Robust Gaussian Process Regression with Huber Likelihood

Anonymous authors
Paper under double-blind review

	SCtMCMC^{+pw}	tLA^{+pw}	HuberMCMC^{+pw}	HuberLA^{+pw}	RCGP^{+pw}	GP^{+pw}	LaplaceMCMC^{+pw}
			$\varepsilon \sim \mathcal{N}(0.01, 0.08)$				
RMSE	2.14 (2.32)	1.26 (0.05)	**0.63** (0.14)	1.18 (0.01)	2.04 (0.005)	1.56 (0.01)	0.71 (0.11)
MAE	1.07 (0.62)	0.82 (0.04)	**0.47** (0.08)	0.81 (0.01)	1.78 (0.003)	1.38 (0.02)	0.50 (0.07)
			$\varepsilon \sim$ Student-$t(10)$				
RMSE	2.25 (1.56)	1.31 (0.00)	**0.61** (0.18)	1.19 (0.00)	2.06 (0.00)	1.58 (0.00)	0.74 (0.16)
MAE	1.11 (0.35)	0.86 (0.00)	**0.47** (0.11)	0.83 (0.00)	1.79 (0.00)	1.40 (0.00)	0.53 (0.10)
			$\varepsilon \sim$ Laplace$(0, 0.1)$				
RMSE	1.75 (1.05)	1.31 (0.00)	**0.65** (0.15)	1.19 (0.00)	2.06 (0.002)	1.58 (0.00)	0.69 (0.16)
MAE	1.00 (0.24)	0.86 (0.00)	**0.49** (0.10)	0.83 (0.00)	1.08 (0.00)	1.40 (0.00)	0.50 (0.09)
			$\varepsilon \sim$ Student-$t(1)$ (Cauchy)				
RMSE	1.33 (0.12)	1.31 (0.00)	**0.67** (0.23)	1.19 (0.00)	2.06 (0.00)	1.57 (0.00)	0.71 (0.16)
MAE	0.90 (0.05)	0.86 (0.00)	**0.50** (0.15)	0.83 (0.00)	1.05 (0.00)	1.39 (0.00)	0.52 (0.09)

Table 1: RMSE and MAE values on the Neal dataset for the Case 1 (with standard errors in parenthesis). Bold values highlight the best performance with the lowest RMSE and MAE.

	SCtMCMC^{+pw}	tLA^{+pw}	HuberMCMC^{+pw}	HuberLA^{+pw}	RCGP^{+pw}	GP^{+pw}	LaplaceMCMC^{+pw}
			$\varepsilon \sim \mathcal{N}(0.01, 0.08)$				
RMSE	0.91 (0.04)	0.97 (0.14)	1.14 (0.23)	**0.84** (0.09)	2.04 (0.004)	0.90 (0.02)	0.90 (0.09)
MAE	0.62 (0.03)	0.59 (0.05)	0.76 (0.16)	**0.57** (0.06)	1.68 (0.002)	0.67 (0.02)	0.61 (0.04)
			$\varepsilon \sim$ Student-$t(10)$				
RMSE	0.85 (0.02)	0.81 (0.00)	0.98 (0.12)	**0.80** (0.08)	2.05 (0.00)	0.88 (0.00)	0.87 (0.12)
MAE	0.61 (0.01)	0.**56** (0.00)	0.67 (0.08)	**0.56** (0.06)	1.70 (0.001)	0.66 (0.00)	0.62 (0.08)
			$\varepsilon \sim$ Laplace$(0, 0.1)$				
RMSE	0.85 (0.02)	**0.81** (0.00)	0.99 (0.13)	0.83 (0.07)	2.05 (0.002)	0.88 (0.00)	0.84 (0.07)
MAE	0.61 (0.01)	**0.55** (0.00)	0.68 (0.10)	0.58 (0.05)	1.69 (0.00)	0.66 (0.00)	0.59 (0.04)
			$\varepsilon \sim$ Student-$t(1)$ (Cauchy)				
RMSE	0.86 (0.04)	0.82 (0.01)	0.91 (0.10)	**0.79** (0.06)	2.05 (0.00)	0.88 (0.01)	0.83 (0.07)
MAE	0.62 (0.03)	0.56 (0.01)	0.61 (0.08)	**0.55** (0.05)	1.69 (0.00)	0.66 (0.01)	0.58 (0.05)

Table 2: RMSE and MAE values on the Neal dataset for the Case 3 (with standard errors in parenthesis). Bold values highlight the best performance with the lowest RMSE and MAE.

