Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Robust Gaussian process regression with G-confluent likelihood

Martin Lindfors [*,**] **Tianshi Chen** [**] **Christian A. Naesseth** [***]

[*] *Department of Electrical Engineering, Linköping University, Linköping, 581 83, Sweden (e-mail: martin.lindfors@liu.se).*
[**] *School of Science and Engineering and Shenzhen Research Institute of Big Data. The Chinese University of Hong Kong, Shenzhen, 518172, China (e-mail: tschen@cuhk.edu.cn)*
[***] *Data Science Institute, Columbia University, New York, 10027, USA*

Abstract:
For robust Gaussian process regression problems where the measurements are contaminated by outliers, a likelihood/measurement noise model with heavy-tailed distributions should be used to improve the prediction performance. In this paper, we propose to use a G-confluent distribution as the measurement noise model and a coordinate ascent variational inference method to infer the overall statistical model. In contrast with the commonly used Student's t distribution, the G-confluent distribution can also be written as a Gaussian scale mixture, but its inverse scale follows a Beta distribution rather than a Gamma distribution, and its main advantage is that it is more flexible for modeling outliers while being equally suitable for variational inference. Numerical simulations based on benchmark data show that the G-confluent distribution performs better than or as well as the Student's t distribution.

Keywords: Bayesian methods, Machine learning, Nonparametric methods, Gaussian process regression, Outliers, Variational inference

1. INTRODUCTION

For regression problems, the data could be contaminated by outliers, that is, intermittent and large deviations from typical values of the measurements, e.g., [Huber, 2011]. Outliers may occur, for example, because of failures in the measurements or omission of certain regression variables in the problem. To improve the prediction performance, the outliers should be handled carefully when inferring the model. One way to handle outliers is to use a likelihood/measurement noise model with heavy-tailed distributions, such as the Student's t or Laplace distribution.

Gaussian process regression is a fundamental regression method and has wide applications in many engineering fields, e.g., [Rasmussen and Williams, 2006]. One of its advantages is that when the measurement noise model is Gaussian, the inference can be computed analytically. However, in the presence of outliers in the data, a non-Gaussian measurement noise model has to be used and as a result, the inference is not analytically tractable and approximate inference methods have to be applied. The standard/most common choice of the measurement noise model is the Student's t distribution, e.g., [Jylänki et al., 2011, Kuss, 2006, Tipping and Lawrence, 2005] and other choices include the Laplace distribution, e.g., [Kuss, 2006]. Approximate inference methods include Monte Carlo [Neal, 1997], expectation propagation [Jylänki et al., 2011], variational inference [Kuss, 2006, Tipping and Lawrence, 2005], and the Laplace approximation [Kuss, 2006, Rasmussen and Williams, 2006].

In this paper, we consider Gaussian process regression in the presence of outliers and we propose to use a G-confluent distribution as the measurement noise model and a coordinate ascent variational method to infer the overall statistical model, which is shown to converge a stationary point. In contrast with the Student's t distribution, the G-confluent distribution can also be written as a Gaussian scale mixture, but its inverse scale follows a Beta distribution rather than a Gamma distribution, and its main advantage is that it is more flexible to model outliers, while being equally suitable for variational inference. The proposed G-confluent distribution is inspired by [Wang et al., 2017]. Numerical simulations based on benchmark data show that the G-confluent distribution performs better than or as well as the Student's t distribution.

2. PROBLEM STATEMENT AND FORMULATION

We consider the following robust regression problem

⋆ Tianshi Chen is the corresponding author. This work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. This work is also supported in part by the National Natural Science Foundation of China under contract No. 61773329, the Thousand Youth Talents Plan funded by the central government of China, the Shenzhen Projects Ji-20170189 (JCY20170411102101881) funded by the Shenzhen Science and Technology Innovation Council, the President's grant under contract No. PF. 01.000249 and the Start-up grant under contract No. 2014.0003.23 funded by the Chinese University of Hong Kong, Shenzhen.

$$y_k = f(x_k) + e_k, \quad k = 1, \ldots, N, \tag{1}$$

where $y_k \in \mathbb{R}$ is the measurement, $x_k \in \mathbb{R}^n$ is the input, $e_k \in \mathbb{R}$ is the measurement noise and $f(\cdot)$ is the unknown function to be inferred. Ideally, the measurement noises e_k, $k = 1, \ldots, N$, are assumed to be i.i.d. Gaussian distributed with mean 0 and variance $R > 0$. In this paper, we consider the case where some of e_k, $k = 1, \ldots, N$, are contaminated by outliers and our goal is to infer $f(\cdot)$ as well as possible in terms of the prediction performance based on the data y_k, x_k, $k = 1, \ldots, N$.

