Student-t Process Regression with Student-t Likelihood

Qingtao Tang[†], Li Niu[‡], Yisen Wang[†], Tao Dai[†], Wangpeng An[†], Jianfei Cai[‡], Shu-Tao Xia[†]
[†] Department of Computer Science and Technology, Tsinghua University, China
[‡] School of Computer Science and Engineering, Nanyang Technological University, Singapore
{tqt15,dait14,wangys14,awp15}@mails.tsinghua.edu.cn; xiast@sz.tsinghua.edu.cn
lniu002@e.ntu.edu.sg; asjfcai@ntu.edu.sg

Abstract

Gaussian Process Regression (GPR) is a powerful Bayesian method. However, the performance of GPR can be significantly degraded when the training data are contaminated by outliers, including target outliers and input outliers. Although there are some variants of GPR (*e.g.*, GPR with Student-t likelihood (GPRT)) aiming to handle outliers, most of the variants focus on handling the target outliers while little effort has been done to deal with the input outliers. In contrast, in this work, we aim to handle both the target outliers and the input outliers at the same time. Specifically, we replace the Gaussian noise in GPR with independent Student-t noise to cope with the target outliers. Moreover, to enhance the robustness *w.r.t.* the input outliers, we use a Student-t Process prior instead of the common Gaussian Process prior, leading to Student-t Process Regression with Student-t Likelihood (TPRT). We theoretically show that TPRT is more robust to both input and target outliers than GPR and GPRT, and prove that both GPR and GPRT are special cases of TPRT. Various experiments demonstrate that TPRT outperforms GPR and its variants on both synthetic and real datasets.

1 Introduction

Gaussian Process Regression (GPR) is a powerful Bayesian method with good interpretability, non-parametric flexibility, and simple hyper-parameter learning [Rasmussen, 2006]. Due to its nice properties, GPR has been successfully applied to many fields, such as reinforcement learning [Rasmussen *et al.*, 2003], computer vision [Liu and Vasconcelos, 2015], spatio-temporal modeling [Senanayake *et al.*, 2016].

In GPR, the basic model is $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$, where $\boldsymbol{y} = \{y_i\}_{i=1}^n$ is the target vector, $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ is the collection of input vectors, and $\boldsymbol{\epsilon}$ is the noise. The latent function f is given a Gaussian Process prior and $\boldsymbol{\epsilon}$ is assumed to be independent and identically distributed (*i.i.d.*) Gaussian noise. In practice, as the number of input vectors is finite, the latent variables $f(\boldsymbol{X})$ follow a multivariate Gaussian distribution. Due to the thin-tailed property of the Gaussian distribution, GPR performs poorly on the data from heavy-tailed distributions or with outliers. However, real-world data often exhibit heavy-tailed phenomena [Nair *et al.*, 2013] and contain outliers [Bendre *et al.*, 1994; Niu *et al.*, 2015; 2016].

In order to handle the outliers, heavy-tailed distributions (*e.g.*, Laplace distribution, mixtures of Gaussians, and Student-t distribution) have been introduced into GPR. In particular, Laplace noise is used in [Kuss, 2006] while mixed two forms of Gaussian corruption are used in [Naish-Guzman and Holden, 2008]. In [Neal, 1997; Vanhatalo *et al.*, 2009; Jylänki *et al.*, 2011], the noise $\boldsymbol{\epsilon}$ is assumed to follow the Student-t distribution (GPRT). However, all these methods are only robust to the target outliers, but not robust to the outliers in the inputs $\boldsymbol{X}$, since the latent variables $f(\boldsymbol{X})$ are still assumed to follow the Gaussian distribution.

Related to the robustness *w.r.t.* the outliers in the inputs $\boldsymbol{X}$, some works [Shah *et al.*, 2014; Solin and Särkkä, 2015; Tang *et al.*, 2016] rely on Student-t Process to handle the input outliers. Particularly, the method in [Shah *et al.*, 2014; Solin and Särkkä, 2015] replaces the Gaussian Process with the Student-t Process and incorporates the noise term into the kernel function (TPRK) for computational simplicity. Following [Shah *et al.*, 2014; Solin and Särkkä, 2015], an input dependent Student-t noise (TPRD) is proposed in [Tang *et al.*, 2016]. Note that Tang *et al.* [2016] prove that TPRK, TPRD, and GPR have the same predictive mean if the kernel has a certain property named β property, which is actually satisfied by most kernels. Taking the frequently used kernels implemented by GPML [Rasmussen and Nickisch, 2010] (the most popular toolbox in Gaussian Process community) as examples, 24 out of 28 kernels have β property, for which the above Student-t Process based methods (*i.e.*, TPRK and TPRD) have the same predictive value as GPR and thus fail to deal with the input outliers effectively.

In this paper, with the aim to handle both the input outliers and the target outliers at the same time, we propose Student-t Process Regression with Student-t Likelihood (TPRT). In our model, the latent function f is assumed to be a Student-t Process prior while the noise is assumed to be an independent Student-t noise, instead of the noise incorporated into kernel as in [Shah *et al.*, 2014; Solin and Särkkä, 2015] or dependent noise as in [Tang *et al.*, 2016]. In addition to owning all the advantages of GPR, such as good interpretability, non-

parametric flexibility, and simple hyper-parameter learning, our proposed TPRT method is robust to both input and target outliers, because the Student-t Process prior contributes to robustness to the input outliers while the independent Student-t noise assumption can cope with the target outliers. One challenge of our TPRT method is that the inference is analytically intractable. To solve the inference problem, we utilize Laplace approximation for computing the posterior and marginal likelihood. The computational cost of TPRT is roughly the same as that of GPRT, which also requires approximate inference. From the perspective of posterior and marginal likelihood, we show that TPRT is more robust than GPR and GPRT. Besides, both GPR and GPRT are proved to be special cases of TPRT. Finally, extensive experiments also demonstrate the effectiveness of our TPRT method on both synthetic and real datasets.