Theorem 1. *Let $\mathcal{D} = (\boldsymbol{x}_i, y_i)_{i=1}^n$ be a dataset with distinct covariates $\boldsymbol{x}_i \in \mathcal{X}$ and response $y_i \in \mathcal{Y}$, where $n < \infty$. The kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ is positive definite, with elements $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ defined by a continuous kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Assume the Huber likelihood function $p_H(\mathbf{y}|\mathbf{f}, \boldsymbol{\sigma})$ based on strictly convex and continuous Huber loss $\rho(r_i) : \mathbb{R} \to \mathbb{R}$. Then the posterior distribution $p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \sigma)$ is unimodal.*

Proof. The GP-Huber posterior distribution is proportional to the expression:

$$p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \sigma) \propto \exp\left(-\frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} - \sum_{i=1}^{n} \rho(y_i - f_i)\right),$$

where ρ denotes the Huber loss function, which is continuous and strictly convex. The derivative of the log-posterior with respect to $\mathbf{f}$ is:

$$\nabla_{\mathbf{f}} \log p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \sigma) \propto -\mathbf{K}^{-1}\mathbf{f} - \nabla_{\mathbf{f}}\rho(y_i - f_i),$$

	SCtMCMC$^{-\text{pw}}$	tLA$^{-\text{pw}}$	HuberMCMC$^{-\text{pw}}$	HuberLA$^{-\text{pw}}$	RCGP$^{-\text{pw}}$	GP$^{-\text{pw}}$	LaplaceMCMC$^{-\text{pw}}$
			$\varepsilon \sim \mathcal{N}(0.01, 0.08)$				
RMSE	2.74 (1.84)	**1.23** (0.10)	1.24 (0.04)	1.30 (0.01)	2.03 (0.01)	1.68 (0.02)	1.29 (0.12)
MAE	1.10 (0.52)	**0.75** (0.09)	0.88 (0.03)	0.89 (0.01)	1.69 (0.008)	1.46 (0.02)	0.88 (0.06)
			$\varepsilon \sim$ Student-$t(10)$				
RMSE	1.92 (1.54)	1.18 (0.02)	**1.17** (0.04)	1.29 (0.00)	2.07 (0.00)	1.70 (0.00)	1.26 (0.10)
MAE	0.92 (0.39)	**0.71** (0.03)	0.84 (0.04)	0.88 (0.00)	1.70 (0.00)	1.48 (0.00)	0.89 (0.06)
			$\varepsilon \sim$ Laplace$(0, 0.1)$				
RMSE	3.26 (1.83)	1.20 (0.04)	**1.16** (0.10)	1.29 (0.00)	2.08 (0.00)	1.70 (0.00)	1.17 (0.08)
MAE	1.26 (0.47)	**0.74** (0.05)	0.84 (0.06)	0.88 (0.00)	1.70 (0.00)	1.48 (0.00)	0.82 (0.06)
			$\varepsilon \sim$ Student-$t(1)$ (Cauchy)				
RMSE	3.40 (1.82)	1.16 (0.06)	**1.12** (0.08)	1.29 (0.00)	2.08 (0.00)	1.70 (0.00)	1.19 (0.10)
MAE	1.31 (0.47)	**0.69** (0.06)	0.81 (0.05)	0.89 (0.00)	1.70 (0.00)	1.48 (0.00)	0.84 (0.07)

Table 3: Results for Case 2 (standard errors in parentheses)