2.1 Gaussian Process Model

The unknown function $f(\cdot)$ in (1) is modeled as a Gaussian process with mean zero and covariance function $k(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that is also called the kernel function. We define the function value vector $F = [f_1, \ldots, f_N]^\top$ with $f_k = f(x_k)$ and the regression matrix $X = [x_1, \ldots, x_N]$. Then F is distributed according to a multivariate Gaussian distribution

$$p(F|\eta_F) = \mathcal{N}(F|0, K(X|\eta_F)), \tag{2}$$

with mean 0 and covariance matrix $K(X|\eta_F)$ which has been parameterized as a function of X and the hyper-parameter η_F, which is constrained in a set Ω_{η_F}. The structure of $K(X|\eta_F)$ depends on the chosen kernel $k(\cdot, \cdot)$ and the squared exponential kernel is used here, e.g., [Rasmussen and Williams, 2006]. Thus for $i, j = 1, \ldots, N$, the (i, j)th element of $K(X|\eta_F)$ is

$$[K(X|\eta_F)]_{i,j} = c \exp\left(- \sum_{m=1}^{n} \frac{|x_{i,m} - x_{j,m}|^2}{2\lambda_m} \right), \tag{3a}$$

$$\eta_F = [c, \lambda_1, \ldots, \lambda_n]^\top, \tag{3b}$$

where $x_{i,m}$ is the mth element of x_i, and $\eta_F \in \Omega_{\eta_F} = \{c > 0, \lambda_m > 0, m = 1, \ldots, n\}$.

2.2 Measurement Noise Model with Gaussian Scale Mixture

To handle possible outliers in the measurement noises e_k, $k = 1, \ldots, N$, a measurement noise model with heavy-tailed distributions should be used instead of the Gaussian distribution $\mathcal{N}(0, R)$, such as the Student's t and Laplace distributions. Interestingly, both the Student's t and Laplace distributions can be written as Gaussian scale mixtures. More specifically, the measurement noises e_k, $k = 1, \ldots, N$, are modeled as Gaussians with PDF

$$p(e_k|R, z_k) = \mathcal{N}(e_k|0, R/z_k), k = 1, \cdots, N, \tag{4}$$

where $z_1, \ldots, z_N$ are latent variables assumed to be i.i.d., their PDF is parameterized by a parameter η_Z, and η_Z is constrained in a set Ω_{η_Z}. When z_k is Gamma distributed, e_k is Student's t distributed and when z_k^{-1} is exponentially distributed, e_k is Laplace distributed.

Here, we also use the Gaussian scale mixture (4) as the measurement noise model, but we choose to use a Beta distribution for the latent variables $z_1, \ldots, z_N$, which makes the measurement noise e_k has a G-confluent distribution. In contrast with the Student's t and Laplace distributions, the Beta distribution has two independent parameters, giving the G-confluent distribution a more flexible structure for outlier modeling, which will be discussed and justified in detail in Section 3.

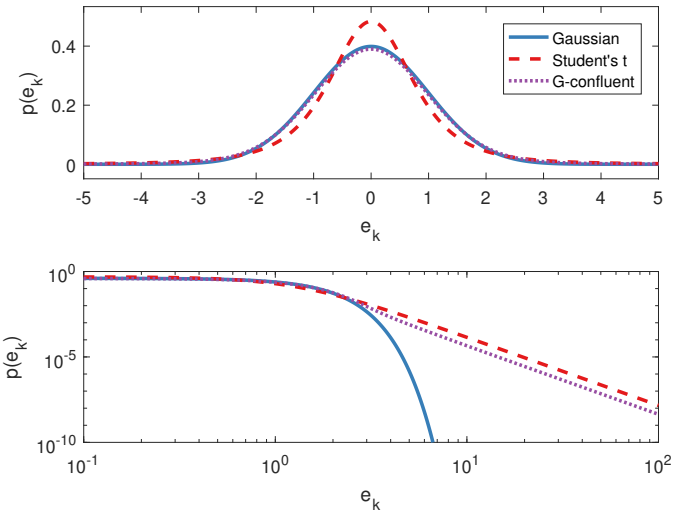

Fig. 1. Probability density functions (PDFs) of the Gaussian, Student's t, and G-confluent distributions. Top: linear-linear scale. Bottom: log-log scale. The parameters are $\mu = 0$, $\nu = 3$, $a = 1.5$ and $b = 0.3$, while the scale parameter R has been chosen to set the variance equal to 1.

