Student’s \( t \)-process Regression on the Space of Probability Density Functions

Yusuke Uchiyama\(^1\), Hiroki Oka\(^1\) and Ayumu Nono\(^2\),
\(^1\)MAZIN Inc.
3-29-14 Nishi-Asakusa, Taito-ku, Tokyo 111-0035, Japan
\(^2\)Graduate School of Engineering, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
E-mail: uchiyama@mazin.tech

Abstract

This study provides an extension of the Student’s \( t \)-process regression (TPR) on the space of probability density functions as a method of system identification for the data set consist of noisy inputs and deterministic outputs with additive noises. With introducing the distance metrics of the probability density functions, the TPR can be naturally extended to the space of the probability density functions and thus prediction and hyper parameter estimation can be implemented by the same fashion of the ordinary model. In addition, with a numerical example of the proposed model, we introduce the Markov Chain Monte Carlo method for hyper parameter estimation.

1 Introduction

Regression for a relation between given pairs of inputs and outputs is one of the key topics in the area of system identification [1]. In this topic, two directions of regression are known to exist, parametric and non-parametric regression models [2]. Parametric regression model assumes that the system considered can be identified by a prescribed function with its parameters. Thus, the parameter estimation of the prescribed function with given data set is considered to be the main part of the parametric regression. For huge amount of observed data set the parametric regression often suffers from the problem of over-fitting, especially for the cases of nonlinear regressions. On the other hand, the non-parametric regression model does not assume the class of functions to estimate the relation between the given inputs and outputs. In this methodology, regression model is identified by probabilistic model on Bayesian sense [3]. Thus, the parameters of the probabilistic model are inferred as random variables, which are sampled from appropriate posterior distributions.

As a family of the non-parametric regression models, Gaussian process regression (GPR) has been proposed [4]. GPR maps input data into infinite dimensional Hilbert space by kernel functions, where Bayesian regression with the Gaussian distribution is implemented. In the Bayesian sense, parameter estimation and prediction for unknown output of the GPR are presented. The central limits theorem derives the GPR from neural networks of both single [5] and multi layers [6]. In addition, latent variable modelings of the GPR and these filtering scheme have been developed.

Despite the applications of the GPR for the problems of system identification, it is restricted to the observed data under Gaussian noise. As is well known, observed data in real world often exhibits non-Gaussian fluctuations, such as, price of financial assets, turbulent fluids, and, chemical reactions. To extend the concept of the GPR to such non-Gaussian fluctuations, the Student’s \( t \)-process regression (TPR) has been proposed [7]. The TPR is realized by marginalization of the conditional GPR with a gamma distributed random precision or a Wishart distributed precision matrix. As with the GPR, the TPR is extended to latent variable modeling [8].

Input of both the GPR and TPR is assumed to be deterministic scalar or vector value. However, realistic input data is often observed as a random variable, which is considered to be a deterministic value with observed noise or purely random variable sampled from a probability density function. Recently, the GPR for the data set consist of probability density functions as inputs and scalar values as outputs [9]. Such situation occurs in the case that system identification for observation by measurement instruments on a moving vehicle. To deal with a probability density function as input, several distance metrics for two functions are introduced to the framework of the GPR. Furthermore, a stochastic control problem has been proposed as an application of this methodology [10].

In this study, we extend the TPR for the space of probability density functions. Section 2 provides brief explanation of related works. In sec. 3, we incorporate distance metrics of probability density functions into kernel functions of the TPR, and introduce a method of hyper parameter estimation based on the Markov Chain Monte Carlo (MCMC). Section 4 exhibits a numerical example of the proposed method. Section 5 is dedicated to conclusions of this study and perspective
of our future works.

2 Related Works

2.1 Gaussian process regression

This part gives a brief introduction of the GPR for deterministic input. For a reader who needs a much
more thorough introduction, we recommend reference [4].

