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

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

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 frequently not the case for the real world data analysis. Using the conventional model in such situatiln easily results in over- or under-fitting, and therefore some human supervision has been usually necessary. In this work, we propose to optimize a model itself to stabilize the 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,

$$p(\mathbf{y}^{(k)}|\mathscr{M}) = \int p(\mathbf{y}^{(k)}|\theta,\mathscr{M})p(\theta|\mathscr{M})d\theta \tag{1}$$

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

The robustness of the model $\mathscr{M}$ might be measured by an expectation of this merginal likelihood, $\mathbb{E}_{p(\mathbf{y})}[\log p(\mathbf{y}|\mathscr{M})]$, where $p(\mathbf{y})$ is the true distribution of $\mathbf{y}$ that will generate data in the future. We show that the maximization of this expectation is identical to the minimization of Kullback-Leibler divergence between the true data distribution $p(\mathbf{y})$ and the modeled data distribution $p(\mathbf{y}|\mathscr{M})$, and therefore the unbiased generative modeling is essential.

A strategy we propose here is to construct a flexible generative model, i.e. the latent function form and the noise distribution, with neural networks and optimize their weights to fit our generative model to a large amount of data. We applied this strategy to Thomson scattering data in Large Helical Device and found that our model outperforms the conventional analysis methods that does not take into account the data distribution, especially in terms of the robustness.