2.3 Overall Statistical Model

We group the measurements, latent variables, and hyper-parameters as

$$Y = [y_1, \cdots, y_N]^\top, \tag{5}$$

$$Z = [z_1, \cdots, z_N]^\top, \tag{6}$$

$$\eta = [\eta_F^\top, R, \eta_Z^\top]^\top, \tag{7}$$

where the hyper-parameter η is constrained in $\Omega_\eta = \Omega_{\eta_F} \times \Omega_R \times \Omega_{\eta_Z}$ with $\Omega_R = \{R > 0\}$.

The joint statistical model can be written as

$$p(F, Z, Y|\eta) = p(F|\eta_F) \prod_{k=1}^{N} p(y_k|f_k, R, z_k) p(z_k|\eta_Z), \tag{8}$$

where $p(F|\eta_F)$ is given in (2),

$$p(y_k|f_k, R, z_k) = \mathcal{N}(y_k|f_k, R/z_k), \tag{9}$$

and $p(z_k|\eta_Z)$ is the PDF of z_k. We skip the conditioning on X for brevity.

2.4 Model Inference

Our goal is to infer the model (8) for F, Z, and a point estimate η^* of the hyper-parameters η, given the data Y and X. The empirical Bayes method is applied and given by

$$p(F, Z|Y, \eta^*) = \frac{p(F, Z, Y|\eta^*)}{p(Y|\eta^*)}, \tag{10a}$$

$$\eta^* = \arg\max_{\eta \in \Omega_\eta} \log p(Y|\eta), \tag{10b}$$

$$p(Y|\eta) = \int p(F, Z, Y|\eta) dF dZ. \tag{10c}$$

The integral (10c) cannot be analytically computed for the given problem, implying that (10) is intractable. In Section 5, we solve the inference problem approximately by a variational method.

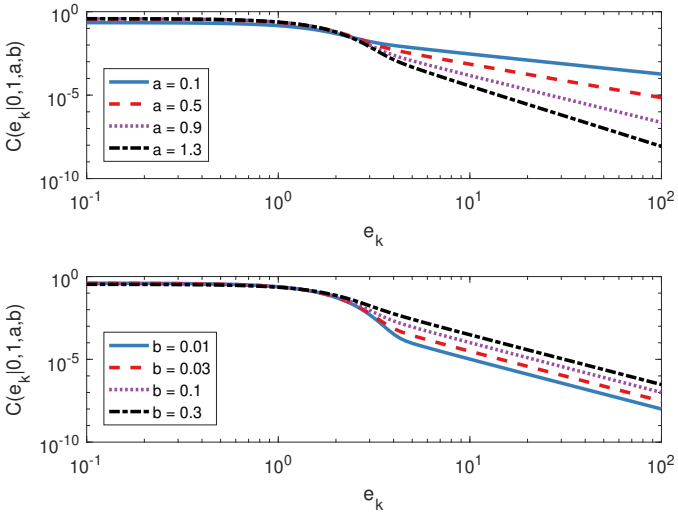

Fig. 2. Right tail of the PDF of the G-confluent distribution $\mathcal{C}(e_k|0,1,a,b)$ in log-log scale. Top: with $b = 0.1$ and different values of a given in the legend. Bottom: with $a = 1$ and different values of b given in the legend.

3. MEASUREMENT NOISE MODEL WITH A G-CONFLUENT DISTRIBUTION

As mentioned in Section 2.2, we use a Gaussian scale mixture (4) as the noise model. The novelty here is that the i.i.d. latent variables $z_1, \ldots, z_N$ are assumed to be Beta distributed, i.e.,

$$p(z_k|\eta_Z) = \mathcal{B}(z_k|a,b) = \frac{\Gamma(a+b)z_k^{a-1}(1-z_k)^{b-1}}{\Gamma(a)\Gamma(b)}, \quad (11)$$

where the hyper-parameter $\eta_Z = [a,b]^\top$ is constrained to a set $\Omega_{\eta_Z} = \{a > 0, b > 0\}$, and $\Gamma(\cdot)$ is the Gamma function. With (11), it can be shown that e_k has a G-confluent distribution whose PDF is denoted by $\mathcal{C}(e_k|0,R,a,b)$ and takes the following form

