Bayesian Soft X-Ray Tomography using Non-stationary Gaussian Processes

Dong Li[1], J.Svensson[1], H.Thomsen[1], F.Medina[2], A. Werner[1], R. Wolf[1]

[1]*Max Planck Institute for Plasma Physics, Teilinstitut D-17491 Greifswald,Germany*
[2]*Asociación EURATOM-CIEMAT, Madrid, Spain*

In this study, a Bayesian based non-stationary Gaussian Process method for the inference of soft X-ray emissivity distribution along with its associated uncertainties, has been developed. For the investigation of equilibrium condition and fast magnetohydrodynamic (MHD) behaviors in nuclear fusion plasmas, it is of importance to infer, especially in the plasma center, spatially resolved soft X-ray profiles from a limited number of noisy line integral measurements. For this ill-posed inversion problem, Bayesian probability theory can provide a posterior probability distribution over all possible solutions under given model assumptions. Specifically, the use of a non-stationary Gaussian Process to model the emission allows the model to adapt to the varying length scales of the underlying diffusion process. In contrast to other conventional methods, the prior regularization is realized in a probability form which enhances the capability of uncertainty analysis, in consequence, scientists who concern the reliability of their results will benefit from it. Under the assumption of normally distributed noise, the posterior distribution evaluated at a discrete number of points becomes a multivariate normal distribution whose mean and covariance are analytically available, making inversions and calculation of uncertainty fast. Additionally, the hyper-parameters embedded in the model assumption can be optimized through a Bayesian Occam's Razor formalism and thereby automatically adjust the model complexity. This method is shown to produce convincing reconstructions and good agreements with independently calculated results from the Maximum Entropy (MaxEnt) and Equilibrium-Based Iterative Tomography Algorithm (EBITA) methods.

I. Introduction

In plasma diagnostics the analysis of soft X-ray radiation is a very useful way to explore transient MHD phenomena. Since the plasma is optically thin for soft X-ray radiation, and the radiation can be recorded with high sampling frequency, the tomographic inversion from a set of sight lines can resolve mode structures, rotating perturbations, disruptions etc., even in the plasma center. A typical setup consists of pinhole cameras with photon-diode detectors and filters, which are opaque for visible and infrared wavelengths. For higher energies in the hard X-ray and gamma-range which are mostly filtered out by a thin Beryllium coil in front of each detector, the sensitivity of the detectors strongly decreases. The measured quantity is thus integrated in the spectral range as well as along the sight lines.

The subject of this work is to infer a most probable reconstruction among a manifold of conceivable solutions through a number of noisy line integrated signals. Historically, the techniques for tomographic inversions for plasma physics applications started from a standard Abel inversion[1] method which was used for circular plasma cross sections in tokamaks. Other methods based on linear least squares techniques[2] and restricted Fourier analysis[3-6] have been developed for asymmetrically elongated plasmas. The EBITA[7] and MaxEnt[8] methods with which our results will be compared, are based on numerical iterative algorithms and able to recover structures with localized perturbations from high harmonics. However, MaxEnt has the drawback of high computation time cost due to the iterative and nonlinear numerical techniques. Though EBITA can overcome the drawback of high computation time, it uses additional information about the toroidal magnetic flux surfaces, which need to be derived from equilibrium calculations.

In this paper we demonstrate a Bayesian based non-stationary Gaussian Process (GP) tomography technique[9], to reconstruct the soft X-ray emissivity distribution from a number of noisy line integral measurements. Most applied numerical inversion techniques are based on representing the unknown emissivity function $f(\bar{r})$ ($\bar{r}$ being the spatial coordinate) as some parametric functions, (e.g. linear or polynomial) whose parameters will then be optimized by minimizing a combination of a misfit function and a regularizer. In contrast, the approach described in the following, realizes a non-parametric model by using a Gaussian Process[10,11] to represent a prior probability over the underlying function $f(\bar{r})$. The regularization of $f(\bar{r})$ is directly controlled by the properties of this Gaussian Process prior. Once the measurements are attained, this prior is updated to a posterior probability through multiplying with the likelihood for the measured data. Here we will assume a Gaussian error on the measurements, which gives a multivariate normal

likelihood model. The maximum of the posterior probability distribution provides the single most likely reconstruction, and the posterior covariance gives the uncertainty of this solution.

This non-stationary GP tomography method does not involve nonlinearity or numerical iterations, which makes a real time application feasible. An estimation of uncertainties of the solution is readily available from an analytic posterior. This uncertainty will depend on the errors of the measurements, the coverage of sight lines and the prior model assumption.

II. Method

2.1 Soft X-ray diagnostics in W7-AS

For the stellarator Wendelstein 7-AS (W7-AS) which was in operation until 2003[12], a soft X-ray imaging system, consisting of eight pinhole cameras, each containing a Silicon photon detector, was used to measure the emissivity within a poloidal cross section. The measured value of a line integrated signal is proportional to the number of photons collected within a solid angle subtended by one detector. The experimental setup of this diagnostic system is illustrated in FIG.1.

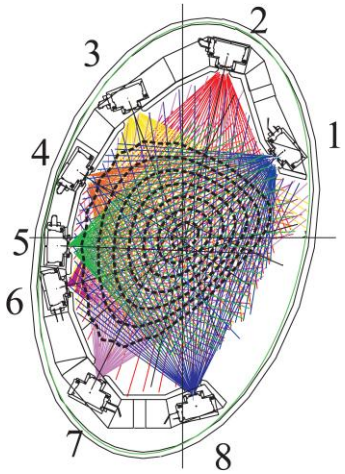

FIG.1. A schematic view of the miniature soft X-ray system (MiniSoX diagnostic system) in Wendelstein 7-AS shows the eight compact detector arrays with a total of 256 sight lines in one poloidal cross-section, achieving a substantial coverage[13,14]. The dashed black lines indicate the magnetic flux surfaces of a typical plasma in Wendelstein 7-AS.

In the following, the emissivity distribution is expressed as a function $f(\vec{r})$ over the 2D poloidal cross section. The actual data d_l obtained from one detector element (indexed by l) is the predicted data d_l^p plus a noise term. The mapping between $f(\vec{r})$ and predicted data d_l^p is given by:

$$d_l^p = c_l \int_{S_l} ds \cdot f(\vec{r}), \quad l=1,2,...,M \tag{1}$$

where, the M integrals are carried out along the paths S_l, of the sight lines. The calibration factors c_l relates to the slight differences in spectral efficiencies and solid angles between the detectors.

