# Automatic Robust Regression Analysis of Fusion Plasma Experiment Data based on Generative Modelling

K. Fujii[1], C. Suzuki[2], and M. Hasuo[1]

[1]Department of Mechanical Engineering and Science, Graduate School of Engineering, Kyoto University, Kyoto 615-8540, Japan

[2]National Institute for Fusion Science, Gifu 509-5292, Japan

## Introduction

The first step to realize an automatic data analysis for fusion plasma experiment is automatically fitting noisy data measured routinely. A textbook example of fitting procedures is the minimization of the squared difference between the measured data and some parameterized functions such as polynomial. This model implicitly assumes that both the noise distribution and the latent function form are already known, however, it is always not the case for the real world data analysis.

Figure 1 shows a typical example of the Thomson scattering electron density data meaured for Large Helical Device plasma. The fit result with the least squared method is almost meaningless (Fig.1 (b)), because of the above deviation. In particular the noise on the data is clearly non-Gaussian, while the least squares method assumes homoscedastic Gaussian noise.

Many algorithms have been proposed to fit the data with non-Gaussian noise more robustly, e.g. Huber regression. These methods work much better than the least squares, as shown in Fig.1 (c) and (d), but they sometimes show the under- or over-fitted results, i.e. they neglect significant pattern on the data as noise or vice versa. Therefore some human supervision has been always necessary.

In this work, we point out that such the unstableness is caused by the deviation between the probabilistic model that is implicitly assumed in the fitting analysis and the true data distribution. For the purpose of finding the most robust and accurate fitting algorithm, we optimize a model so that its generative model fits the data distribution.

## Theoretical perspective of the robust analysis

Based on Bayesian statistics, the goodness of a model $\mathscr{M}$ for particular ($k$-th) data $\mathbf{y}^{(k)}$ can be measured by the merginal likelihood,

$$p(\mathbf{y}^{(k)}|\mathscr{M}) = \int p(\mathbf{y}^{(k)}|\theta,\mathscr{M})p(\theta|\mathscr{M})d\theta \qquad (1)$$
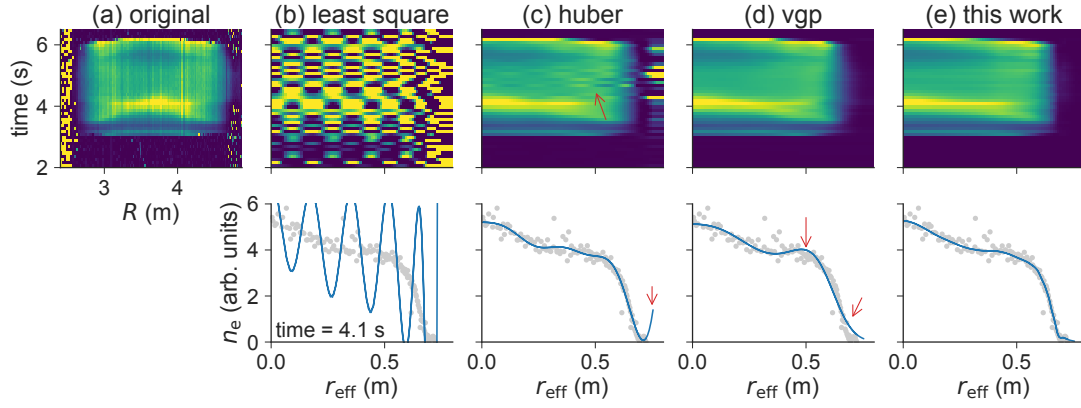
Figure 1: *(a) A typical electron density data obtained by Thomson scattering systm for LHD plasma, as a function of time and the major radius (R). A fit results by several methods are shown in (b)–(e). The top figures show the results as a spatio-temporal evolution, while the bottom figures show the data (gray markers) and the fit results (solid curves) for a single frame. (b) results by the least squares method with Gaussian basis, (c) results by the Huber regression method with Gaussian basis. (d) results by the variational Gaussian process regression with student t-distribution, (e) results by this work.*

where, $p(\mathbf{y}^{(k)}|\theta)$ is likelihood of data $\mathbf{y}^{(k)}$ with given fitting parameter $\theta$. Noise distribution and the latent function form are implicitly included in the likelihood and the prior distribution $p(\theta|\mathcal{M})$.