$$\mathcal{C}(e_k|0,R,a,b) = \int \mathcal{N}(e_k|0,R/z_k)\,\mathcal{B}(z_k|a,b)dz_k$$
$$= \frac{\Gamma(a+b)\Gamma(a+\frac{1}{2})}{\Gamma(a)\Gamma(a+b+\frac{1}{2})\sqrt{2\pi R}} \cdot M(a+\tfrac{1}{2}, a+b+\tfrac{1}{2}, -\tfrac{1}{2R}e_k^2), \quad (12)$$

where M is the confluent hypergeometric function, see [Olver et al., 2010, Ch. 13] and Appendix **??**. This analytic expression for the PDF of the G-confluent distribution is a key feature of the proposed measurement noise model. The the corresponding derivation can be found in detail in Appendix **??**.

4. MORE FLEXIBLE FOR OUTLIER MODELING

In contrast with the Student's t distribution, the main advantage of the G-confluent distribution (12) is that it is more flexible for outlier modeling because its two parameters, a and b, can be both used to adjust the tail behavior, as can be seen in Figure 1. A small b keeps the G-confluent distribution similar to the Gaussian near the mode, but a small a allows for heavy tails, which cannot be realized by the Student's t distribution. The role of a and b is further illustrated in Figure 2, which shows that a and b control the slope and height of the tail of the PDF, respectively. Finally, it is worth noting that as $a \to \infty$ with b kept constant, or as $b \to 0$ with a kept constant, it holds that $\mathcal{C}(e_k|0,R,a,b) \to \mathcal{N}(e_k|0,R)$.

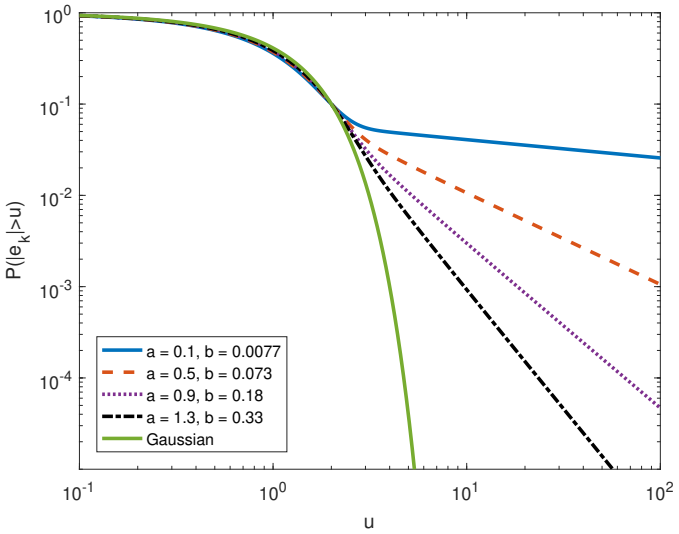

Fig. 3. Probability $P(|e_k| > u)$, where $e_k \sim \mathcal{C}(0,1,a,b)$. The values of a and b are given in the legend, and are such that $P(|e_k| > 2) = 0.1$. The corresponding probability for Gaussian distribution is shown for comparison.

4.1 Able to Adjust Relative Size and Occurrence Probability of Outliers

In Figure 3, the probability $P(|e_k| > u)$ has been illustrated as a function of u, where $e_k \sim \mathcal{C}(0,1,a,b)$. The probability is computed using the CDF of the G-confluent distribution, see Appendix **??**. The parameter a is varied, while b has been chosen as a function of a to let $P(|e_k| > 2) = 10^{-1}$ (note that the curves intersect at $u = 2$). As shown in Figure 3, this allows for a similar shape of the distribution close to the mode, and a constant $P(|e_k| > 2) = 10^{-1}$ (which can be thought of as the occurrence probability of outliers). The probability $P(|e_k| > 10)$ (which is related to the relative size of outliers) varies from 10^{-3} to $0.5 \cdot 10^{-1}$. This illustrates that by varying a and b, we can adjust the occurrence probability of outliers and the relative size of outliers in an uncoupled way, distinguishing the G-confluent distribution from the Student's t distribution.

5. MODEL INFERENCE WITH VARIATIONAL EM

A variational Bayes method is used to compute approximations of (10a) and (10b) using a lower bound of $\log p(Y|\eta)$. This corresponds to variational expectation maximization (EM), see, e.g., [Beal and Ghahramani, 2003, Neal and Hinton, 1998].