To calculate the integral of Eq.(1), the function domain is subdivided into a number of discrete elements. The emissivity area is chosen so as to cover the plasma region in case of plasma shifts and expansions. For the following we have used $N=30\times30=900$ rectangular cells, assuming homogeneous emission in each cell, whose size should ensure to be small enough to justify a constant emissivity within each cell. The reconstruction results won't be sensitive to the number of cells if the size of each cell is guaranteed to be comparable to the resolution of the realistic emissivity distribution. The emissions from the pixels around the limiting wall are set to be zero, as a boundary condition. With this discretization, the matrix formulation of Eq.(1) can be written as:

$$\vec{d}_M = \overline{\overline{R}}_{M\times N} \cdot \vec{f}_N \tag{2}$$

where, the column vector $\vec{f}_N$ is the discretized emissivity function $f(\vec{r})$ and the contribution matrix $\overline{\overline{R}}_{M\times N}$ arises from the forward calculation, whose element R_{lk} is the contribution from a unit emission in cell k to the calibrated measurement l. Since the number of unknowns N is much larger than M, a direct inversion would suffer from the problems of existence and uniqueness of the solutions[15]. A possible solution might also be very sensitive to small changes in the measured data. In the formulation of Bayesian probability theory, these problems are solved by expressing regularizing assumptions as a prior probability density function (pdf), which in our case is realized through a Gaussian Process.

2.2 Gaussian Process

A Gaussian Process is a generalization of a multivariate normal distribution over functions, and is defined by a mean function, $\mu(\vec{r})$ and a covariance function, $k(\vec{r}_i,\vec{r}_j)$. $k(\vec{r}_i,\vec{r}_j)$ defines the covariance between function values at any two locations $\vec{r}_i,\vec{r}_j$ and controls the properties (e.g. smoothness, differentiability etc.) of random sample paths of $f(\vec{r})$ under the process. The Gaussian Process thus fully describes the properties of the underlying function (as a mean and a regularizing covariance function), and no further parameterization of the underlying function is necessary. The pdf of $f(\vec{r})$ over any discrete set of locations will be a multivariate normal

distribution with its mean, $E(f(\vec{r}_i))=\mu(\vec{r}_i)$ and covariance function, $\mathrm{cov}(f(\vec{r}_i), f(\vec{r}_j))=\mathrm{k}(\vec{r}_i,\vec{r}_j)$.

The choice of a good covariance function is crucial for the Gaussian Process inversion since it determines how the regularization is imposed on the underlying function. For the case when the smoothness of the underlying function can be assumed to be the same everywhere, a stationary (non position-dependent) covariance function can work effectively. Two widely used stationary covariance functions are the so called squared exponential (Eq.(3)), and the Matérn covariance function (Eq.(4)). For such stationary covariance functions, the covariance is dependent only on the distance between two locations, $d_{ij}=\left\|\vec{r}_i-\vec{r}_j\right\|$. The argument l in those covariance functions corresponds to a length scale of spatial variation of the function $f(\vec{r})$. σ_f controls the variance/amplitude of the function at a given location.

$$k_{SE}\left(d_{ij}\right)=\sigma_f^2\exp\left(-\frac{d_{ij}^2}{2l^2}\right),\quad d_{ij}=\left\|\vec{r}_i-\vec{r}_j\right\| \tag{3}$$

$$k_{\mathrm{Matern}}\left(d_{ij}\right)=\sigma_f^2\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}d_{ij}}{l}\right)^{\nu}K_{\nu}\left(\frac{\sqrt{2\nu}d_{ij}}{l}\right) \tag{4}$$

where, Γ is the gamma function and K_ν is a modified Bessel function. ν and l are non-negative parameters of the covariance function. The limitation of the stationary covariance functions arises when the smoothness of the underlying function varies between different locations. In this case, a non-stationary covariance becomes necessary and can be made to locally adapt to the varying smoothness, as described in section 2.3. The construction and usage of various stationary and non-stationary covariance functions can be found in various fields, such as machine learning[16], geophysics (kriging)[17] etc. For our work, a non-stationary covariance function has been applied to adapt to the varying smoothness from a diffusion process of the emissivity distribution from plasmas. This is accomplished through a specification of locally adaptive length scales that are inferred by an independent latent Gaussian Process over the local length scales.

2.3 A non-stationary covariance function

To address the problem of varying smoothness in an emissivity distribution, the following non-stationary extension of the squared exponential covariance function has been used[18,19]

$$k(\vec{r}_i,\vec{r}_j)=\sigma_f^2\left|\overline{\overline{\Sigma}}(\vec{r}_i)\right|^{1/4}\left|\overline{\overline{\Sigma}}(\vec{r}_j)\right|^{1/4}\left|\frac{\overline{\overline{\Sigma}}(\vec{r}_i)+\overline{\overline{\Sigma}}(\vec{r}_j)}{2}\right|^{-1/2}\cdot\exp\left[-(\vec{r}_i-\vec{r}_j)^T\left(\frac{\overline{\overline{\Sigma}}(\vec{r}_i)+\overline{\overline{\Sigma}}(\vec{r}_j)}{2}\right)^{-1}(\vec{r}_i-\vec{r}_j)\right] \tag{5}$$

where, $\overline{\overline{\Sigma}}(\vec{r}_i)$ is a 2D matrix describing the local length scales (and possible local correlations) of the function at location $\vec{r}_i$. Under an isotropic assumption, we have a diagonal $\overline{\overline{\Sigma}}(\vec{r}_i)$:

$$\overline{\overline{\Sigma}}(\vec{r}_i)=\begin{bmatrix}l_R^2(\vec{r}_i) & 0\\ 0 & l_Z^2(\vec{r}_i)\end{bmatrix} \tag{6}$$