	SCtMCMC$^{-\text{pw}}$	tLA$^{-\text{pw}}$	HuberMCMC$^{-\text{pw}}$	HuberLA$^{-\text{pw}}$	RCGP$^{-\text{pw}}$	GP$^{-\text{pw}}$	LaplaceMCMC$^{-\text{pw}}$
			$\varepsilon \sim \mathcal{N}(0.01, 0.08)$				
RMSE	3.45 (1.78)	1.17 (0.06)	1.28 (0.02)	1.18 (0.05)	1.99 (0.003)	1.68 (0.02)	**1.13** (0.11)
MAE	1.28 (0.50)	**0.67** (0.06)	0.87 (0.02)	0.85 (0.04)	1.74 (0.002)	1.45 (0.02)	0.79 (0.07)
			$\varepsilon \sim$ Student-$t(10)$				
RMSE	2.43 (1.49)	1.21 (0.15)	1.24 (0.00)	**1.15** (0.09)	1.99 (0.00)	1.72 (0.00)	1.17 (0.10)
MAE	1.06 (0.45)	**0.76** (0.14)	0.91 (0.00)	0.83 (0.06)	1.75 (0.00)	1.47 (0.00)	0.82 (0.07)
			$\varepsilon \sim$ Laplace$(0, 0.1)$				
RMSE	3.53 (1.69)	1.19 (0.10)	1.24 (0.00)	**1.14** (0.10)	1.99 (0.00)	1.72 (0.00)	1.17 (0.07)
MAE	1.33 (0.49)	**0.74** (0.10)	0.90 (0.00)	0.82 (0.06)	1.75 (0.00)	1.47 (0.00)	0.82 (0.06)
			$\varepsilon \sim$ Student-$t(1)$ (Cauchy)				
RMSE	2.53 (1.82)	1.13 (0.04)	1.21 (0.03)	**1.08** (0.09)	2.00 (0.00)	1.71 (0.00)	1.15 (0.08)
MAE	1.04 (0.51)	**0.66** (0.05)	0.88 (0.02)	0.78 (0.07)	1.76 (0.00)	1.47 (0.00)	0.81 (0.06)

Table 4: Results for Case 4 (standard errors in parentheses)

and for each component f_i, the derivative becomes:

$$h_i(f_i) = \frac{-(y_i - f_i)}{\sqrt{1 + (y_i - f_i)^2}} - v_i,$$

where v_i represents the i^{th} component of $\mathbf{v} = \mathbf{K}^{-1}\mathbf{f}$. The term $\frac{-(y_i - f_i)}{\sqrt{1+(y_i-f_i)^2}}$ is strictly monotonic in f_i, as its derivative is positive for all f_i. Its behavior at the limits is given by:

$$\frac{-(y_i - f_i)}{\sqrt{1 + (y_i - f_i)^2}} = \begin{cases} 0 & \text{if } f_i \to y_i, \\ -1 & \text{if } f_i \to \infty. \end{cases}$$

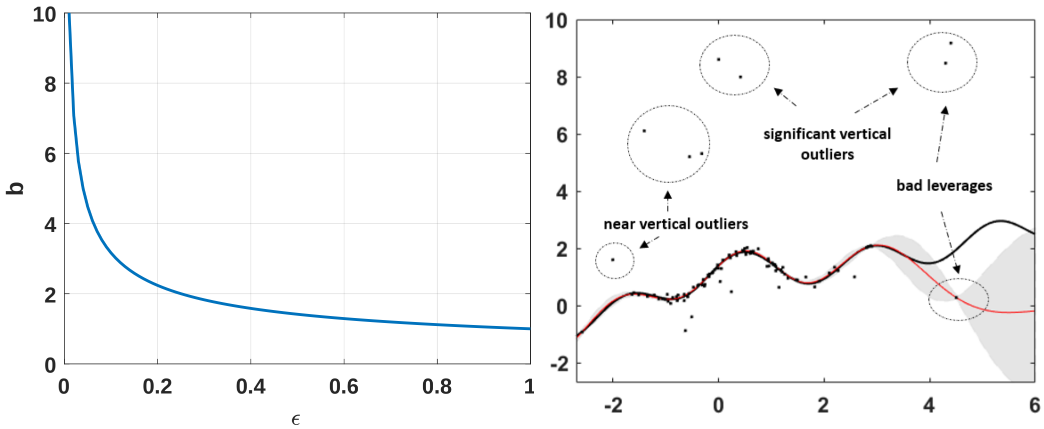
The second term, v_i, arises from the precision matrix $\mathbf{K}^{-1}$, which is symmetric and positive definite. By the spectral theorem, $\mathbf{K}^{-1}$ can be diagonalized as $\mathbf{K}^{-1} = \mathbf{Q}\Lambda\mathbf{Q}^\top$, where Λ is the diagonal matrix of positive eigenvalues, and $\mathbf{Q}$ is an orthogonal matrix. This ensures that v_i is a linear combination of the entries of $\mathbf{f}$ and is therefore continuous and differentiable in f_i. Combining these terms, $h_i(f_i)$ is strictly monotonic because the first term is monotonic and the second term is linear. By the intermediate value theorem, $h_i(f_i)$ crosses zero exactly once because:

$$\lim_{f_i \to y_i} h_i(f_i) = 0 \quad \text{and} \quad \lim_{f_i \to \infty} h_i(f_i) = -1.$$