2 Background

In this section, we will briefly introduce Gaussian Process Regression (GPR), provide the definitions of Student-t distribution and Student-t Process, and then compare Gaussian Process (GP) with Student-t Process (TP).

2.1 Review of GPR

In a regression problem, we have a training set $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ of n instances, where $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ and $\boldsymbol{x}_i$ denotes a d-dim input vector; $\boldsymbol{y} = \{y_i\}_{i=1}^n$ and y_i denotes a scalar output or target. In GPR, we have

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

where ϵ_i $(i = 1, 2, ..., n)$ is assumed to be *i.i.d.* Gaussian noise and the latent function f is given a GP prior, implying that any finite subset of latent variables $\boldsymbol{f} = \{f(\boldsymbol{x}_i)\}_{i=1}^n$ follow a multivariate Gaussian distribution, *i.e.*, $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{K}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{K})$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{K}$ is the covariance matrix. Specifically, $\boldsymbol{\mu}$ is an n-dim vector which is usually assumed to be $\mathbf{0}$ for simplicity, and $\boldsymbol{K}$ is the covariance matrix with $\boldsymbol{K}_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}_k)$, in which k is a kernel function and $\boldsymbol{\theta}_k = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kl})$ is the set of kernel parameters. As ϵ_i is *i.i.d.* Gaussian noise, given the latent variables $\boldsymbol{f}$, the likelihood can be represented as

$$p(\boldsymbol{y}|\boldsymbol{f}, \sigma) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma^2\boldsymbol{I}), \tag{2}$$

where σ^2 is the variance of the noise and $\boldsymbol{I}$ is an $n \times n$ identity matrix. Based on Bayes' theorem, we can obtain the following marginal likelihood by integrating over $\boldsymbol{f}$:

$$p(\boldsymbol{y}|\boldsymbol{X}, \sigma, \boldsymbol{K}) = \mathcal{N}(\boldsymbol{y}|\mathbf{0}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\Sigma} = \boldsymbol{K} + \sigma^2\boldsymbol{I}$. Then, the hyper-parameters σ and $\boldsymbol{\theta}_k$ can be learned by minimizing the negative logarithm marginal likelihood

$$-\ln p(\boldsymbol{y}|\boldsymbol{X}, \sigma, \boldsymbol{K}) = \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{n}{2}\ln 2\pi. \tag{4}$$

After learning the hyper-parameters σ and $\boldsymbol{\theta}_k$, given an input $\boldsymbol{x}_* \in \mathbb{R}^d$, the predictive mean is

$$\mathbb{E}(f_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) = \boldsymbol{k}_*^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{k}_* = \{k(\boldsymbol{x}_i, \boldsymbol{x}_*; \boldsymbol{\theta}_k)\}_{i=1}^n$. Please refer to [Rasmussen, 2006] for more details.

2.2 Student-t Distribution and Student-t Process

The Student-t distribution [McNeil, 2006] we use in this paper is defined as follows.

Definition 1. An n-dim random vector $\boldsymbol{x} \in \mathbb{R}^n$ follows the n-variate Student-t distribution with degrees of freedom $\nu \in \mathbb{R}_+$, mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$, and correlation matrix $\boldsymbol{R} \in \Pi(n)$ if its joint probability density function (PDF) is given by

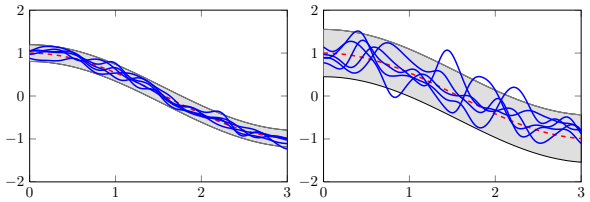
$$St(\boldsymbol{x}|\nu, \boldsymbol{\mu}, \boldsymbol{R}) = \frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|\boldsymbol{R}|^{1/2}} \cdot \left[1 + \frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{R}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+n}{2}}.$$

Given the definition of Student-t distribution, we can have the definition of Student-t Process [Shah *et al.*, 2014].

Definition 2. The process f is a Student-t Process (TP) on $\mathcal{X}$ with degrees of freedom $\nu \in \mathbb{R}_+$, mean function m: $\mathcal{X} \to \mathbb{R}$, and kernel function k: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if any finite subset of function values have a multivariate Student-t distribution, *i.e.*, $\boldsymbol{f} = \{f(\boldsymbol{x}_i)\}_{i=1}^n \sim St(\nu, \boldsymbol{\mu}, \boldsymbol{K})$ where $\boldsymbol{K} \in \Pi(n)$ with $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}_k)$ and $\boldsymbol{\mu} \in \mathbb{R}^n$ with $\boldsymbol{\mu}_i = m(\boldsymbol{x}_i)$. We denote that the process f is a Student-t Process with degrees of freedom ν, mean function m, and kernel function k as $f \sim TP(\nu, m, k)$.