5.1 Variational EM

The posterior (10a) is approximated as

$$p(F,Z|Y,\eta) \approx q(F,Z) = q_F(F)q_Z(Z), \quad (13)$$

where $q_F(F)$ and $q_Z(Z)$ are general PDFs. We select $q_F(F)$, $q_Z(Z)$, and η^* through maximizing the evidence lower bound,

$$\max_{q(F,Z)\in\Omega_q,\,\eta\in\Omega_\eta} \mathscr{L}(q(F,Z);p(F,Z,Y|\eta)), \tag{14}$$

where the evidence lower bound is given by

$$\mathscr{L}(q(F,Z);p(F,Z,Y|\eta)) = \mathbb{E}_{q(F,Z)} \log \frac{p(F,Z,Y|\eta)}{q(F,Z)}. \tag{15}$$

Variational EM solves (14) using block coordinate ascent: by iteratively applying a variational E-step and a variational M-step until convergence [Neal and Hinton, 1998, Tzikas et al., 2008].

The variational E-step consists of computing an updated $q^{(j)}(F,Z) = q_F^{(j)}(F) q_Z^{(j)}(Z)$ for a given $\eta^{(j-1)}$ as follows:

$$q_F^{(j)}(F) = \operatorname*{arg\,max}_{q_F(F)\in\Omega_{q_F}} \mathscr{L}(q_F(F) q_Z^{(j-1)}(Z); p(F,Z,Y|\eta^{(j-1)})), \tag{16a}$$

$$q_Z^{(j)}(Z) = \operatorname*{arg\,max}_{q_Z(Z)\in\Omega_{q_Z}} \mathscr{L}(q_F^{(j)}(F) q_Z(Z); p(F,Z,Y|\eta^{(j-1)})), \tag{16b}$$

where $\Omega_{q_F} = \{q_F(F) > 0 \,|\, \int q_F(F) dF = 1\}$, $\Omega_{q_Z} = \{q_Z(Z) > 0 \,|\, \int q_Z(Z) dZ = 1\}$.

The variational M-step consists of computing an updated value $\eta^{(j)}$ for a given $q^{(j)}(F,Z)$ as follows:

$$\eta^{(j)} = \operatorname*{arg\,max}_{\eta\in\Omega_\eta} \mathscr{L}(q^{(j)}(F,Z); p(F,Z,Y|\eta)). \tag{17}$$

5.2 Variational E-step (16): Computing $q^{(j)}(F,Z)$

In this step, we compute the update (16). It follows from, e.g., [Bishop, 2006, Ch. 10] that it can be written as $\log q_F^{(j)}(F) = \mathbb{E}_{q_Z^{(j-1)}} \log p(F,Z,Y|\eta^{(j-1)}) + c_1$ and $\log q_Z^{(j)}(Z) = \mathbb{E}_{q_F^{(j)}} \log p(F,Z,Y|\eta^{(j-1)}) + c_2$, where the constants c_1 and c_2 must be chosen such that $q_F^{(j)}(F)$ and $q_Z^{(j)}(Z)$ integrate to 1.

Equation (16a) can then be written as

$$q_F^{(j)}(F) = \mathcal{N}(F|\mu_F^{(j)}, P_F^{(j)}), \tag{18a}$$

$$P_F^{(j)} = \left(D^{(j-1)} + K(X|\eta_F^{(j-1)})^{-1}\right)^{-1}, \tag{18b}$$

$$\mu_F^{(j)} = P_F^{(j)} D^{(j-1)} Y, \tag{18c}$$

$$D^{(j-1)} = \frac{1}{R^{(j-1)}} \operatorname{diag}\left[\mathbb{E}_{q_Z^{(j-1)}} z_1, \cdots, \mathbb{E}_{q_Z^{(j-1)}} z_N\right]. \tag{18d}$$

Equation (16b) can be written as

$$q_Z^{(j)}(Z) = \prod_{k=1}^N \mathcal{E}(z_k|a_{z,k}^{(j)}, b_{z,k}^{(j)}, c_{z,k}^{(j)}), \tag{19a}$$

$$a_{z,k}^{(j)} = a^{(j-1)} + \frac{1}{2} \tag{19b}$$

$$b_{z,k}^{(j)} = b^{(j-1)}, \tag{19c}$$

$$c_{z,k}^{(j)} = -\frac{1}{2R^{(j-1)}} \mathbb{E}_{q_F^{(j)}}(y_k - f_k)^2. \tag{19d}$$

Here, $\mathcal{E}$ denotes the PDF of the exponentially skewed Beta distribution

$$\mathcal{E}(x|a,b,c) = \frac{\Gamma(a+b) x^{a-1}(1-x)^{b-1} e^{cx}}{\Gamma(a)\Gamma(b) M(a, a+b, c)}, \tag{20}$$

where $0 \le x \le 1$, $a > 0$, $b > 0$, $c \in \mathbb{R}$. We discuss the exponentially skewed Beta distribution further in Appendix **??**.