Suppose \( \mathcal{X} \) be a subspace of \( D \) dimensional Euclidean space \( \mathbb{R}^D \). A random map \( f : \mathcal{X} \to \mathbb{R} \) is a Gaussian process if, for an arbitrary natural number \( N \), \( \{f(x_1), f(x_2), \cdots, f(x_N)\} \) follows the Gaussian distribution with a mean vector

\[
m = \begin{bmatrix} m(x_1) \\
m(x_2) \\
\vdots \\
m(x_N) \end{bmatrix},
\]

and a covariance matrix

\[
K = \begin{bmatrix}
k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) \\
k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) \\
\vdots & \vdots & \ddots & \vdots \\
k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N)
\end{bmatrix}
\]

(2)

where \( m : \mathcal{X} \to \mathbb{R} \) is a mean function and \( k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \) is a kernel function, which is a symmetric positive definite function. The notation \( f \sim \mathcal{GP}(m, K) \) denotes \( f \) is a Gaussian process.

Given the pairs of inputs and outputs \( \mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\} \), the relation between the input \( x \in \mathbb{R}^D \) and output \( y \in \mathbb{R} \) is often inferred by the following form:

\[
y = f(x) + \epsilon,
\]

(3)

where \( \epsilon \) is a Gaussian noise with zero mean and variance \( \sigma_0 \). Since both \( y \) and \( f \) follows the Gaussian distributions,

\[
p(y|X) = \int p(y|f)p(f|X)df
\]

(4)

is also the Gaussian distribution given as

\[
p(y|X) = \mathcal{N}(m, \sigma_0 I + K).
\]

(5)

Note that the mean function of the GPR is often set to be zero without loss of generality.

In general kernel functions of the GPR contain hyper parameters. To estimate the hyper parameters, various methods have been developed, i.e., maximum-likelihood, variational inference, Monte Carlo methods.

For instance the log-likelihood of the GPR is given as

\[
\log p(y|X) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\sigma_0 I + K| \\
- \frac{1}{2} (y - m)^T (\sigma_0 I + K)^{-1} (y - m).
\]

(6)

The gradient of the log-likelihood yields the estimate of the hyper parameters. In the Bayesian sense, the prior \( p(\theta) \) is added to the log-likelihood as

\[
\log p(\theta|X, y) = \log p(y|X) + \log p(\theta) + \text{const},
\]

(7)

where additional term \( \log p(\theta) \) regularizes optimization [3].

For an additional input \( x^* \) prediction for the corresponding output is implemented by the conditional distribution of the GPR. Given the pairs of inputs and outputs \( \{X, y\} \), the conditional distribution of the GPR for the additional input \( x^* \) and the predicted output \( y^* \) is derived as

\[
p(y^*|x^*, X, y) = \mathcal{N}(m^*, k^*),
\]

(8)

where predicted the mean \( m^* \) and variance \( k^* \) are given by

\[
m^* = m(x^*) + k(x, x^*)^T (\sigma_0 I + K(X, X))^{-1} (y - m),
\]

(9)

\[
k^* = k(x^*, x^*) - k(x, x^*)^T (\sigma_0 I + K(X, X))^{-1} k(x, x^*).
\]

(10)

By these relations, the mean and variance of the predicted distribution are updated.

2.2 Student’s t-process regression

Although the GPR has many applications in the areas of system identification, machine learning, robotics and so on, the use of it is restricted to the data following the nature of Gaussian fluctuations. For non-Gaussian fluctuations, in particular, following heavy-tailed probability distributions, the TPR has been developed as a modification of the GPR.