2.3 Comparison of GP and TP

In [Shah *et al.*, 2014], it has been proved that GP is a special case of TP with degrees of freedom $\nu \to +\infty$. Among all the elliptical processes with an analytically representable density, TP is the most general one, which implies its expressiveness for nonparametric Bayesian modeling. The comparison of TP and GP is illustrated in Figure 1 ([Shah *et al.*, 2014]), from which we can see that TP allows the samples (blue solid) to be away from the mean (red dashed) while the samples of GP gather around the mean. This indicates that the outliers (usually away from the mean) will not have much effect on the mean of TP, but will affect the mean of GP severely as GP enforces the samples to be close to the mean. Since we make prediction mainly with the predictive mean in practice, TP is expected to be more robust to outliers than GP.



(a) Gaussian Process (b) Student-t Process

Figure 1: A comparison of TP and GP with identical mean function (red dashed), kernel function, and hyper-parameters. Degrees of freedom ν of TP is 5. The blue solid lines are the samples from the processes and the gray shaded area represents 95% confidence interval.

3 Our TPRT Method

3.1 Assumptions of TPRT

Now, we introduce our Student-t Process Regression with Student-t Likelihood (TPRT) model. In the regression problem (1), we assume

$$f \sim TP(\nu_1, m, k), \tag{6}$$

$$\epsilon_i \overset{ind}{\sim} St(\epsilon_i|\nu_2, 0, \sigma^2), \quad i = 1, 2, \ldots, n. \tag{7}$$

Compared with GPR, TPRT has the following two differences:

- The latent function f is assumed to be TP, which is more robust than GP. From Definition 2, we know that the latent variables $\boldsymbol{f}$ follow the multivariate Student-t distribution instead of Gaussian distribution. It has been theoretically analyzed in [Box and Tiao, 1962; O'Hagan, 1979] that Student-t distribution is a robust alternative to Gaussian distribution. Specifically, Student-t distribution can reject up to N outliers when there are at least $2N$ observations in total. With sufficient data, the degrees of freedom ν_1, which controls how heavy the tail of the distribution is, can be estimated based on maximum marginal likelihood, yielding an adaptive robust procedure. Based on the above discussions, the latent variables $\boldsymbol{f} = \{f(\boldsymbol{x}_i)\}_{i=1}^n$ are expected to be more robust to input outliers.
- Secondly, the noise ϵ_i $(i = 1, 2, \ldots, n)$ is assumed to follow the *i.i.d.* Student-t distribution instead of the *i.i.d.* Gaussian distribution or the dependent Student-t distribution. With (1) and (7), we can learn that the likelihood $p(y_i|f(\boldsymbol{x}_i))$ is a Student-t distribution, which accounts for the robustness to target outliers. The reasons for using *i.i.d.* Student-t noise instead of other heavy-tailed distributions are as follows: (a) Student-t distribution is a natural generalization of Gaussian distribution. When degrees of freedom ν_2 approaches $+\infty$, Student-t distribution reduces to Gaussian distribution. (b) The robustness of Student-t distribution has been theoretically proved in [Box and Tiao, 1962; O'Hagan, 1979], as mentioned above. (c) The *i.i.d.* Student-t noise assumption (7) has been successfully used in some previous works such as Bayesian Linear Regression [West, 1984], spatio-temporal modeling [Chen *et al.*, 2012]. Thus, we expect that TPRT is also robust to outliers in the targets $\boldsymbol{y}$.

To the best of our knowledge, TPRT is the first improvement of GPR that considers the robustness to both input and target outliers. As discussed in Section 1, the previous approaches in [Kuss, 2006; Naish-Guzman and Holden, 2008; Vanhatalo *et al.*, 2009; Jylänki *et al.*, 2011] replace the *i.i.d.* Gaussian noise with heavy tailed noise, only aiming at robustness to outliers in the targets $\boldsymbol{y}$. Student-t Process is used in [Shah *et al.*, 2014; Solin and Särkkä, 2015; Tang *et al.*, 2016], but the robustness to input outliers is compromised by their dependent noise assumption.

In contrast, our method combines the strengths of these two types of methods and avoids their weaknesses. Specifically, on one hand, the Student-t Process assumption (see (6)) enhances the robustness to input outliers. On the other hand, independent Student-t noise (likelihood) (see (7)) tackles the thin-tailed issue of Gaussian noise and avoids the problems of the dependent noise.

One challenge of our proposed method is that the inference is analytically intractable. Inspired by the Laplace approximation in Gaussian Process Classification [Rasmussen, 2006] and GPRT [Vanhatalo *et al.*, 2009], we propose a Laplace approximation for the posterior and the marginal likelihood, which are required for prediction and hyper-parameter learning separately. Note that a similar approach has been considered by West [1984] in the case of robust linear regression and by Rue *et al.* [2009] in their integrated nested Laplace approximation. In the following, for ease of presentation, we collect all the hyper-parameters into $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k, \nu_1, \nu_2, \sigma\}$ for the rest of this paper.

3.2 Approximation for the Posterior

In this section, we derive the approximation for the conditional posterior of the latent variables $\boldsymbol{f}$, and compare it with the posterior of GPRT and GPR.