5.3 Variational M-step (17): Computing $\eta^{(j)}$

We now solve (17) for $\eta^{(j)}$. Using (8) and (13), the evidence lower bound in (17) can be written as

$$\begin{aligned}\mathscr{L}(q^{(j)}(F,Z); p(F,Z,Y|\eta)) \\ &\mathbb{E}_{q_F^{(j)}} \log p(F|\eta_F) + \sum_{k=1}^N \mathbb{E}_{q_Z^{(j)}} \log p(z_k|\eta_Z) \\ &+ \sum_{k=1}^N \mathbb{E}_{q_Z^{(j)}} \mathbb{E}_{q_F^{(j)}} \log p(y_k|f_k, R, z_k) \\ &- \mathbb{E}_{q_F^{(j)}} \log q_F^{(j)}(F) - \mathbb{E}_{q_Z} q_Z^{(j)}(Z).\end{aligned} \tag{21}$$

This expression is analytically tractable. Only the first component depends on η_F, only the second on η_Z, and only the third on R. We can thus maximize these three components separately. This yields the updates

$$\begin{aligned}\eta_F^{(j)} = \operatorname*{arg\,max}_{\eta_F\in\Omega_{\eta_F}} \Big(&- \tfrac{1}{2} \operatorname{Tr}(K(X|\eta_F)^{-1} \mathbb{E}_{q_F^{(j)}} F F^\top) \\ &- \tfrac{1}{2} \log\det K(X|\eta_F) \Big),\end{aligned} \tag{22}$$

$$\begin{aligned}\eta_Z^{(j)} = \operatorname*{arg\,max}_{\eta_Z\in\Omega_{\eta_Z}} \Big(&(a-1) \sum_{k=1}^N \mathbb{E}_{q_Z^{(j)}} \log z_k \\ &+ (b-1) \sum_{k=1}^N \mathbb{E}_{q_Z^{(j)}} \log(1-z_k) \\ &+ N\left(\log\Gamma(a+b) - \log\Gamma(a) - \log\Gamma(b)\right) \Big),\end{aligned} \tag{23}$$

$$R^{(j)} = \frac{1}{N} \sum_{k=1}^N \mathbb{E}_{q_Z^{(j)}} z_k \, \mathbb{E}_{q_F^{(j)}} (y_k - f_k)^2. \tag{24}$$

The variational M-step consists of computing (22)–(24). Thus, it contains two low-dimensional optimization steps and an analytic update.

5.4 Summary of the Proposed Method and Its Convergence Properties

The proposed method is summarized in Algorithm 1.

6. NUMERICAL EXPERIMENTS

We evaluate the proposed method (Algorithm 1, denoted as GP, G-Confluent) in comparison with a method implemented in the same way, except that the measurement noise model is chosen to be the Student's t distribution (denoted as GP, Student's t, following e.g. [Jylänki et al., 2011, Kuss, 2006, Tipping and Lawrence, 2005]).

6.1 Method Configuration

The squared exponential kernel (3) is used for all evaluations for ease of comparison.

Algorithm 1 Gaussian process regression with G-confluent measurement noise model using variational EM.

Initialize $q_F^{(0)}(F), q_Z^{(0)}(Z), \eta^{(0)}$. Set $j = 0$.
repeat
$j = j + 1$
Variational E-step (16)*:* Compute $q_F^{(j)}(F)$ using (18), and $q_Z^{(j)}(Z)$ using (19).
Variational M-step (17)*:* Compute $\eta_F^{(j)}$ using (22), $\eta_Z^{(j)}$ using (23), and $R^{(j)}$ using (24).
Evidence lower bound: Compute $\mathscr{L}^{(j)} = \mathscr{L}(q^{(j)}(F,Z); p(F,Z,Y|\eta^{(j)}))$ using (21).
until convergence.
return the achieved limit points $q_F^*(F)$, $q_Z^*(Z)$, η^*.

Table 1. Initialization values for η, where $\hat{s}$ is the estimated noise standard deviation using Gaussian process regression with a Gaussian likelihood.