where, $l_R^2(\vec{r}_i), l_Z^2(\vec{r}_i)$ are the square of the local length scales along R and Z directions respectively. We here have equal length scales, $l_R^2(\vec{r}_i), l_Z^2(\vec{r}_i)=l^2(\vec{r}_i)$ in both directions (isotropic assumption). In principle, an anisotropic assumption can be implemented by adjusting the non-diagonal elements of the matrix in Eq.(6). For example, a conventional regularization used in plasma tomography is to regularize the emission constant along the magnetic flux surfaces, which is particularly essential and helpful when available measurements are too few. Here, a similar regularization can be realized by manipulating the matrix in Eq.(6) to set the length scale along the flux surface much smaller than in the perpendicular direction, so that the diffusion will be much stronger along the contour flux surfaces. However, such a kind of constrain often appears to be over strong and even incurs biased reconstructions when it conflicts with the reality, so in this work with sufficient lines of sight, we relax this artificial constraint. From the exponential part of Eq.(5), the covariance between two locations is thus calculated from the average of two local matrices at locations $\vec{r}_i$ and $\vec{r}_j$. In this way, the covariance is location dependent and the local characteristics at both locations influence the overall modeled covariance. As will be described later (section 2.3), we want the local length scales to adapt to the measurements, so we need to find a low-dimensional representation of the length scales as a function of position. We do this (see ref.18) by modeling the underlying length scales as another, secondary/latent, Gaussian Process. Two Gaussian Processes are thus used together: one latent stationary process GP_l to model the local length scales, and a second non-stationary process GP_f for the emissivity function $f(\vec{r})$, using the first process for its local length scales. Each of the these two processes has a number of hyper-parameters associated with it, which can be taken as fixed values, or optimized from the available data, as described further on in section

2.5. For a list of notations used to describe the different parts of these two processes, see Table 1.

Table 1: Notations used in overall processes about the model assumption.

Non-stationary GP over $f(\bar{r})$	GP_f
Hyper-parameters of GP_f	$\bar{\theta}_f = <\sigma_f, \varepsilon>$
Measurement data set	$\bar{d}_M \in \mathbb{R}^M$
Prediction of underlying function	$\bar{f}^* \in \mathbb{R}^N$ at locations $\bar{r}^*$
Latent GP over local length scales	GP_l
Support local length scales	$\bar{l}^s \in \mathbb{R}^P$ at selected locations
Prediction of local length scales	$l^* \in \mathbb{R}^N$ at location $\bar{r}^*$
Hyper-parameters of GP_l	$\bar{\theta}_l = <\sigma_l, l_l, \bar{l}^s>$
Joint hyper-parameters	$\bar{\theta} = <\bar{\theta}_f, \bar{\theta}_l>$

In above table, σ_f is a hyper-parameter included in the prior covariance (see Eq.(5)) and ε is a parameter used to describe the measurement error.

For the length scale process GP_l, we need to specify local length scales, $\bar{l}^s(\bar{r})$ at a number of support positions, and from which we infer the local length scale at any other position using GP_l. These support length scales will be treated as hyper-parameters of the full model, to be optimized (as described in section 2.5) together with all other hyper-parameters of the model (Table 1). For our problem, to keep the dimensionality low, we have specified support length scales only in the central region (l^c), and in the edge region (l^e), each region having a single support length scale, and regions in between are then interpolated through the latent process. The length scale for the latent process has been taken as 0.5m, approximately the size of the emissivity region. Since this process is a GP, the local length scales at any discrete location form a multivariate normal distribution:

$$l(\bar{r}) \sim N(\bar{\mu}(\bar{r}), \bar{\bar{k}}(\bar{r}_i, \bar{r}_j)) \tag{7}$$

where, $\bar{\mu}(\bar{r})$ is the prior mean of the local length scales and the $k(\bar{r}_i, \bar{r}_j)$ is a stationary squared exponential covariance function as in Eq.(3). Eq.(7) can be imagined as an overall distribution of both the support $\bar{l}^s(\bar{r})$ and inferred local length scales $\bar{l}^*(\bar{r})$. To separate these known $\bar{l}^s(\bar{r})$ and unknown $\bar{l}^*(\bar{r})$ quantities explicitly, we decompose them into two sub vectors as the following equivalent form[20] :

$$\begin{bmatrix} \bar{l}^s(\bar{r}) \\ \bar{l}^*(\bar{r}) \end{bmatrix} \sim N\left(\begin{bmatrix} \bar{\mu}^s(\bar{r}) \\ \bar{\mu}^*(\bar{r}) \end{bmatrix}, \begin{bmatrix} \bar{\bar{\Sigma}} & \bar{\bar{\Sigma}}_*^T \\ \bar{\bar{\Sigma}}_* & \bar{\bar{\Sigma}}_{**} \end{bmatrix} \right) \tag{8}$$

where, $\bar{\mu}^s(\bar{r})$ and $\bar{\mu}^*(\bar{r})$ are the mean of $\bar{l}^s(\bar{r})$ and $\bar{l}^*(\bar{r})$. The sub matrices $\bar{\bar{\Sigma}}$, $\bar{\bar{\Sigma}}_*$ and $\bar{\bar{\Sigma}}_{**}$ constitute the compound matrix $\bar{\bar{k}}(\bar{r}_i, \bar{r}_j)$ in Eq.(7). In detail, $\bar{\bar{\Sigma}}_*$ is the covariance between the support and inferred local length scales. Similarly, $\bar{\bar{\Sigma}}_{**}$ is the covariance between the inferred local length scales. With P denoting the number of support positions, and N the number of positions where length scales are to be inferred, the sub matrices can be summarized as follows:

$$\bar{\bar{\Sigma}} = \begin{bmatrix} k(\bar{r}_1, \bar{r}_1) & k(\bar{r}_1, \bar{r}_2) & \cdots & k(\bar{r}_1, \bar{r}_P) \\ k(\bar{r}_2, \bar{r}_1) & k(\bar{r}_2, \bar{r}_2) & \cdots & k(\bar{r}_2, \bar{r}_P) \\ \vdots & \vdots & \ddots & \vdots \\ k(\bar{r}_P, \bar{r}_1) & k(\bar{r}_P, \bar{r}_2) & \cdots & k(\bar{r}_P, \bar{r}_P) \end{bmatrix}_{P \times P} \tag{9}$$

$$\bar{\bar{\Sigma}}_{**} = \begin{bmatrix} k(\bar{r}_1^*, \bar{r}_1^*) & \cdots & k(\bar{r}_1^*, \bar{r}_N^*) \\ \vdots & \ddots & \vdots \\ k(\bar{r}_N^*, \bar{r}_1^*) & \cdots & k(\bar{r}_N^*, \bar{r}_N^*) \end{bmatrix}_{N \times N} \tag{10}$$