The robustness of the model $\mathcal{M}$ against the data that will be obtained in the future, $\mathbf{y}^{(k+1)}$, $\mathbf{y}^{(k+2)}$, $\mathbf{y}^{(k+3)}$, ..., is measured by an expectation of this merginal likelihood,

$$\lim_{N \to \infty} \frac{1}{N}\left[\sum_{i=1}^{N} \log p(\mathbf{y}^{(k+i)}|\mathcal{M})\right] = \int p(\mathbf{y}) \log p(\mathbf{y}^{(k+i)}|\mathcal{M}) d\mathbf{y} = -\mathrm{KL}\left[p(\mathbf{y})||p(\mathbf{y}|\mathcal{M})\right] + \mathrm{const.}$$

(2)

where $p(\mathbf{y})$ is the true distribution of $\mathbf{y}$ that will generate data in the future. This expectation is identical to the minus of Kullback-Leibler divergence between the true data distribution $p(\mathbf{y})$ and the modeled data distribution $p(\mathbf{y}|\mathcal{M})$, where Kullback-Leibler divergence $\mathrm{KL}[q(x)||p(x)] = \int q(x) \log \frac{q(x)}{p(x)} dx$ is a measure of the distance between the two probability distributions $q(x)$ and $p(x)$. In other words, the data distribution by the best model $p(\mathbf{y}|\mathcal{M}^{\mathrm{best}})$ is closest to the true data distribution.

## Model

A method we propose here is to construct a flexible generative model, i.e. the latent function form, the prior distribution, and the noise distribution, with neural networks and optimize their weights to fit our generative model to the true data distribution estimated from the large amount

of the existing experimental data. We applied this method to Thomson scattering data in Large Helical Device.

For the latent function form, we assume the following generalized linear combination,

$$f(r, \theta^{(k)} | \mathcal{M}) = \exp\left[\sum_i^M c_i^{(k)} \phi_i(r | \mathcal{M})\right], \tag{3}$$

where $c_i^{(k)}$ is coefficients of the linear summation, which is specific for the measurement data $k$. On the other hand, $\phi_i(r | \mathcal{M})$ is the basis function that is common for all the data, which is expressed by a neural network.

For the prior distribution of $\theta_i^{(k)}$, we adopted the independent Gaussian distribution with variance 1,

$$p(\theta_i | \mathcal{M}) = \mathcal{N}(\theta_i | \mu_i, 1) \tag{4}$$

where $\mu_i$ is the mean of the prior distribution, which also belongs to $\mathcal{M}$.

We assume the noise distribution by the following form of the likelihood,

$$p(y_i | f, \Theta^{\text{lik}}) = \mathcal{S}t\big(y_i | f, \sigma(f | \mathcal{M}), \nu(f | \mathcal{M})\big), \tag{5}$$

where $\mathcal{S}t(x | \mu, \sigma, \nu)$ is student's $t$-distribution for random variable $x$ with mean $\mu$, scale parameter $\sigma$ and the degree of freedom $\nu$. For modeling the heteroscedasticity of the noise, we assume that both the scale parameter and the degree of freedom depends on the true function value $f$. Since the forms of $\sigma(f | \mathcal{M})$ and $\nu(f | \mathcal{M})$ are unknown, we model them by neural networks, where their weights are also regarded as the model prameters.

We optimized these model parameters, i.e. the neural network weights in the latent function form and the likelihood, and the mean of the prior distribution, so that they maximize Eq. 2.