Parameter	Initialization Value(s)
a	$\{1, 2, 3\}$
b	0.1
ν	$\{2, 4, 6\}$
R	$\{0.1, 1, 10\} \cdot \hat{s}^2$
c	$\{e^{-3}, 1, e^3\}$
λ_m	1

Since the optimization problem (14) is non-convex, we propose an initialization procedure as follows: Run the algorithm initialized at several values of the hyper-parameter η, given in Table 1. The density $q_F^{(0)}(F)$ is initialized using Gaussian process regression with a Gaussian likelihood [Rasmussen and Williams, 2006]. The density $q_Z^{(0)}(Z)$ is initialized so that $\mathbb{E}_{q_{z_k}^{(0)}} z_k = 0.9$ for $k = 1, \ldots, N$. For each initial value, run the loop in Algorithm 1 for 10 iterations. Then, we select the resulting approximation with the highest evidence lower bound $\mathscr{L}$ as an initialization for Algorithm 1 and run it until convergence. As for the stopping criteria, we consider the increase in the evidence lower bound $\mathscr{L}^{(j)}$ and the iterations are deemed to have converged when $\mathscr{L}^{(j)} - \mathscr{L}^{(j-1)} < 10^{-6}$. We assume that each regressor and output has sample mean 0 and sample variance 1. The data sets were normalized if this did not hold.

6.2 Performance Assessment

We evaluate the two methods using cross-validation. The data is randomly partitioned into training data (X, Y), and test data $(\tilde{X}, \tilde{Y})$, where $\tilde{Y} = [\tilde{y}_1, \cdots, \tilde{y}_{N_t}]^\top$, $\tilde{X} = [\tilde{x}_1, \cdots, \tilde{x}_{N_t}]$, and N_t is the number of test data points. The model is inferred from the training data. We then evaluate the root mean square error (RMSE) on test data, given by

$$\text{RMSE}(\tilde{Y}) = \sqrt{\sum_{k=1}^{N_t} (\tilde{y}_k - \hat{f}_k)^2}, \tag{25}$$

where $\hat{f}_k = \mathbb{E}_{p(F|Y)} f(\tilde{x}_k)$ is the posterior predictive mean given the training data. Further, we evaluate the pointwise predictive log likelihood (PLL) on test data, given by

$$\text{PLL}(\tilde{Y}) = \sum_{k=1}^{N_t} \log p(\tilde{y}_k|Y) = \sum_{k=1}^{N_t} \log \int p(\tilde{y}_k|F) p(F|Y) dF. \tag{26}$$

The integrals are computed using 50 point Gauss-Hermite quadrature [Abramowitz and Stegun, 1964, Ch. 25].

6.3 Benchmark Data

The two methods are tested on four benchmark data sets.

Boston Housing Data We study the data set of housing prices in Boston originally published in [Harrison Jr and Rubinfeld, 1978]. The data set contains 506 observations. Each measurement y_k is the median price of houses in different parts of the Boston metropolitan area, associated with a regressor x_k containing 13 input variables per area. We randomly partition the data into a set of 200 training points and 306 test points, which is repeated for 20 times, leading to 20 Monte Carlo simulations.

Concrete Data We study the data set of concrete strength from [Yeh, 1998]. The data set contains 1030 observations, and each measurement y_k is the compressive strength of a certain batch of high performance concrete. Each batch is associated with a regressor x_k containing 8 input variables related to the components used to make the concrete. The data is randomly partitioned into a set of 200 training points and 830 test points, which is repeated for 20 times, leading to 20 Monte Carlo simulations.

Friedman Data We study the synthetic example proposed in [Friedman, 1991]. For each index k, we simulate regressors $x_k = [x_{k,1}, \ldots, x_{k,10}]^\top$ as $x_{k,i} \sim \mathcal{U}(0,1)$, $i = 1, \ldots, 10$, where $\mathcal{U}(0,1)$ is the uniform distribution on $[0,1]$. We then select the true function as $f^*(x_k) = 10 \sin(\pi x_{k,1} x_{k,2}) + 20(x_{k,3} - 0.5)^2 + 10 x_{k,4} + 5 x_{k,5}$ which does not depend on $x_{k,i}$, $i = 6, \ldots, 10$ and leads to a feature selection problem.

To generate outlier-free data, we follow [Kuss, 2006] and generate $y_k^* = f^*(x_k) + e_k$, $e_k \sim \mathcal{N}(0,1)$, $k = 1, \ldots, 200$. Then, 10 of the first 100 measurements and 10 of the second 100 measurements are replaced with outliers $y_k^o \sim \mathcal{N}(15, 3)$, respectively, which lead to the 100 training data points and 100 test data points, respectively. The above data generation procedure is repeated for 100 times, leading to 100 Monte Carlo simulations.

