0.3706	0.7947	418

Table 2. The ML estimation results on the Servo data set with the diagonal squared exponential kernel

Model	$\ln \ell$	$\ln \sigma_f$	$\ln \sigma$	MSE	iteration
GPR	0.5588	0.6278	-1.2413	0.3041	1500
TPRD(ν=3)	0.5588	0.6278	-1.2413	0.3041	767
TPRD(ν=10)	0.5588	0.6278	-1.2413	0.3041	598
TPRD(ν=1000)	0.5588	0.6278	-1.2413	0.3041	2265
TPRD(ν=100000)	0.5588	0.6278	-1.2413	0.3041	2004

Table 3. The ML estimation results on the Stock data set with the linear kernel

Model	$\ln \ell$	$\ln \sigma$	MSE	iteration
GPR	1.1757	-0.7758	0.1954	1406
TPRD(ν=3)	1.1757	-0.7758	0.1954	496
TPRD(ν=10)	1.1757	-0.7758	0.1954	322
TPRD(ν=1000)	1.1757	-0.7758	0.1954	2484
TPRD(ν=100000)	1.1757	-0.7758	0.1954	3386

Table 4. The ML estimation results on the Stock data set with the diagonal squared exponential kernel

Model	$\ln \ell$	$\ln \sigma_f$	$\ln \sigma$	MSE	iteration
GPR	2.7766	1.7299	-0.7949	0.1377	5000
TPRD(ν=3)	2.7767	1.7300	-0.7949	0.1377	1002
TPRD(ν=10)	2.7767	1.7300	-0.7949	0.1377	887
TPRD(ν=1000)	2.7752	1.7282	-0.7949	0.1377	5000
TPRD(ν=100000)	2.7659	1.7170	-0.7953	0.1378	5000

From Table 1-4, we can see that the ML estimation of hyperparameters of TPRD is indeed the same as that of GPR. The slight difference in Table 4 is caused by that fact the maximum iteration is reached before the optimization point is reached. And the MSE of TPRD and GPR is identical, which implies that they have the same predictive mean.

Another interesting phenomenon is that the convergence rate is indeed influenced by ν. On most data sets, when ν is set to 10, TPRD has the smallest number of iteration, faster than GPR. Empirically, we recommend that ν is set around 10. The underlying reason may be that the tail thickness is appropriate for most data when ν is around 10, considering that ν controls the thickness of the tail.

The result may be remarkable, since the computational cost is an important problem of GPR. As the main computational cost lies on solving the inverse of the covariance matrix, most efforts are focused on the covariance matrix, e.g., from the sparsity [4], distributed method [3], and exploiting the structure of covariance matrix [17]. Our model provides an iteration-less way, which may be able to combine with the covariance matrix way.

4.4 Predictive mean of GPR, TPRD and TPRK with the β-property kernel

Now, we use the experiments to verify the Theorem 3 and Corollary 1 that the predictive mean of TPRK is identical to that of GPR and TPRD with the β-property kernel and ML estimation.

We compare TPRK($\nu = 3, 10, 1000, 100000$) with GPR and TPRD($\nu = 10$) on the Servo data set. The following are the experimental results.

From Table 5 and 6, we see that on each β-property kernel, whatever ν is, the MSE of TPRK is the same as that of GPR and TPRD, which implies these three models have the same predictive mean.

Besides, different from TPRD, ν influences the ML estimation of hyperparameters of TPRK. Next, we exploit how the ν influences them. From Section 3.2, we know that if the ML estimation of TPRD's hyperparameters is $(\sigma_t^0, \theta_1^0, \theta_2^0, \ldots, \theta_l^0)$, the ML estimation of TPRK's hyperparameters is $(\sigma_a^0, \theta_1'^0, \theta_2'^0, \ldots, \theta_l'^0)$, where

$$\sigma_a^0 = \sqrt{\frac{\nu}{\nu - 2}} \sigma_t^0, \tag{41}$$

$$\frac{\nu - 2}{\nu} g(\theta_1'^0) = g(\theta_1^0). \tag{42}$$

Table 5. The ML estimation results on the Servo data set with the linear kernel

Model	$\ln \ell$	$\ln \sigma$	MSE
GPR	0.5866	-0.3603	0.6628
TPRD(ν=10)	0.5866	-0.3603	0.6628
TPRK(ν=3)	0.03733	0.1890	0.6628
TPRK(ν=10)	0.4751	-0.2487	0.6628
TPRK(ν=1000)	0.5856	-0.3593	0.6628
TPRK(ν=100000)	0.5866	-0.3603	0.6628