In particular, from (6) and Definition 2, we have

$$\boldsymbol{f} \sim St(\nu_1, \boldsymbol{\mu}, \boldsymbol{K}).$$

Similarly as in GPR, the mean $\boldsymbol{\mu}$ is assumed to be $\mathbf{0}$ for simplicity. Then, we can have

$$p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta}) = \frac{\Gamma[(\nu_1+n)/2]}{\Gamma(\nu_1/2)\nu_1^{n/2}\pi^{n/2}|\boldsymbol{K}|^{1/2}} \cdot \left(1 + \frac{1}{\nu_1}\boldsymbol{f}^\top \boldsymbol{K}^{-1}\boldsymbol{f}\right)^{-\frac{\nu_1+n}{2}}. \tag{8}$$

Considering (7) and (1), we have the likelihood

$$p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^n \frac{\Gamma[(\nu_2+1)/2]}{\Gamma(\nu_2/2)\nu_2^{1/2}\pi^{1/2}\sigma} \cdot \left[1 + \frac{1}{\nu_2}\left(\frac{y_i - f(\boldsymbol{x}_i)}{\sigma}\right)^2\right]^{-\frac{\nu_2+1}{2}}. \tag{9}$$

Based on (8) and (9), the posterior of the latent variables $\boldsymbol{f}$,

$$p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}) \propto p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{f}, \boldsymbol{\theta}), \tag{10}$$

is analytically intractable. To solve the posterior of the latent variables $\boldsymbol{f}$, denoting the unnormalized posterior $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{f}, \boldsymbol{\theta})$ as $\exp(\Psi(\boldsymbol{f}))$, we use a second order Taylor expansion of $\Psi(\boldsymbol{f})$ around the mode of the unnormalized posterior as follows,

$$\Psi(\boldsymbol{f}) \simeq \Psi(\hat{\boldsymbol{f}}) - \frac{1}{2}\left((\boldsymbol{f} - \hat{\boldsymbol{f}})^\top \boldsymbol{A}^{-1}(\boldsymbol{f} - \hat{\boldsymbol{f}})\right), \tag{11}$$

where $\hat{\boldsymbol{f}}$ is the mode of $\Psi(\boldsymbol{f})$

$$\hat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{f}, \boldsymbol{\theta}) = \arg\min_{\boldsymbol{f}} \ln \mathcal{Q}, \tag{12}$$

$$\ln \mathcal{Q} = \sum_{i=1}^n \frac{\nu_2+1}{2} \ln\left[1 + \frac{1}{\nu_2}\left(\frac{y_i - f(\boldsymbol{x}_i)}{\sigma}\right)^2\right] + \frac{\nu_1+n}{2} \ln\left(1 + \frac{1}{\nu_1}\boldsymbol{f}^\top \boldsymbol{K}^{-1}\boldsymbol{f}\right),$$

$\boldsymbol{A}^{-1}$ is the negative Hessian matrix of $\Psi(\boldsymbol{f})$ at $\hat{\boldsymbol{f}}$:

$$\begin{aligned}\boldsymbol{A}^{-1} &= -\nabla\nabla \ln p\left(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta}\right) p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{f},\boldsymbol{\theta}\right) \\ &= (\nu_1+n)\frac{\boldsymbol{K}^{-1}(\nu_1+\hat{\boldsymbol{f}}^\top\boldsymbol{K}^{-1}\hat{\boldsymbol{f}}) - 2\boldsymbol{K}^{-1}\hat{\boldsymbol{f}}\hat{\boldsymbol{f}}^\top\boldsymbol{K}^{-1}}{\left(\nu_1+\hat{\boldsymbol{f}}^\top\boldsymbol{K}^{-1}\hat{\boldsymbol{f}}\right)^2} \\ &\quad +\boldsymbol{W},\end{aligned}$$

where $\boldsymbol{W}$ is a diagonal matrix with

$$\boldsymbol{W}_{ii} = -(\nu_2+1)\frac{\left(y_i-\hat{f}_i\right)^2-\nu_2\sigma^2}{\left[\left(y_i-\hat{f}_i\right)^2+\nu_2\sigma^2\right]^2}, \quad i=1,2,\ldots,n.$$

Based on (10) and (11), we can reach

$$p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\left((\boldsymbol{f}-\hat{\boldsymbol{f}})^\top\boldsymbol{A}^{-1}(\boldsymbol{f}-\hat{\boldsymbol{f}})\right)\right), \quad (13)$$

which is the unnormalized PDF of $\mathcal{N}(\hat{\boldsymbol{f}},\boldsymbol{A})$. Therefore, we obtain the approximation, $q(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta})$, for $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta})$

$$p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}) \simeq q(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}) \triangleq \mathcal{N}(\hat{\boldsymbol{f}},\boldsymbol{A}), \quad (14)$$

which will be used for making predictions in Section 3.4.

Comparison with GPRT: Note that the Laplace approximation is also utilized in GPRT [Vanhatalo *et al.*, 2009] and the $\hat{\boldsymbol{f}}$ in their approximation can be written as

$$\begin{aligned}\hat{\boldsymbol{f}} &= \arg\min_{\boldsymbol{f}} \ln \mathcal{Q}', \quad (15)\\ \ln \mathcal{Q}' &= \sum_{i=1}^{n}\frac{\nu_2+1}{2}\ln\left[1+\frac{1}{\nu_2}\left(\frac{y_i-f\left(\boldsymbol{x}_i\right)}{\sigma}\right)^2\right] \\ &\quad +\boldsymbol{f}^\top\boldsymbol{K}^{-1}\boldsymbol{f}.\end{aligned}$$

We can see that the only difference between (12) and (15) is that the term $\frac{\nu_1+n}{2}\ln\left(1+\frac{1}{\nu_1}\boldsymbol{f}^\top\boldsymbol{K}^{-1}\boldsymbol{f}\right)$ in (12) is the result of a log transformation of the term $\boldsymbol{f}^\top\boldsymbol{K}^{-1}\boldsymbol{f}$ in (15). If there are input outliers, the term $\boldsymbol{f}^\top\boldsymbol{K}^{-1}\boldsymbol{f}$ would be disturbed and the log transformation can reduce the disturbance. Therefore, the mean of approximate posterior of TPRT (*i.e.*, (12)) is more robust to input outliers than that of GPRT (*i.e.*, (15)). In fact, similar log transformations have been widely used in data analysis and statistics [Box and Cox, 1964].