From the following formula about conditioning a joint normal distribution[21],

$$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \bar{a} \\ \bar{b} \end{bmatrix}, \begin{bmatrix} \bar{\bar{A}} & \bar{\bar{C}} \\ \bar{\bar{C}}^T & \bar{\bar{B}} \end{bmatrix} \right) \Rightarrow \bar{x} \Big| \bar{y} \sim N\left(\bar{a} + \bar{\bar{C}}\bar{\bar{B}}^{-1}(\bar{y} - \bar{b}), \bar{\bar{A}} - \bar{\bar{C}}\bar{\bar{B}}^{-1}\bar{\bar{C}}^T \right) \tag{11}$$

a conditional normal distribution over the inferred length scales $\bar{l}^*$ can be derived from the joint normal distribution(Eq.(8)) as follows:

$$\bar{l}^* \Big| \bar{l}^s \sim N\left(\bar{\mu}(\bar{l}^*) + \bar{\bar{\Sigma}}_*^T \bar{\bar{\Sigma}}^{-1} (\bar{l}^s - \bar{\mu}(\bar{l}^s)), \bar{\bar{\Sigma}}_{**} - \bar{\bar{\Sigma}}_*^T \bar{\bar{\Sigma}}^{-1} \bar{\bar{\Sigma}}_* \right) \tag{12}$$

With the prior mean values of both the support and inferred length scales $\mu(\bar{l}^s)$ and $\mu(\bar{l}^*)$ are set to be zero, Eq.(12) can be rewritten as:

$$\bar{l}^* \Big| \bar{l}^s \sim N(\bar{\bar{\Sigma}}_*^T \bar{\bar{\Sigma}}^{-1} \bar{l}^s, \bar{\bar{\Sigma}}_{**} - \bar{\bar{\Sigma}}_*^T \bar{\bar{\Sigma}}^{-1} \bar{\bar{\Sigma}}_*) \tag{13}$$

The mean value of the inferred local length scales $\bar{l}^*$ is given by the mean of the posterior normal distribution in Eq.(13):

$$\bar{l}^* = \bar{\bar{\Sigma}}_*^T \bar{\bar{\Sigma}}^{-1} \bar{l}^s \tag{14}$$

Eq.(14) is the expression used to predict the local length scales $\bar{l}^*$ from the support length scales $\bar{l}^s$. When the support local length scales $\bar{l}^s(\bar{r})$ are optimized as hyper-parameters, they are simultaneously used to infer the values of $\bar{l}^*(\bar{r})$. Consequently, all the local length scales are optimized together to attain a distribution of local length scales which is used as the underlying length scales for the Gaussian Process describing the emissivity function.

2.4 Bayesian formulae

For a Bayesian approach to an inversion problem, first a prior need to be constructed and applied to regularize the unknown quantity $\bar{f}_N$, which is realized in this method by the covariance of a multivariate normal probability $p(\bar{f}_N)$. The usage of a parameterized Gaussian process family (such as the squared exponential used in this work) assists in finding a proper regularizer through its formulation of regularizing properties in terms of a priori length scales of the underlying function. As shown in section [2.3], these length scales can furthermore be estimated directly from the data, making it possible to tune this regularizer automatically. In addition, a likelihood distribution $p(\bar{d}_M \,|\, \bar{f}_N)$ acts as a misfit between model predictions and measurements. The combination of the prior and likelihood leads to a posterior probability $p(\bar{f}_N \,|\, \bar{d}_M)$ which accounts for both the prior regularization and the constraints from measurements, according to Bayes formula[22,23]:

$$p(\bar{f}_N \,|\, \bar{d}_M) = \frac{p(\bar{d}_M \,|\, \bar{f}_N) \times p(\bar{f}_N)}{p(\bar{d}_M)} \tag{15}$$

where, the prior $p(\bar{f}_N)$ denotes the probability of a possible solution $\bar{f}_N$. The likelihood $p(\bar{d}_M \,|\, \bar{f}_N)$ includes information from data $\bar{d}_M$ and measures the probability of yielding that data given $\bar{f}_N$. The marginal likelihood $p(\bar{d}_M)$ (also called the model evidence) is a marginalization of the joint distribution $p(\bar{d}_M, \bar{f}_N)$ with respect to $\bar{f}_N$ through a marginal integration:

$$\int p(\bar{d}_M \,|\, \bar{f}_N) \cdot p(\bar{f}_N) df = \int p(\bar{d}_M, \bar{f}_N) df = p(\bar{d}_M) \tag{16}$$

Evaluated at a number of discrete positions, the Gaussian Process prior becomes a multivariate normal distribution, with mean $\bar{m}_f$ and covariance matrix $\bar{\bar{\Sigma}}_f$:

$$p\left(\bar{f}_N \,|\, \sigma_f, \bar{\theta}_l\right) = \frac{1}{(2\pi)^{N/2} \left|\bar{\bar{\Sigma}}_f\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\bar{f}_N - \bar{m}_f\right)^T \bar{\bar{\Sigma}}_f^{-1} \left(\bar{f}_N - \bar{m}_f\right)\right] \tag{17}$$

where, σ_f and $\bar{\theta}_l$ are the hyper-parameters about prior model assumption. $\bar{m}_f$ is here set to be zero, and $\bar{\bar{\Sigma}}_f$ includes the hyper-parameters σ_f and $\bar{\theta}_l$. The regularization of the underlying function is thus imposed through the prior covariance function $\bar{\bar{\Sigma}}_f$ defining the variance and correlation of the function values between any two positions on our cell grid. Because of the numerical approximation of the integrals in Eq.(1), we are only interested in the function values at the grid cell positions, which are the positions where the multivariate normal distribution in Eq.(17) is defined.

The measurement noise is assumed to be independently normally distributed, giving the following likelihood distribution:

$$\begin{aligned} p\left(\bar{d}_M \,|\, \bar{f}_N, \varepsilon\right) &= \frac{1}{(2\pi)^{M/2} \left|\bar{\bar{\Sigma}}_d\right|^{1/2}} \\ &\times \exp\left[-\frac{1}{2}\left(\bar{\bar{R}}\bar{f}_N - \bar{d}_M\right)^T \bar{\bar{\Sigma}}_d^{-1} \left(\bar{\bar{R}}\bar{f}_N - \bar{d}_M\right)\right] \end{aligned} \tag{18}$$

where, $\bar{\bar{\Sigma}}_d$ is a diagonal covariance matrix with each element being the data variance ε^2,which accounts for uncertainties due to measurement errors (see section 3.1). The contribution matrix $\bar{\bar{R}}$ is a $M \times N$ matrix and as said before its elements represent the contributed proportion of the N individual emissivity pixels to one of the M chord measurements. Multiplication of the prior (Eq.(17)) and likelihood (Eq.(18)) leads to a posterior expressed as[24]:

$$\begin{aligned} p\left(\bar{f}_N \,|\, \bar{d}_M, \bar{\theta}\right) &\propto p\left(\bar{d}_M \,|\, \bar{f}_N, \varepsilon\right) p\left(\bar{f}_N \,|\, \sigma_f, \bar{\theta}_l\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\left(\bar{\bar{R}}\bar{f}_N - \bar{d}_M\right)^T \bar{\bar{\Sigma}}_d^{-1} \left(\bar{\bar{R}}\bar{f}_N - \bar{d}_M\right)\right.\right. \\ &\left.\left. + \left(\bar{f}_N - \bar{m}_f\right)^T \bar{\bar{\Sigma}}_f^{-1} \left(\bar{f}_N - \bar{m}_f\right)\right)\right] \end{aligned} \tag{19}$$

where, $\bar{\theta} = \{\sigma_f, \varepsilon, \bar{\theta}_l\}$ is the combination of all the hyper-parameters (Table 1) and need to be optimized as described in the next section.