## Results

Figure 2 shows the first four basis functions, that are estimated from the optimization. Three basis functions ($\phi_0$, $\phi_2$, $\phi_3$) have different spatial resolutions at $r < 0.6$, where $\phi_0$ is most smooth, while $\phi_3$ has an oscillation structure. Sharp structures around $r \approx 0.7$ m can be seen in these basis functions. They might represent the intrinsic magnetic island structure in LHD. $\phi_1$ has a steep gradient at $r > 0.7$ m, which is used to represent the steep gradient near the last closed flux surface.
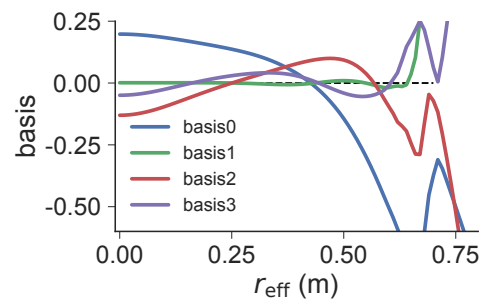


Figure 2: *First four basis functions $\phi(r)$ estimated from LHD Thomson scattering data.*

All the basis functions have a low spatial resolution in the core region ($r_{\text{eff}} < 0.6$ m) while have a high resolution in the edge region ($r_{\text{eff}} \approx 0.65$ m).

Figure 3 shows typical fitting results for a testing data. Black curve shows the median of the posterior distribution of latent functions, while the dark and light shadows show the 65% and 95% region of the likelihood. Note that with Gaussian distribution, the 95% region is nearly two times wider than the 65% region, while with student-$t$ distribution with less degree of freedom, the difference becomes larger.

The likelihood width is estimated to have a dependence on the latent function values, i.e. for the width is large in the low temperature and density region ($r_{\text{eff}} > 0.7$ m) also in the very high temperature region ($T_{\text{e}} > 5$ keV). This dependence is consistent with the characteristics of the Thomson scattering instruments. Most of the training data stay whithin the 95% region. It can be seen that the region corresponding to the outlying points is much larger than the ordinary points.

In Fig. (d), we also showed the fit result by our method. Our method is most robust against the outliers, but still keeping the detailed structure in the data.

## Summary

We have developed a method to learn a robust regression algorithm from the data. We demonstrated that our model trained in this work outperforms other conventional robust analysis methods in terms of the robustness and the accuracy.

The model developed and trained in this work was already integrated in the LHD automatic analysis system. When a new Thomson scattering data arrives, our program runs automatically and provides the fitting result based on the trained model. The results are being used by more than 80 other automatic analysis programs in LHD.
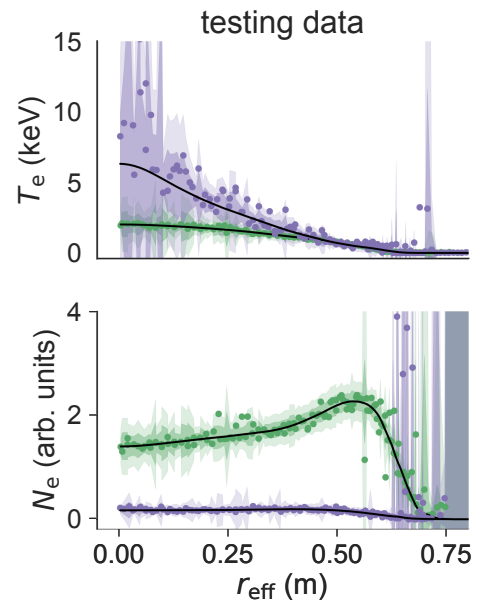


Figure 3: *Typical data of $T_{\text{e}}$ and $N_{\text{e}}$ data measured by LHD Thomson scattering system. Our fit results are shown by solid curves. Shaded regions indicate the coverage probabilities of the likelihood, where light and dark regions indicate 95% and 65% coverage probability.*