Table 6. The ML estimation results on the Servo data set with the diagonal squared exponential kernel

Model	$\ln \ell$	$\ln \sigma_f$	$\ln \sigma$	MSE
GPR	0.5977	0.5297	-1.1933	0.1397
TPRD(ν=10)	0.5977	0.5297	-1.1933	0.1397
TPRK(ν=3)	0.5977	1.079	-0.6440	0.1397
TPRK(ν=10)	0.5977	0.6413	-1.082	0.1397
TPRK(ν=1000)	0.5977	0.5307	-1.192	0.1397
TPRK(ν=100000)	0.5977	0.5297	-1.193	0.1397

Taking TPRK(ν=3) for example, we check whether the experimental results are consistent with the relationship above.

For the hyperparameters θ_1^0 and $\theta_1'^0$, from Table 6, we know

$$\begin{aligned} \frac{\nu-2}{\nu} g(\theta_1'^0) &= \frac{\nu-2}{\nu}(\sigma_f)^2 \\ &= \frac{\nu-2}{\nu}\exp(2\sigma_f) \\ &= \frac{3-2}{3}\exp(2\times 1.079) \\ &= 2.8846 \\ g(\theta_1^0) &= (\theta_1^0)^2 \\ &= \exp(2\times 0.5297) \\ &= 2.8846, \end{aligned} \tag{43}$$

which verifies the Eq. (42).

For σ_a^0 and σ_t^0, we have

$$\begin{aligned} \sqrt{\frac{\nu}{\nu-2}}\sigma_t^0 &= \sqrt{\frac{3}{3-2}}\exp(-1.1933) \\ &= 0.5252 \\ \sigma_a^0 &= \exp(\ln \sigma_a^0) \\ &= \exp(-0.6440) \\ &= 0.5252, \end{aligned} \tag{44}$$

which is consistent with the Eq. (41).

As the ML estimation of TPRD's hyperparameters is not influenced by ν, from the relationship, we know that ν does not affect the ML estimation of hyperparameters of TPRK, except σ_a and $\theta_1'^0$. From Table 6, it is clear that ν indeed does not affect the estimation of ℓ.

4.5 Robustness of TPRD with non-β-property kernel

Now, we evaluate the robustness of TPRD on all the data sets (Section 4.1) with the squared exponential kernel, which does not have the β-property. Table 7 reports the MSE of TPRD and GPR. The MSE of data sets on which TRPD is significantly better than GPR is in boldface. We see that on the data set Airfoil, Concrete, Servo and Stock, TPRD has similar performances with GPR. Then in each training dataset, 5% of the instances are chosen randomly and each value of the target variable in these instances is randomly added or subtracted by 3 standard derivations of the target variable. It's clear that TPRD performs more robustly than GPR when the data are contaminated by the outliers. And on the data sets Space, Yacht, Wine, TPRD outperfoms GPR. Just as we state in Section 2.2, the negative logarithm marginal likelihood of TPRD is a logarithm function of the data-fit term, while that of GPR is a linear function. Therefore, TPRD is more robust than GPR theoretically and empirically, with the non-β-property kernel.