Eq.(19) can be explicitly rewritten as a multivariate normal distribution:

$$\begin{aligned} &p\left(\bar{f}_N \,|\, \bar{d}_M, \bar{\theta}\right) = \\ &\frac{1}{(2\pi)^{N/2} \left|\bar{\bar{\Sigma}}_f^{post}\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\bar{f}_N - \bar{m}_f^{post}\right)^T \left(\bar{\bar{\Sigma}}_f^{post}\right)^{-1} \left(\bar{f}_N - \bar{m}_f^{post}\right)\right] \end{aligned} \tag{20}$$

with its posterior mean vector,

$$\bar{m}_f^{post} = \bar{m}_f + \left(\bar{\bar{R}}^T \bar{\bar{\Sigma}}_d^{-1} \bar{\bar{R}} + \bar{\bar{\Sigma}}_f^{-1}\right)^{-1} \bar{\bar{R}}^T \bar{\bar{\Sigma}}_d^{-1} \left(\bar{d}_M - \bar{\bar{R}}\bar{m}_f\right) \tag{21}$$

and its posterior covariance matrix,

$$\bar{\bar{\Sigma}}_f^{post} = \left(\bar{\bar{R}}^T \bar{\bar{\Sigma}}_d^{-1} \bar{\bar{R}} + \bar{\bar{\Sigma}}_f^{-1}\right) \tag{22}$$

where, $\bar{m}_f^{post}$ coincides with the maximum posterior (MAP) point, and provides a single most probable solution of $\bar{f}_N$. The uncertainty of the solution is provided by the

covariance matrix, whose diagonal elements can be used as marginal errors of the solution. A direct way to visualize the uncertainties of the result (that includes posterior variance of the reconstruction), is to sample directly from the posterior probability distribution, to get a range of possible solutions of the reconstruction under the posterior. Sampling from a multivariate normal distribution is easily done through a Cholesky decomposition of the covariance matrix:

$$\overline{\overline{\Sigma}}_f^{post} = \overline{\overline{L}}\,\overline{\overline{L}}^T \tag{23}$$

where, $\overline{\overline{L}}$ is a lower triangular matrix. Samples can then be taken through:

$$\overline{m}_f^{post} + \overline{\overline{L}}\,\overline{n} \tag{24}$$

where, $\overline{n}$ is a vector of N independently normal random variables with zero mean and unit variance.

2.5 Bayesian Occam's Razor optimization

The hyper-parameters $\overline{\theta} = \{\sigma_f, \varepsilon, \overline{\theta}_l\}$ in the prior and likelihood have to be optimized in the light of the measurement data. This is accomplished by maximizing the posterior probability over the hyper-parameters conditioned on the measurements. σ_f and $\overline{\theta}_l$ come from the non-stationary prior covariance function (Eq. (5)). The posterior probability over the hyper-parameters $\overline{\theta}$ is given by $p(\overline{\theta}|\overline{d}_M) \propto p(\overline{d}_M|\overline{\theta})p(\overline{\theta})$, which is proportional to $P(\overline{d}_M|\overline{\theta})$ if a flat prior on $\overline{\theta}$ was assumed. $P(\overline{d}_M|\overline{\theta})$ is the marginal likelihood found in the denominator of the right hand side of Bayes rule (Eq.(15)). Since both the prior and the likelihood are multivariate normal distributions, and the model is linear, the integral in Eq.(16) can be carried out analytically and gives another multivariate normal distribution over $\overline{d}_M$ under a given model assumption $\overline{\theta}$, resulting in the following expression for the logarithm of the marginal likelihood term:

$$\log p(\overline{\theta}|\overline{d}_M) \propto \log p(\overline{d}_M|\overline{\theta}) = -\frac{1}{2}\log\left(\left|\overline{\overline{\Sigma}}_d\right|\right) - \frac{1}{2}\left(\overline{d}_M - \overline{\mu}_d\right)^T \overline{\overline{\Sigma}}_d^{-1}\left(\overline{d}_M - \overline{\mu}_d\right) - \frac{M}{2}\log(2\pi) \tag{25}$$

where, $\overline{\overline{\Sigma}}_d$, $\overline{\mu}_d$ are functions of the hyper-parameters. Maximizing this expression with respect to the hyper-parameters $\overline{\theta}$ gives the value of the hyper-parameters that have highest probability in light of the data. This optimization procedure will automatically penalize over complex models (models with small length scales) that would overfit the data. This happens since over complex models are able to explain a larger range of data sets (but the marginal likelihood has to be normalized to 1 in the space of the data as a probability distribution), thus each one having a lower probability than a typical dataset explained by a less complex model. If the model can not fit the data satisfactorily, the probability of that data under the model will also be low, so optimization of the marginal likelihood will result in a tradeoff between model complexity and data fit. In practice, an optimal model assumption is achieved by maximizing the probability in Eq.(25) using a multivariate optimization algorithm e.g. conjugate gradient on this equation to find the optimal values of the hyper-parameters.

III. Performance and results

To assess the performance of our non-stationary GP method relative to other existing methods, we first use simulated data for a benchmark with the standard MaxEnt method[8]. Afterwards, we demonstrate the performance of this method using experimental data from the stellarator devices W7-AS[13] and TJ-II[25].