In what follows, we consider the case that the mean function of the GPR to be zero for the sake of brevity. Suppose the covariance matrix of the GPR be multiplied by a parameter as \( \lambda^{-1} K \), of course, it includes the case \( \lambda^{-1}(\sigma_0 I + K) \) considered in the previous subsection. The parameter \( \lambda \) is assumed to be a random variable which is subjected to the gamma distribution with a positive real parameter \( \nu \) as

\[
p(\lambda) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \lambda^{\nu/2-1}e^{-\nu\lambda/2},
\]

(11)

where \( \Gamma(\cdot) \) is the Gamma function. Then the \( \lambda \) conditional Gaussian distribution is given as

\[
p(y|X, \lambda) = \frac{1}{(2\pi)^{N/2}|K|^{1/2}} \exp\left(-\frac{1}{2} y^T K^{-1} y\right).
\]

(12)

Marginalizing this conditional Gaussian distribution with respect to \( \lambda \),

\[
p(y|X) = \int_0^\infty p(y|X, \lambda)p(\lambda)d\lambda,
\]

(13)
the Student’s $t$-distribution is derived as
\[ p(y|X) = \frac{\Gamma \left( \frac{\nu + N}{2} \right)}{(\pi \nu)^{N/2} \Gamma \left( \frac{\nu}{2} \right)} \left( 1 + \frac{1}{\nu} yK^{-1}y \right)^{-\frac{\nu + N}{2}}. \] (14)

The parameter $\nu$ denotes the degree of freedom of the Student’s $t$-distribution. At the limit of $\nu \to \infty$, the TPR is reduced to the GPR.

The TPR contains hyper parameters in its mean and kernel functions and the degree of freedom. As with those of the GRP, the hyper parameters of the TPR can be estimated by several methods with its likelihood. The log-likelihood of the TPR is given as follows:
\[ \log p(y|X) = \log \Gamma \left( \frac{\nu + N}{2} \right) - \frac{N}{2} \log (\pi \nu) - \log \Gamma \left( \frac{\nu}{2} \right) - \frac{\nu + N}{2} \log \left( 1 + \frac{1}{\nu} yK^{-1}y \right). \] (15)

With the use of this log-likelihood, the hyper parameters of the TPR can be estimated.

Prediction of the TPR is also implemented by conditional distribution. Given the data set $\{X, y\}$, the prediction distribution of the TPR for an additional input $x^*$ is derived as
\[ p(y^*|x^*, X, y) = \frac{\Gamma \left( \frac{\nu^* + m^*}{2} \right)}{(\pi \nu^*)^{N/2} \Gamma \left( \frac{\nu^*}{2} \right)} \times \left[ 1 + \frac{1}{\nu^*} (y^* - m^*)^2 \right]^{-\frac{\nu^* + N}{2}}, \] (16)
where $m^*$, $k^*$, and $\nu^*$ are given by
\[ m^* = k(X, x^*)^T K(X, X)^{-1} y, \] (17)
\[ k^* = \frac{\nu - y^T K^{-1} y - 2}{\nu - N - 2} \times \left[ k(x^*, x^*) - k(X, x^*)^T K(X, X)^{-1} k(x^*, x^*) \right], \] (18)
\[ \nu^* = \nu + N. \] (19)

It is seen that the variance and degree of freedom of the prediction distribution are affected by the given data set.

3 Proposed Method

In this section we propose TPR on the space of probability density functions by incorporating distance metrics for probability density functions into kernel functions of the TPR. In addition, parameter estimation for the proposed model is implemented by MCMC.

3.1 Distance metrics on the space of probability density functions

Before our proposing framework, we introduce some distance metrics on the space of probability density functions. These are incorporated into the kernel functions of the TPR later.

In the theory of functional analysis, the $L^p$ norm is a most used distance metric, which is defined by
\[ d(f, g) = \left( \int_\Omega |f(x) - g(x)|^p d\mu(x) \right)^{\frac{1}{p}}, \] (20)
where $\Omega$ is the domain of the probability density functions $f(\cdot)$ and $g(\cdot)$, and $\mu(\cdot)$ is a measure on functional space considered.