Comparison with GPR: For GPR, as the posterior distribution is symmetric, the mean of the posterior is equal to the mode. Thus, the mean of the posterior can be written as

$$\arg\min_{\boldsymbol{f}} \boldsymbol{f}^\top\boldsymbol{K}^{-1}\boldsymbol{f} + \sum_{i=1}^{n}\left(\frac{y_i-f\left(\boldsymbol{x}_i\right)}{\sigma}\right)^2. \quad (16)$$

Comparing (12) and (16), we can see that both terms in (12) take the log transformation of the corresponding terms in (16). Thus, when there are outliers in the targets $\boldsymbol{y}$ or in the inputs $\boldsymbol{X}$, (12) would be more robust than (16). Note that Student-t distribution is a more general form of Gaussian distribution and the degrees of freedom ν_1,ν_2, controlling the heavy tail, also affect the log transformation. In particular, when ν_1,ν_2 approach $+\infty$, (12) reduces to (16), which means there is no log transformation.

3.3 Approximation for the Marginal Likelihood

In this section, we introduce the approximation for the marginal likelihood of TPRT, which is needed for learning the hyper-parameters, and compare it with the marginal likelihood of GPR and GPRT. In particular, the marginal likelihood of TPRT is

$$\begin{aligned}p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) &= \int p\left(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta}\right) p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{f},\boldsymbol{\theta}\right)\mathrm{d}\boldsymbol{f} \quad (17)\\ &= \int \exp\left(\Psi(\boldsymbol{f})\right)\mathrm{d}\boldsymbol{f},\end{aligned}$$

which is also analytically intractable. Similar to that as in Section 3.2, we utilize the Laplace approximation. Considering the approximation of $\Psi(\boldsymbol{f})$ in (11), we can rewrite (17) as

$$p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) \simeq \exp\left(\Psi(\hat{\boldsymbol{f}})\right)\int \exp\Big(-\frac{1}{2}\big((\boldsymbol{f}-\hat{\boldsymbol{f}})^\top\cdot \boldsymbol{A}^{-1}(\boldsymbol{f}-\hat{\boldsymbol{f}})\big)\Big)\mathrm{d}\boldsymbol{f},$$

which can be evaluated analytically. Then we obtain an approximation, $-\ln q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})$, for the negative log marginal likelihood as follows,

$$\begin{aligned}-\ln p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) &\simeq -\ln q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) \\ &\triangleq \ln\mathcal{Q}|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} + \frac{1}{2}\ln|\boldsymbol{B}| + c, \quad (18)\end{aligned}$$

where $\mathcal{Q}$ is the same as in (12), $\boldsymbol{B} = \boldsymbol{K}\boldsymbol{A}^{-1}$, and $c = -\sum_{i=1}^{n}\ln\frac{\Gamma[(\nu_2+1)/2]}{\Gamma(\nu_2/2)\nu_2^{1/2}\pi^{1/2}\sigma} - \ln\frac{\Gamma[(\nu_1+n)/2]}{\Gamma(\nu_1/2)\nu_1^{n/2}\pi^{n/2}} - \frac{n}{2}\ln(2\pi)$. Note that (18) is differentiable *w.r.t.* $\boldsymbol{\theta}$ and thus the hyper-parameters $\boldsymbol{\theta}$ can be estimated by minimizing (18).

Comparisons with GPR and GPRT: The negative log marginal likelihood of GPR is provided in (4) and the approximate negative log marginal likelihood of GPRT [Vanhatalo *et al.*, 2009] is

$$\ln\mathcal{Q}'|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} + \frac{1}{2}\ln|\boldsymbol{B}'| + c', \quad (19)$$

where $\mathcal{Q}'$ is the same as in (15), $\boldsymbol{B}' = \boldsymbol{I}+\boldsymbol{K}\boldsymbol{W}$, and $c' = -\sum_{i=1}^{n}\ln\frac{\Gamma[(\nu_2+1)/2]}{\Gamma(\nu_2/2)\nu_2^{1/2}\pi^{1/2}\sigma}$. Comparing the negative log marginal likelihoods of GPR, GPRT, and TPRT, we can observe that the main difference of their negative log marginal likelihoods is also caused by the log transformation, similar to the difference of their posteriors discussed in Section 3.2. Due to the log transformation, the negative log marginal likelihood of TPRT is more stable when there are outliers in the inputs or the targets. Thus, the estimation of the hyper-parameters $\boldsymbol{\theta}$, which is the solution of minimum negative log marginal likelihood, is less affected by the outliers.