3.1 Error model

FIG.2 shows time traces of signals from a central and an edge channel from the beginning of the discharge in W7-AS. The variance increases abruptly once the plasma discharge starts, which happens for both channels, even though the low-level edge channel has a mean value close to zero, thus not measuring any emission from the edge. The error model we have used for this work assumes a constant variation of all channels, with a standard deviation given by the hyper-parameter ε. A single constant error level ε is assumed for data from all detectors at the same time slice during the same plasma discharge. To avoid having to perform a full optimization for each time slice in real time application, we have used the optimal standard deviation found through the evidence optimization in Eq.(25) on a large number of pulses and time slices (FIG.3), and compared that with the average signal level. This gives an approximately linear relationship, as shown in FIG.3, between the optimized standard deviation and the average signal level. We can then use this heuristic relationship to read off an approximate optimal error level from the average signal level, and keep the rest of the hyper-parameters constant, to achieve real time speed inversions. The optimal value occurs around 12.5% of the average of the signal strength. Therefore, 12.5% of the average of one data set is used in the following as an approximate value of a most likely standard deviation of each single data set. Note that the large value of 12.5% is due to the significant portion of low-level data among each data set in W7-AS, which also roughly amount to 2.5% of the maximum data.

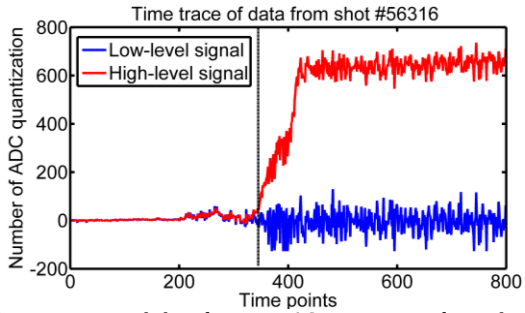

FIG.2: Experimental data from W7-AS: time trace of signals from two different detectors during discharge #56316 shows the variance of data increase abruptly once the discharge starts.

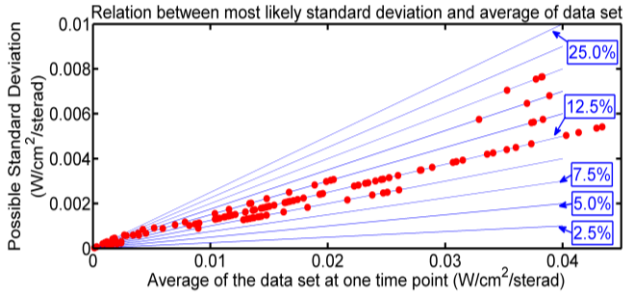

FIG.3: Scatter plot showing the linear relation between the most likely standard deviation of one data set and the average of this data set. The possible values of the standard deviation are chosen to be within a certain maximum percentage level (from 2.5% to 25%) of the average of one data set. The most likely standard deviations of many data sets appear around the 12.5% of the average of data.

3.2 Benchmark using simulated data

The simulated emissivity distribution shown in FIG.4 is based on the magnetic flux surfaces of a standard W7-AS magnetic configuration[26] at the toroidal location where the soft X-ray diagnostic system is installed. The artificial line integral data can be calculated (cf. Fig.5) using the forward model of Eq.(2). A normally distributed random noise with a zero mean and a 12.5% of the average of the data set as standard deviation is added to this artificial data set. This artificial noisy data set is then used as input data for the calculation of reconstructions by the non-stationary GP and MaxEnt methods.

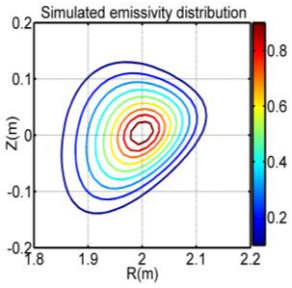

FIG.4: A 2D emissivity distribution in the poloidal plane, where the MiniSoX-Tomography system was located, was simulated based on a typical magnetic configuration in W7-AS.

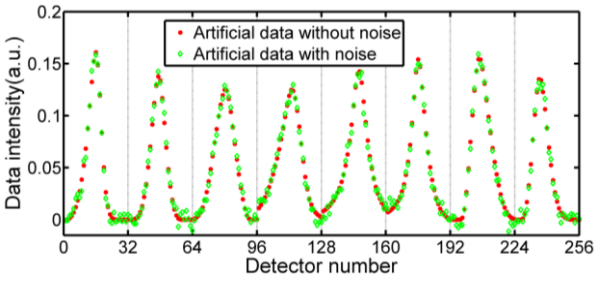

Fig.5: To compare the different inversion methods, the artificial line integral data calculated from a simulated emissivity distribution is taken as input data of the different methods. Red dots: artificial data without errors. Green diamonds: artificial data with independently normally distributed random noise, used as input data.

The 2D distribution of the optimal local length scales from the non-stationary GP (Eq.(14)) is presented in FIG.6. Fig.7, FIG.8 and FIG.9 show the comparison of results obtained by the non-stationary GP and MaxEnt methods. The root-mean-square deviation (RMSD) between the two reconstructions (by non-stationary GP and MaxEnt) and the simulated emissivity distribution are 0.16 and 0.26 respectively, indicating the precision of the overall condition of the two reconstructions concerning both overall magnitude and location. The location of the emissivity peak is well reconstructed with both methods. Also the difference of the shape in the central region is small. However, for the regions with low intensity, the left edge of the MaxEnt reconstruction becomes wiggly and does not coincide with the simulated emission. This is caused by the use of the entropic prior in the MaxEnt method, which assumes that emissivity of neighboring pixels are uncorrelated, leading to noisy reconstructions especially for low signal measurements. In contrast, the better reconstruction of GP case is caused by the non-stationary covariance in the prior, which correctly smoothes the edge of reconstruction by using lager length scales. To investigate the reconstruction more intensively both at center and edge regions, FIG.8 shows 1D profiles of the reconstruction in Fig.7 intercepted at $R = 2.0m$ and $R = 1.9m$. A better agreement of two reconstructed profiles with the simulated profile in FIG.8 (a) than (b) is mainly attributed to the higher measurement density in the center than the edge. The non-stationary GP additionally provides the 95% confidence intervals of the reconstruction shown by the error bars to indicate the reliability of the reconstruction and accounts for the misfits between the inferred reconstruction and simulated emissivity. Since the non-stationary GP uses a boundary condition assuming the emission from the first wall is zero, the derived error bars of the reconstruction around the boundary become accordingly smaller. This shows that the uncertainties of the reconstruction depend both on the prior model assumptions and the quality of measurements. In FIG.8(b), the profile reconstructed by MaxEnt tends to have small positive values at the two edges and misfits the simulated profile, whereas the profile by the non-stationary GP still coincides well enough with the simulated profile.