As a modification of the Kullback-Leibler divergence to be a distance measure, the Jensen-Shannon divergence has been proposed. For the probability density functions $f(\cdot)$ and $g(\cdot)$, the Jensen-Shannon divergence is defined by
\[ d(f, g) = \frac{1}{2} \int_\Omega \left( f(x) \log \frac{f(x)}{h(x)} + g(x) \log \frac{g(x)}{h(x)} \right) dx, \] (21)
where $h(\cdot) = (f(\cdot) + g(\cdot))/2$.

The Wasserstein metric is defined by
\[ d(f, g) = \inf_h \left( \int_{\Omega \times \Omega} l(x, y)^p h(x, y) dxdy \right)^{\frac{1}{p}}, \] (22)
where $l(x, y)$ is the Euclidean distance between the vectors $x$ and $y$, and $h(x, y)$ is a joint probability density function whose marginal distributions are given by $\int_{\Omega_y} h(x, y) dy = f(x)$ and $\int_{\Omega_x} h(x, y) dx = g(y)$. Recently, the Wasserstein metric has been intensively applied in machine learning community, especially for the metric of deep neural networks [11].

3.2 Student’s $t$-process regression with distance metrics of probability density functions

The kernel functions of the TPR are symmetric positive definite bi-variate functions described by
\[ k(x, x') = k(||x - x'||), \] (23)
where $||\cdot||$ denotes the Euclidean norm for a vector. That is, the kernel function is considered to be a function of the distance between two vectors.

According to the above notion of the kernel functions, the domain of input of the TPR can be extended from vector values to probability density functions. Suppose $d(f, g)$ be a distance between the probability density functions $f(\cdot)$ and $g(\cdot)$, the corresponding kernel function is defined by
\[ k(f, g) = k(d(f, g)), \] (24)
where $k(\cdot)$ is a symmetric positive definite function with respect to the distance measure $d(f, g)$. Various kernel functions can be adapted to the TPR, for instance, the radial basis function is defined by
\[ k_{\text{RBF}}(d(f, g)) = \alpha^2 \exp \left( -\frac{d(f, g)^2}{l^2} \right), \] (25)
where $\alpha$ and $l$ are positive real parameters. This kernel function is the most used with both the GPR and TPR. The Matérn kernel functions are also applicable, which are defined by

$$ k_{M32}(d(f, g)) = \alpha^2 \left( 1 + \frac{\sqrt{3}d(f, g)}{l} \right) \times \exp \left( -\frac{\sqrt{3}d(f, g)}{l} \right) \tag{26} $$

and

$$ k_{M52}(d(f, g)) = \alpha^2 \left( 1 + \frac{\sqrt{5}d(f, g)}{l} + \frac{5d(f, g)}{3l^2} \right) \times \exp \left( -\frac{\sqrt{5}d(f, g)^2}{2l} \right). \tag{27} $$

These are known as the Matérn-32 and Matérn-52 kernel functions, respectively. As with the radial basis function, $\alpha$ and $l$ are positive real parameters.

### 3.3 Hyper parameter estimation

Here we provide the method of hyper parameter estimation. Based on the Bayesian inference, we consider the hyper parameters to be random variables sampled from posterior distributions. Let the set of the hyper parameters in the kernel functions and the degree of freedom of the Student’s $t$-distribution be $\theta$, the probability density function of the corresponding TPR is denoted by $p(y|\mathcal{P}, \theta)$, where $\mathcal{P}$ is a set of input probability density functions and $y$ is the corresponding output vector, respectively. Giving a prior distribution for the hyper parameters as $p(\theta)$, we can estimate the posterior distribution by the Bayes’ theorem

$$ p(\theta|\mathcal{P}, y) = \frac{p(y|\mathcal{P}, \theta)p(\theta)}{\int p(y|\mathcal{P}, \theta)p(\theta)d\theta}. \tag{28} $$

The integral in the denominator in the left hand side is intractable whereby the hyper parameters are numerically sampled from

$$ p(\theta|\mathcal{P}, y) \propto p(y|\mathcal{P}, \theta)p(\theta) \tag{29} $$

and normalized by these ensemble. To implement sampling efficiently, diverse numerical methods on the MCMC have been developed [12]. With sampled set of $\theta$ the hyper parameters are estimated by the statistics of numerically estimated posterior distribution.