3.4 Making Predictions

Once obtaining the approximate posterior $q(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta})$ and the hyper-parameters $\boldsymbol{\theta}$, we can make predictions. Specifically, given a new input $\boldsymbol{x}_* \in \mathbb{R}^d$, the predictive mean is

computed as

$$\begin{aligned}\mathbb{E}\left(f_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}\right) &= \int \mathbb{E}\left(f_*|\boldsymbol{f},\boldsymbol{X},\boldsymbol{x}_*\right) p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta})\mathrm{d}\boldsymbol{f} \\ &\simeq \int \mathbb{E}\left(f_*|\boldsymbol{f},\boldsymbol{X},\boldsymbol{x}_*\right) q(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta})\mathrm{d}\boldsymbol{f} \\ &= \boldsymbol{k}_*^\top \boldsymbol{K}\hat{\boldsymbol{f}}, \end{aligned} \tag{20}$$

where $\boldsymbol{k}_* = \{k(\boldsymbol{x}_i,\boldsymbol{x}_*;\boldsymbol{\theta}_k)\}_{i=1}^n$ and we use the fact for TP: $\mathbb{E}\left(f_*|\boldsymbol{f},\boldsymbol{X},\boldsymbol{x}_*\right) = \boldsymbol{k}_*^\top \boldsymbol{K}\boldsymbol{f}$ (see Lemma 3 in [Shah *et al.*, 2014]). The predictive variance can be derived in a similar way.

3.5 Implementation Details

When computing the Laplace approximation $q(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta})$ for the posterior of latent variables $\boldsymbol{f}$, the main computational effort lies in finding the mode of the unnormalized log posterior, *i.e.*, solving (12). In [Rasmussen, 2006], Newton's method is used to find the mode. However, in our model, as the negative Hessian matrix of the log posterior is not necessarily positive definite, we cannot use Newton's method and thus use a conjugate gradient algorithm instead.

After obtaining the mode $\hat{\boldsymbol{f}}$ using (12), The hyper-parameters $\boldsymbol{\theta}$ can be estimated by minimizing the approximate negative log marginal likelihood (18). Our implementation is also based on the conjugate gradient optimization, in which we need to compute the derivatives of (18) *w.r.t.* $\boldsymbol{\theta}$. The dependency of the approximate marginal likelihood on $\boldsymbol{\theta}$ is two-fold:

$$\begin{aligned}\frac{\partial \ln q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} &= \sum_{k,l} \frac{\partial \ln q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\partial \boldsymbol{K}_{k,l}} \frac{\partial \boldsymbol{K}_{k,l}}{\partial \boldsymbol{\theta}_i} \\ &\quad + \frac{\partial \ln q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\partial \hat{\boldsymbol{f}}} \frac{\partial \hat{\boldsymbol{f}}}{\partial \boldsymbol{\theta}_i}. \end{aligned} \tag{21}$$

In particular, as $\ln q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})$ is a function of $\boldsymbol{K}$, there is an explicit dependency via the terms involving $\boldsymbol{K}$. Besides, since the change in $\boldsymbol{\theta}$ will cause a change in $\hat{\boldsymbol{f}}$, there is an implicit dependency through the terms involving $\hat{\boldsymbol{f}}$. The explicit derivative of (18) can be easily obtained. The implicit derivative accounts for the dependency of (18) on $\boldsymbol{\theta}$ due to change in the mode $\hat{\boldsymbol{f}}$. As $\hat{\boldsymbol{f}}$ is the solution to (12), we have

$$\left.\frac{\partial \ln \mathcal{Q}}{\partial \boldsymbol{f}}\right|_{\boldsymbol{f}=\hat{\boldsymbol{f}}} = 0. \tag{22}$$

Thus, differentiating (18) *w.r.t.* $\hat{\boldsymbol{f}}$ can be simplified as $\partial \ln |\boldsymbol{B}|/\partial \hat{\boldsymbol{f}}$. Then, we can have

$$\frac{\partial \ln q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\partial \hat{\boldsymbol{f}}} \frac{\partial \hat{\boldsymbol{f}}}{\partial \boldsymbol{\theta}_i} = -\frac{1}{2} \frac{\partial \ln |\boldsymbol{B}|}{\partial \hat{\boldsymbol{f}}} \frac{\partial \hat{\boldsymbol{f}}}{\partial \boldsymbol{\theta}_i}, \tag{23}$$

where $\frac{\partial \hat{\boldsymbol{f}}}{\partial \boldsymbol{\theta}_i}$ can be obtained by differentiating (22) *w.r.t.* $\boldsymbol{\theta}_i$ (see [Kuss and Rasmussen, 2005]). After obtaining the explicit and implicit derivatives of (18) *w.r.t.* $\boldsymbol{\theta}$, we can solve $\boldsymbol{\theta}$ using conjugate gradient descent.

4 Relations to GPR and GPRT

Theorem 1 *GPR is a special case of TPRT with* $\nu_1, \nu_2 \to +\infty$.

Proof. When $\nu_1, \nu_2 \to +\infty$, the TP assumption (6) reduces to GP and the Student-t noise (7) reduces to Gaussian noise. In this case, since the Laplace method approximates the posterior and marginal likelihood with Gaussian distributions, it is not difficult to prove that the obtained posterior and ML using Laplace approximation are exactly the same as those of GPR. Therefore, TPRT $\to$ GPR with $\nu_1, \nu_2 \to +\infty$. □

Theorem 2 *GPRT is a special case of TPRT with* $\nu_1 \to +\infty$.

Proof. The proof is similar to that of Theorem 1 and thus we omit the details here due to the space limitation. □

Theorem 1 and 2 indicate that TPRT is a natural generalization of GPR and GPRT, like Student-t distribution/process is a generalization of Gaussian distribution/process. When tuning ν_1, ν_2 carefully, we can guarantee that the performance of TPRT is at least comparable with the performances of GPR and GPRT. In fact, the hyper-parameters ν_1, ν_2 obtained from maximum marginal likelihood instead of manual tuning can already achieve satisfactory performance, as we will show in the experiments.