It is because the role of the prior increases when the density of measurement is low at the edge region, the MaxEnt reconstruction approaches a default model in the prior, keeping small positive values, to which the MaxEnt solution will reduce in the absence of any data. On the contrary, the non-stationary GP makes inference directly on the underlying function and does not necessarily involve a default model which may make the reconstruction deviate from the true emissivity distribution. FIG.9 shows the predicted data from the reconstruction by non-stationary GP with their 95% confidence intervals which reasonably cover the misfits between predicted and artificially noisy data, indicating an appropriate error level has been defined for the input data.

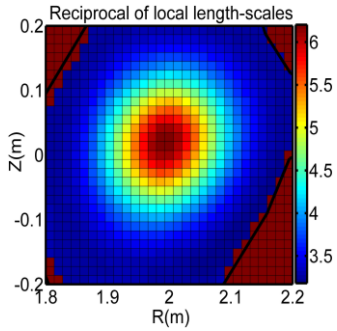

FIG.6: Distribution of the reciprocal of local length scales inferred by a stationary Gaussian Process regression. The black line indicates the boundary of the vacuum vessel.

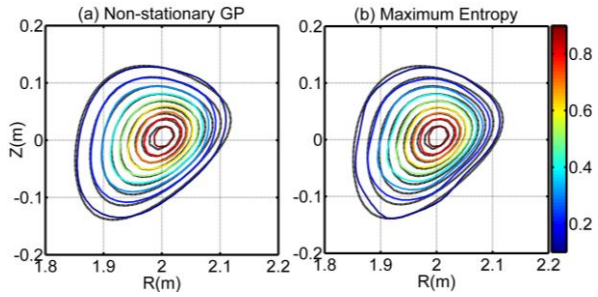

Fig.7: Reconstructions by (a) non-stationary GP and (b) MaxEnt methods using the artificially noisy data, and the errors in both methods are exactly described as how they are added. The black contours show the simulated emissivity distribution for a clear comparison.

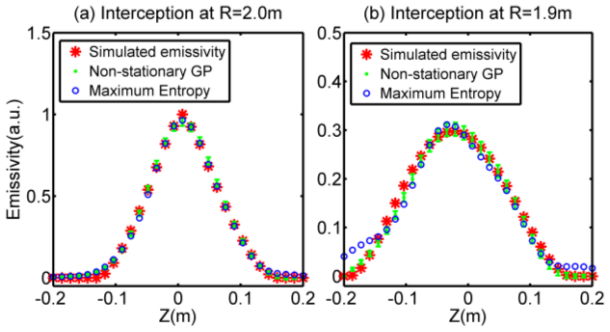

FIG.8: The 1D plots about the profiles intercepted at (a) $R = 2.0m$ and (b) $R = 1.9m$ from the reconstructions by two methods. Red asterisks: simulated emissivity profiles. Green dots: the non-stationary GP reconstruction with 95% confidence intervals. Blue circles: the MaxEnt reconstruction.

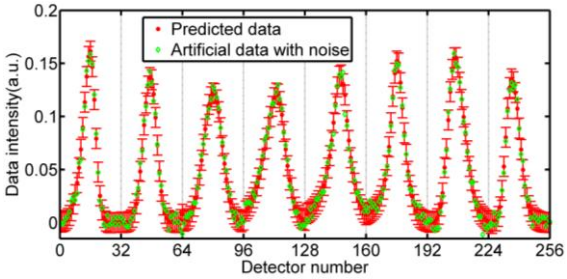

FIG.9: Fit between the data predicted from reconstruction by non-stationary GP and the artificially noisy data. Green diamonds: artificially noisy data used for the inversion of reconstruction. Red dots: predicted data with their 95% confidence intervals.

3.3 Application on W7-AS

To illustrate the application on the W7-AS stellarator, two distinctive data sets with large differences in intensities from two shots are chosen as input data to compare the performance of the two methods. The reconstructions using the first data set with strong intensity are shown in FIG.10. The central regions of both reconstructions have an elliptic shape conforming to the magnetic flux surface. The 3 to 4 cm outward displacement relative to the axis of the flux surface is due to the Shafranov shift. For the edge region of the reconstructions, the non-stationary GP appears to be a smoothly triangular shape which is close to the flux surface, whereas the reconstruction by MaxEnt displays a wiggly boundary as observed in the simulation case. In FIG.11(a), the profiles at $R = 2.0m$ from two reconstructions coincide adequately, however, the discrepancy of the profiles at $R = 1.9m$ in FIG.11 (b) become evident, especially at the right hand side, the profile by MaxEnt even tails up due to the approach to its default model. Since the posterior of the non-stationary GP is a multivariate normal distribution, the spread of 100 samples taken (Eq.(24)) from the posterior can be used to visualize the uncertainties of the reconstruction as shown in FIG.12. In FIG.13 the misfits between predicted data and experimental data are reasonably small and covered by the error bars. The appropriate coverage of the misfits by the error bars also indicates a reasonable error level is defined for the experimental data.

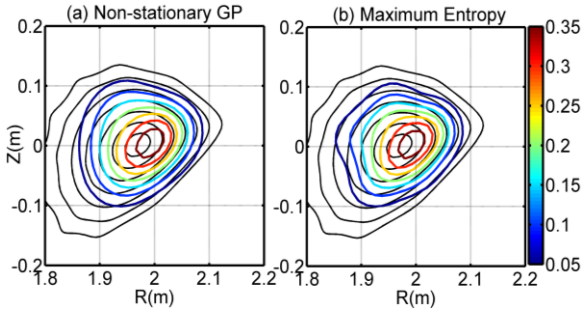

FIG.10: Comparison of the reconstructions from shot number 56316 by (a) non-stationary GP and (b) MaxEnt using the experimental data from W7-AS. The black contours show the flux surface derived from the equilibrium calculation of the vacuum configuration.

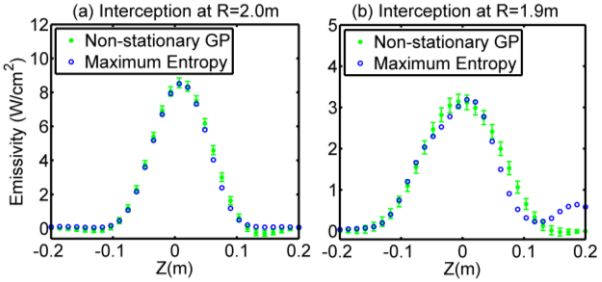

FIG.11: The 1D plots about profiles intercepted at (a) $R = 2.0m$ and (b) $R = 1.9m$ from the reconstructions by two methods using the experimental data. The green dots show the reconstructed profile with 95% confidence intervals given.