Table 7. The MSE of GPR and TPRD

Data Set	GPR MSE	TPRD MSE
Airfoil	0.3153	0.3108
Airfoil with outliers	0.5524	**0.3557**
Concrete	0.1545	0.1665
Concret with outliers	0.5503	**0.2706**
Servo	0.1541	0.1550
Servo with outliers	0.8616	**0.2451**
Stock	0.2744	0.2756
Stock with outliers	0.7833	**0.2928**
Space	0.4258	**0.3386**
Yacht	0.0544	**0.0458**
Wine	0.8362	**0.6580**

5 CONCLUSION

We have proposed a Student-t Process Regression with Dependent Student-t noise model (TPRD) in this paper, which is proved to be a generalization of GPR. In addition, TPRD gives a probabilistic interpretation to the Student-t Process Regression with noise incorporated into the Kernel (TPRK). In fact, by incorporating the noise into the kernel, the noise approximates to the Gassian noise with the variance adjusted to the data-fit term. More importantly, we analyze the influence of different kernels on TPRD and TPRK. Specifically, if the kernel has the β-property, the ML estimation of TPRD's hyperparameters is independent of ν and we also discuss how ν influences the ML estimation of the TPRK's hyperparameters. Besides, the predictive mean of TPRD, TPRK and GPR is identical, which is not expected by [14]. In that case, empirically, ν is a convergence accelerator and TPRD can be faster than GPR. On the other hand, if the kernel does not have the β-property, owing to the dependent noise, experimental results show that TPRD achieves significantly better performances than GPR.

ACKNOWLEDGMENTS

This research is supported in part by the Major State Basic Research Development Program of China (973 Program, 2012CB315803), the National Natural Science Foundation of China (61371078), and the Research Fund for the Doctoral Program of Higher Education of China (20130002110051).

REFERENCES

[1] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[2] Paulo Cortez, Antnio Cerdeira, Fernando Almeida, Telmo Matos, and Jos Reis, 'Modeling wine preferences by data mining from physicochemical properties', *Decision Support Systems*, **47**(4), 547 – 553, (2009). Smart Business Networks: Concepts and Empirical Evidence.

[3] Marc Deisenroth and Jun Wei Ng, 'Distributed gaussian processes', in *Proceedings of The 32nd International Conference on Machine Learning (ICML-15)*, pp. 1481–1490, (2015).

[4] Yarin Gal and Richard Turner, 'Improving the gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs', in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 655–664, (2015).

[5] Somesh Das Gupta and Jun Shao. Mathematical statistics, 2000.

[6] Malte Kuss, *Gaussian process models for robust regression, classification, and reinforcement learning*, Ph.D. dissertation, TU Darmstadt, 2006.

[7] M. Lichman. UCI machine learning repository, 2013.

[8] Alexander J McNeil, 'Multivariate t distributions and their applications', *Journal of the American Statistical Association*, **101**(473), 390–391, (2006).

[9] Saralees Nadarajah and Samuel Kotz, 'Mathematical properties of the multivariate t distribution', *Acta Applicandae Mathematica*, **89**(1-3), 53–84, (2005).

[10] Radford M Neal, 'Monte carlo implementation of gaussian process models for bayesian regression and classification', Technical report, Dept. of statistics and Dept. of Computer Science, University of Toronto, (1997).

[11] Anthony O'Hagan, 'Bayes–hermite quadrature', *Journal of statistical planning and inference*, **29**(3), 245–260, (1991).

[12] Carl Edward Rasmussen, 'Gaussian processes for machine learning', (2006).

[13] Carl Edward Rasmussen and Hannes Nickisch, 'The gpml toolbox version 3.5', (2015).

[14] Amar Shah, Andrew Wilson, and Zoubin Ghahramani, 'Student-t processes as alternatives to gaussian processes', in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-14)*, pp. 877–885, (2014).

[15] Arno Solin and Simo Särkkä, 'State space methods for efficient inference in student-t process regression', in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS-15)*, pp. 885–893, (2015).

[16] Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari, 'Gaussian process regression with student-t likelihood', in *Advances in Neural Information Processing Systems (NIPS-09)*, pp. 1910–1918, (2009).

[17] Andrew Wilson and Hannes Nickisch, 'Kernel interpolation for scalable structured gaussian processes (kiss-gp)', in *Proceedings of The 32nd International Conference on Machine Learning (ICML-15)*, pp. 1775–1784, (2015).

[18] I-Cheng Yeh, 'Modeling slump flow of concrete using second-order regressions and artificial neural networks', *Cement and Concrete Composites*, **29**(6), 474–480, (2007).

[19] Yu Zhang and Dit-Yan Yeung, 'Multi-task learning using generalized t process', in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, volume 9, p. 964, (2010).