### 4 Numerical example

As a numerical example of the proposed method, we estimate hyper parameters by the MCMC. To generate Gaussian distributions as synthetic input data, we utilize a quadratic function defined by

$$ s(\mu, \sigma) = \mu^2 + \sigma^2 \tag{30} $$

as shown in Fig. 1. Given a value for the quadratic function, the family of the Gaussian distributions is constraint on the surface $s(\mu, \sigma) = \text{const.}$ Figure 2 shows the sampled Gaussian distributions whose mean and variance are constrained on the quadratic surface. It is seen that different shapes of the Gaussian distributions are sampled. With the use of the sampled Gaussian distributions as inputs we generate the corresponding outputs from the multivariate Student’s $t$-distribution whose covariance matrix is yielded by the radial basis functions with $\alpha = 5$ and $l = 10$, and the degree of freedom $\nu = 4$.

To estimate the hyper parameters we implemented the MCMC by the Hamiltonian Monte Carlo algorithm [12]. For simplicity, $\alpha$ and $l$ are fixed, namely, only $\nu$ is estimated. Figure 3 shows estimated posterior distribution of $\nu$. The peak of the posterior exists around $\nu = 4$ whereby the MCMC estimates well the degree of freedom.
5 Conclusions

In this study, we extended the TPR on the space of probability density functions as a method of system identification for the data set consist of noisy inputs and deterministic outputs. With introducing the distance metrics on the probability density functions, the TPR can be naturally extended to the space of the probability density functions and thus prediction and hyper parameter estimation can be implemented by the same fashion of ordinary methodology. In particular, with an numerical example of the proposed model, we introduced the MCMC for hyper parameter estimation.

The proposed method is expected to be applied to other problems in the area of systems and information, such as position estimation and optimal control with a little bit modification. Applications for such problems will be our future works.

References

[1] L. Ljung: System identification, Wiley Online Library, N.Y., 1999.

[2] J. Friedman, T. Hastie, Trevor and R. Tibshirani: The elements of statistical learning, Springer series in statistics New York, N.Y., 2001.

[3] C. M. Bishop: Pattern recognition and machine learning, Springer, N.Y., 2006.

[4] C. K. Williams and C. E. Rasmussen: Gaussian processes for machine learning, MIT press Cambridge, MA, 2006.

[5] R. M. Neal: Bayesian learning for neural networks, Springer Science & Business Media, N.Y., 2012.

[6] J. Lee, et.al.: Deep neural networks as gaussian processes, arXiv:1711.00165, 2017.

[7] A. Shah, A. Wilson and Z. Ghahramani: Student-t processes as alternatives to Gaussian processes, Artificial intelligence and statistics, pp. 877–885, 2014.

[8] Y. Uchiyama and K. Nakagawa: TPLVM: Portfolio Construction by Student’s t-Process Latent Variable Model, Mathematics, 3, pp. 449, 2020.

[9] M. Dolgov and U. D. Hanebeck: A Distance-based Framework for Gaussian Processes over Probability Distributions, arXiv:1809.09193, 2018.

[10] J. Mayer, et.al.: Stochastic Optimal Control Using Gaussian Process Regression over Probability Distributions, 2019 American Control Conference (ACC), pp. 4847–4853, 2019.

[11] M. Arjovsky, S. Chintala and L. Bottou: Wasserstein GAN, arXiv:1701.07875, 2017.

[12] S. Brooks, et.al.: Handbook of Markov Chain Monte Carlo, CRC Press, Florida, 2011.