5 Experiments

In this section, we evaluate our TPRT method on both synthetic and real datasets. The experimental results demonstrate the effectiveness of our TPRT method. We also investigate why our TPRT method works on real datasets.

5.1 Datasets

The experiments are conducted on both synthetic and real datasets. For the synthetic datasets, we use Neal Dataset [Neal, 1997] and its variant with input outliers. The training data are described as follows,

- **Neal Data.** This dataset was proposed by [Neal, 1997], containing target outliers. This dataset contains 100 instances with one attribute.
- **Neal with input outliers.** To study the robustness to the input outliers, we add outliers to 5% of inputs of the Neal data, *i.e.*, 5% of elements in the inputs $\boldsymbol{X}$ of Neal data are randomly added or subtracted by 3 standard derivations. We use Neal_Xoutlier to denote Neal with input outliers in Table 2.

In the testing stage, 50 instances are generated from the underlying true function [Neal, 1997] as test data.

For real datasets, we use 6 real datasets from [Alcalá *et al.*, 2010; Lichman, 2013] and their detailed information is listed in Table 1, from which we can observe that 6 real datasets have different numbers of instances and attributes, and cover a wide range of areas. Following [Shah *et al.*, 2014], on Concrete and Bike datasets, a subset of 400 instances are randomly chosen from each dataset for the experiments. For each dataset, 80% of instances are randomly sampled as the training data and the remaining instances are used as the test data.

Table 1: Detailed information of the real datasets.

Dataset	# of Instances	# of Attributes	Area
Diabetes	43	2	Medical
Machine CPU	209	6	Computer
MPG	398	8	Industry
ELE	495	2	Industry
Concrete	1030	9	Physical
Bike	17389	16	Social

5.2 Baselines, Experimental Settings and Computational Complexity

For the kernel, we choose the squared exponential kernel [Rasmussen and Nickisch, 2010], which is one of the most commonly used kernels. As this kernel has the β property, the methods in [Shah *et al.*, 2014; Solin and Särkkä, 2015; Tang *et al.*, 2016] generate the same predictive mean as GPR. Thus, we only report the results of GPR as a representative. We also include GPRT as a baseline, in which the Laplace approximation is also used following [Vanhatalo *et al.*, 2009] for fair comparison. Each experiment is repeated 10 times and the average results of all methods on each dataset are reported. We use the root mean squared error (RMSE) between the predictive mean and the true value for evaluation.

Recall that when learning the hyper-parameters, we use conjugate gradient method, in which the maximum iteration number is set as 100. The initial values for the kernel parameters $\boldsymbol{\theta}_k$ and the variance of the noise σ^2 are set as 1. Note that we use the same initial $\boldsymbol{\theta}_k$ and σ^2 for the baselines GPR and GPRT.

The main computational cost of TPRT lies in solving the inverse of the kernel matrix ($O(n^3)$ time complexity), which can be accelerated by most methods speeding up GPR, such as methods proposed by [Williams and Seeger, 2000; Deisenroth and Ng, 2015; Wilson and Nickisch, 2015].

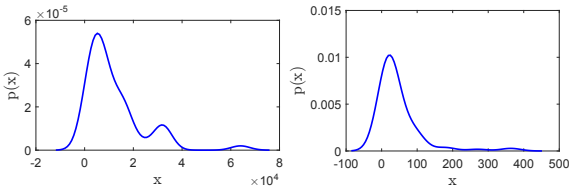
5.3 Experimental Results

The experimental results are summarized in Table 2, from which we can observe that GPRT and TPRT have better performance than GPR on the Neal dataset, which demonstrates the effectiveness of GPRT and TPRT for handling the target outliers. On the Neal_Xoutlier dataset, TPRT outperforms GPR and GPRT significantly and the reason can be explained as follows. The Student-t Process assumption (6) is more tolerant to the input outliers, and the log transformations in the posterior and marginal likelihood reduce the effect of the outliers. On all the real datasets, TPRT also outperforms all the baselines and achieves the best results. Especially on the Machine CPU dataset, the RMSE of TPRT is 13.24% lower than that of GPR.

We further investigate why our TPRT method works on the real datasets by studying the distribution of each real dataset. In particular, we use kernel density estimation [Silverman, 1986] to estimate the probability density function of each attribute for each dataset. One observation is that most datasets (5 out of 6 datasets except the Diabetes dataset) have certain heavy-tailed distributed input attributes, which may be caused by input outliers. Taking Machine CPU and Concrete as examples, Figure 2 reports their estimated probability density function. It is clear that the estimated densities of both datasets have heavy tails in some degrees. As analyzed in Section 3.2 and 3.3, TPRT is more robust to the input outliers than GPR and GPRT. This may explain why TPRT performs better on these datasets. Especially for the Machine CPU dataset, all input attributes exhibit heavy-tail distribution, which may explain why TPRT has much lower RMSE on this dataset as demonstrated in Table 2.

Table 2: The RMSE results of our TPRT method and all baselines on different datasets. The best results on each dataset are denoted in boldface.

Dataset	GPR	GPRT	TPRT
Neal	0.1676	0.0739	**0.0558**
Neal_Xoutlier	0.4392	0.3815	**0.3369**
Diabetes	0.8895	0.8870	**0.8512**
Machine CPU	0.4719	0.4811	**0.4094**
MPG	0.3376	0.3360	**0.3287**
ELE	0.5785	0.5740	**0.5500**
Concrete	0.4013	0.4005	**0.3894**
Bike	0.3445	0.3536	**0.3383**



(a) Machine CPU (b) Concrete

Figure 2: The heavy-tailed phenomenon on the real datasets. Each subfigure shows the results of a representative attribute of the corresponding dataset.