Fuel Consumption Data We study the data set of car fuel consumption [Quinlan, 1993]. The data set contains 398 measurements, of which 6 contained missing data and were removed, and 6 regressors which describe car models were used to predict fuel consumption. We then randomly partition the data into 200 training data points and 192 test data points, which is repeated for 20 times, leading to 20 Monte Carlo simulations.

6.4 Simulation Results and Findings

Simulation results are summarized in Table 2, which shows the mean RMSE and PLL, and Figures 4 and 5, which respectively show the dispersion of the RMSE and PLL. As shown in Table 2, for all four data sets, the noise model

Table 2. Mean of RMSE and PLL values.

RMSE	Boston H.	Concrete	Friedman	Fuel C.
G-confluent	0.872	3.020	5.154	4.272
Student's t	1.014	3.087	5.430	4.324
PLL	Boston H.	Concrete	Friedman	Fuel C.
G-confluent	-683.0	-4895.0	-437.2	-849.5
Student's t	-898.4	-9323.2	-439.9	-1198.8

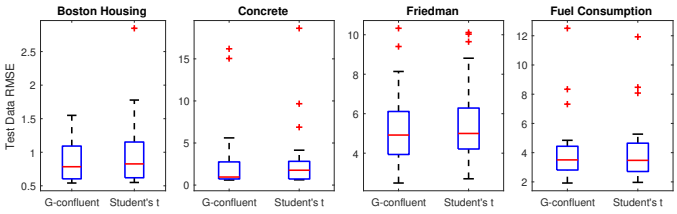

Fig. 4. Boxplots of RMSE on the benchmark data sets.

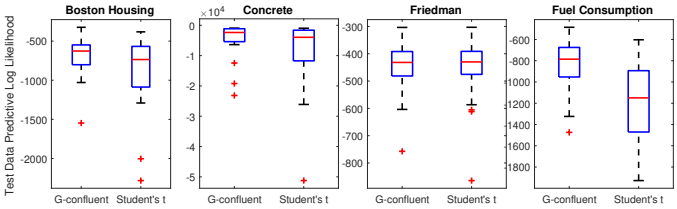

Fig. 5. Boxplots of PLL on the benchmark data sets.

with the G-confluent distribution achieves lower mean of RMSE and larger mean of PLL. In particular, the mean of RMSE is lower for the Boston housing data set and the mean of PLL is larger for all data sets except the Friedman data set. As shown in boxplots of RMSE and PLL in Figures 4 and 5, the noise model with the G-confluent distribution achieves more compact distribution of values of RMSE for all data sets except the concrete data set (which is slightly less compact) and more compact distribution of values of PLL for all data sets except the Friedman data set (which looks similar).

These simulation results show that the proposed noise model with the G-confluent distribution performs better than or as well as the Student's t distribution.

7. CONCLUSION

We have studied robust Gaussian process regression where a G-confluent distribution was proposed as the measurement noise model, suitable for coordinate ascent variational inference. Numerical experiments on benchmark data showed that the G-confluent distribution performs better than or as well as the commonly used Student's t distribution.

REFERENCES

Abramowitz, M. and Stegun, I.A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation.

Beal, M.J. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7, 453–464.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 19(1), 1–67.

Harrison Jr, D. and Rubinfeld, D.L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1), 81–102.

Huber, P.J. (2011). *Robust statistics*. Springer.

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011). Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12(Nov), 3227–3257.

Kuss, M. (2006). *Gaussian process models for robust regression, classification, and reinforcement learning*. Ph.D. thesis, Technische Universität.

Neal, R.M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *arXiv preprint physics/9701026*.

Neal, R.M. and Hinton, G.E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, 355–368. Springer.

Olver, F.W., Lozier, D.W., Boisvert, R.F., and Clark, C.W. (2010). *The NIST Handbook of Mathematical Functions*. Cambridge University Press.

Quinlan, J.R. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, 236–243.

Rasmussen, C.E. and Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.

Tipping, M.E. and Lawrence, N.D. (2005). Variational inference for Student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(1), 123–141.

Tzikas, D.G., Likas, A.C., and Galatsanos, N.P. (2008). The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, 25(6), 131–146.

Wang, Y., Kucukelbir, A., and Blei, D.M. (2017). Robust probabilistic modeling with Bayesian data reweighting. In *International Conference on Machine Learning*, 3646–3655.

Yeh, I.C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12), 1797–1808.