Thus, $h_i(f_i)$ has a unique root, ensuring that the log-posterior $\log p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \sigma)$ is strictly concave. The strict concavity of the log-posterior implies that the posterior distribution $p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}, \sigma)$ has a unique mode, proving that it is unimodal. □ □

Theorem 2. *Under the same assumptions as Theorem 1, the influence of an individual observation y_i on the posterior mean $\mathbb{E}[f(x) \mid \mathbf{y}]$ is bounded, ensuring robustness to outliers:*

$$\left| \frac{\partial}{\partial y_i} \mathbb{E}[f \mid y] \right| \leq \frac{b}{\sigma}.$$



(a) (b)

Figure 1: (a) b vs contamination ε; (b) extreme (significant) outliers, near outliers, and bad leverage points analyzed on the Neal dataset

Proof. The influence function $I(r_i) = -\frac{\partial}{\partial r_i} \log p(y_i \mid f(\boldsymbol{x}_i))$ of the likelihood $p(y_i \mid f(\boldsymbol{x}_i), \sigma^2)$ for inliers ($|r_i| \leq b$) is:

$$I(r_i) = \frac{r_i}{\sigma^2},$$

and for outliers ($|r_i| > b$) becomes:

$$I(r_i) = \frac{\delta \,\mathrm{sign}(r_i)}{\sigma}.$$

We observe that the $I(r_i)$ is bounded i.e. the contribution of any individual observation y_i to $\mathbb{E}[f(x) \mid \mathbf{y}]$ is bounded, ensuring robustness to outliers. $\square$

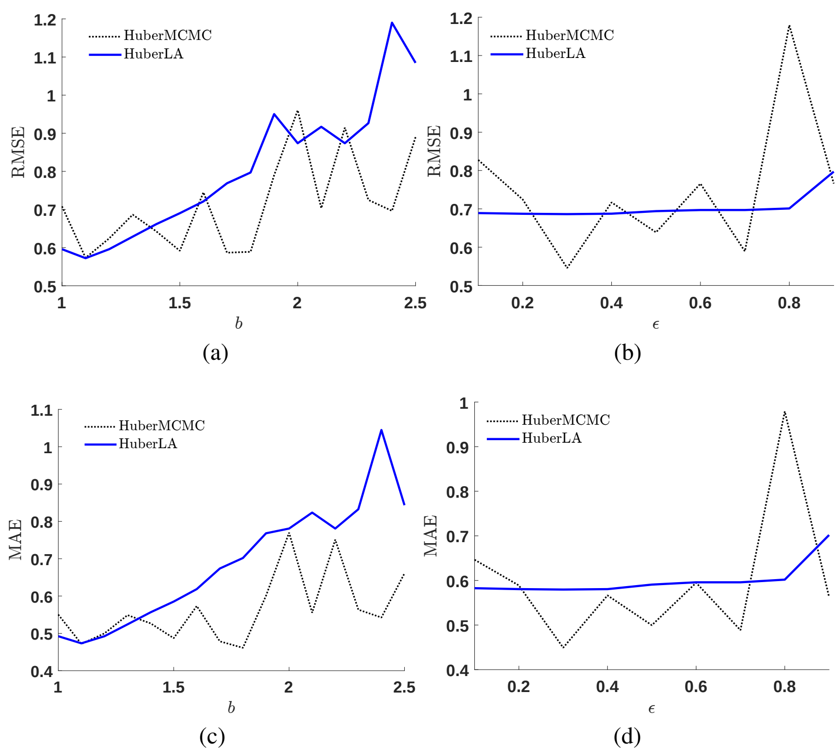


Figure 2: Impact of b and ε on predictions for Neal dataset (case 1 with Student-t noise): (a) RMSE vs b ($\varepsilon = 0.45$); (b) RMSE vs ε ($b = 1.5$); (c) MAE vs b; (d) MAE vs ε.