6 Conclusion

In this paper, motivated by the fact that neither GPR nor its variants are robust to both input and target outliers, we propose a novel Student-t Process Regression with Student-t Likelihood (TPRT), which is robust to the input and target outliers at the same time. Specifically, we propose the Student-t Process prior assumption to handle the input outliers and Student-t likelihood (noise) assumption to handle the target outliers. We derive a Laplace approximation for the inference and analyze why TPRT is robust to both input and target outliers from the views of the posterior and marginal likelihood. We also prove that both GPR and TPRT are special cases of TPRT. Extensive experiments demonstrate the effectiveness of our TPRT method on both synthetic and real datasets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant Nos. 61371078, 61375054, and the R&D Program of Shenzhen under grant Nos. JCYJ20140509172959977, JSGG20150512162853495, ZDSYS20140509172959989, JCYJ20160331184440545. This work is partially supported by NTU-SCSE Grant 2016-2017 and NTU-SUG (CoE) Grant 2016-2018.

References

[Alcalá *et al.*, 2010] J Alcalá, A Fernández, J Luengo, J Derrac, S García, L Sánchez, and F Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287, 2010.

[Bendre *et al.*, 1994] SM Bendre, V Barnett, and T Lewis. Outliers in statistical data, 1994.

[Box and Cox, 1964] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

[Box and Tiao, 1962] G. E. P. Box and G. C. Tiao. A further look at robustness via bayes's theorem. *Biometrika*, 49(3/4):419–432, 1962.

[Chen *et al.*, 2012] Yang Chen, Feng Chen, Jing Dai, T Charles Clancy, and Yao-Jan Wu. Student-t based robust spatio-temporal prediction. In *ICDM*, pages 151–160. IEEE, 2012.

[Deisenroth and Ng, 2015] Marc Deisenroth and Jun Wei Ng. Distributed gaussian processes. In *ICML*, pages 1481–1490, 2015.

[Jylänki *et al.*, 2011] Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust gaussian process regression with a student-t likelihood. *JMLR*, 12(Nov):3227–3257, 2011.

[Kuss and Rasmussen, 2005] Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary gaussian process classification. *JMLR*, 6(Oct):1679–1704, 2005.

[Kuss, 2006] Malte Kuss. *Gaussian process models for robust regression, classification, and reinforcement learning*. PhD thesis, Darmstadt University of Technology, Germany, 2006.

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

[Liu and Vasconcelos, 2015] Bo Liu and Nuno Vasconcelos. Bayesian model adaptation for crowd counts. In *CVPR*, pages 4175–4183, 2015.

[McNeil, 2006] Alexander J McNeil. Multivariate t distributions and their applications. *Journal of the American Statistical Association*, 101(473):390–391, 2006.

[Nair *et al.*, 2013] Jayakrishnan Nair, Adam Wierman, and Bert Zwart. The fundamentals of heavy-tails: Properties, emergence, and identification. *SIGMETRICS Perform. Eval. Rev.*, 41(1):387–388, June 2013.

[Naish-Guzman and Holden, 2008] Andrew Naish-Guzman and Sean Holden. Robust regression with twinned gaussian processes. In *NIPS*, pages 1065–1072, 2008.

[Neal, 1997] 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.

[Niu *et al.*, 2015] Li Niu, Wen Li, and Dong Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, June 2015.

[Niu *et al.*, 2016] L. Niu, X. Xu, L. Chen, L. Duan, and D. Xu. Action and event recognition in videos by learning from heterogeneous web sources. *TNNLS*, PP(99):1–15, 2016.

[O'Hagan, 1979] Anthony O'Hagan. On outlier rejection phenomena in bayes inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 358–367, 1979.

[Rasmussen and Nickisch, 2010] Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning toolbox. *JMLR*, 11(Nov):3011–3015, 2010.

[Rasmussen *et al.*, 2003] Carl Edward Rasmussen, Malte Kuss, et al. Gaussian processes in reinforcement learning. In *NIPS*, volume 4, page 1, 2003.

[Rasmussen, 2006] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[Rue *et al.*, 2009] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[Senanayake *et al.*, 2016] Ransalu Senanayake, Simon OCallaghan, and Fabio Ramos. Predicting spatio–temporal propagation of seasonal influenza using variational gaussian process regression. In *AAAI*, 2016.

[Shah *et al.*, 2014] Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to gaussian processes. In *AISTAT*, pages 877–885, 2014.

[Silverman, 1986] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[Solin and Särkkä, 2015] Arno Solin and Simo Särkkä. State space methods for efficient inference in student-t process regression. In *AISTAT*, pages 885–893, 2015.

[Tang *et al.*, 2016] Qingtao Tang, Yisen Wang, and Shu-Tao Xia. Student-t process regression with dependent student-t noise. In *ECAI*, pages 82–89, 2016.

[Vanhatalo *et al.*, 2009] Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with student-t likelihood. In *NIPS*, pages 1910–1918, 2009.

[West, 1984] Mike West. Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 431–439, 1984.

[Williams and Seeger, 2000] Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *NIPS*, pages 661–667, 2000.

[Wilson and Nickisch, 2015] Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes. In *ICML*, pages 1775–1784, 2015.