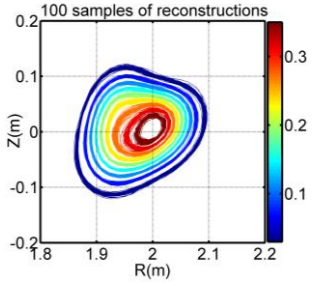

FIG.12: 100 samples of possible reconstructions, drawn from the multivariate normal posterior distribution to visualize the uncertainties of reconstruction.

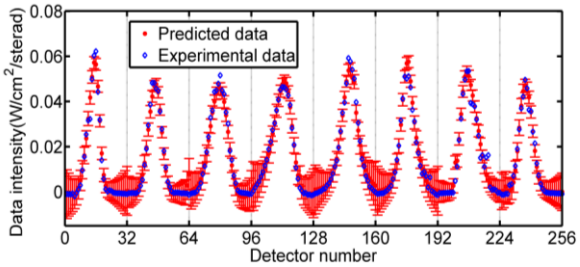

FIG.13: Fit between the predicted data from the reconstruction by non-stationary GP and the used experimental data. Red dots show the predicted data with error bars

To verify the performance of the two methods in face of emissivity distributions with complex structures, here we choose another experimental data from a shot which has complex structures in the center. As FIG.14 shows, the reconstructions from two methods successfully find an m=3 mode structure that distributes symmetrically around the axis of the flux surface.

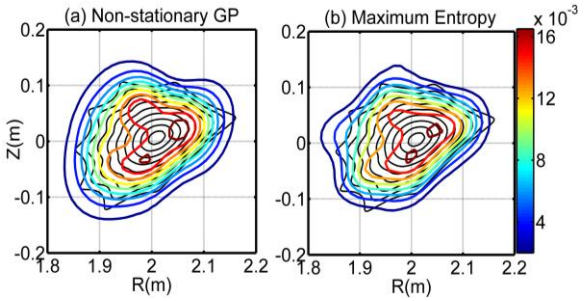

FIG.14: Comparison of the reconstructions by two methods using the experimental data from shot number 53962 at time point 0.3741s. For comparison, black contours show the flux surface derived from the equilibrium calculation of the vacuum configuration.

In W7-AS, the dependence of the maximum achievable thermal/magnetic pressure ratio β on the equilibrium magnetic flux surfaces has been intensively investigated as one experimental issue[27]. Here β , as an indicative parameter, is preferably maximized for higher power production efficiency; the equilibrium flux surfaces are calculated using equilibrium code Variational Moments Equilibrium Code[28] (VMEC), which is a numerical tool widely used for planning experiments and equilibrium analysis. Such a code involves the solution of a set of MHD equations through finding the minimum total energy of the magnetically confined plasma system, so can rapidly solve MHD equilibrium configuration. Since the emission relevant parameters e.g. plasma density, temperature are often assumed to be constant within each contour flux surface, the basic features the reconstructed emissivity distribution will agree well with the equilibrium flux obtained by VMEC, hence the β induced effects on equilibrium flux surfaces can be investigated by tomographic analysis. The reconstructions by non-stationary GP in FIG.15 clearly present an outward shift frequently occurring during the experiments of high β performance[27] and also the consistent structures between reconstructions and equilibrium flux surfaces except a large indentation in the inboard side, which may arise from the movement of the plasma center.

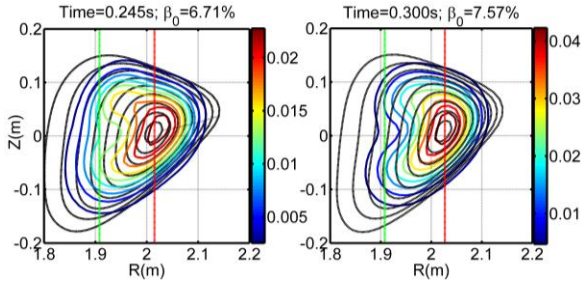

FIG.15: Reconstructions at two different time points with high β_0 in center from shot number 51755 in W7-AS, calculated by the non-stationary GP, shows a large horizontal shift, which also coincide with the equilibrium flux surfaces (black contours) calculated by free boundary

VMEC calculations and also a large indentation frequently occurring in the inboard side. The green and red lines indicate the locations of the magnetic axis of the vacuum and finite β configurations, respectively. Specifically, a strongly inward axis of vacuum configuration is achieved by low magnetic fields and comparably higher vertical fields for high β experiments.

3.4 Application on TJ-II

Another implementation of this non-stationary GP method was carried out at stellarator TJ-II in Spain, which is a medium size stellarator with four periods and a major radius of 1.5 m. It has a helical magnetic axis as a great flexibility in magnetic configuration and a bean-shaped magnetic surface from a combined action of existing magnetic fields. The experimental setup of a soft X-ray diagnostic system, consisting of 5 cameras with 16 detectors each, is illustrated in FIG.16 and a total number of 80 lines of sight passing through strongly bended plasma[25]. The non-stationary GP reconstruction in FIG.17(a) is compared with the EBITA[7] reconstruction in FIG.17(b) with the flux surfaces given by the dashed lines. Both reconstructions coincide with the magnetic flux surface satisfactorily and concentrate in the inner region of the plasma, reaching a good agreement both regarding shape and location. Note that EBITA uses the flux as complementary information for its calculation of the reconstruction.

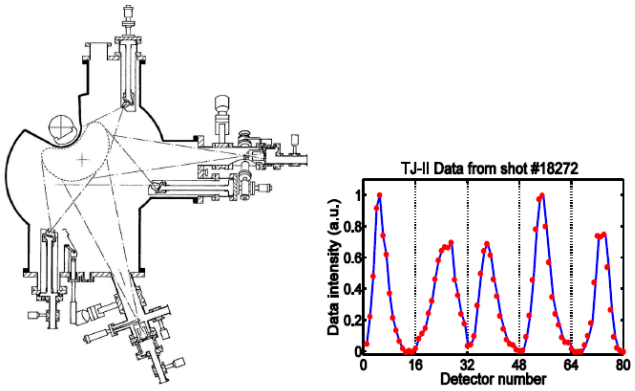

FIG.16: (a) Schematic diagram of the setup of a soft X-ray diagnostic system in TJII, which consists of five detector arrays with each array having 16 detectors. (b) Experimental data from a typical discharge.

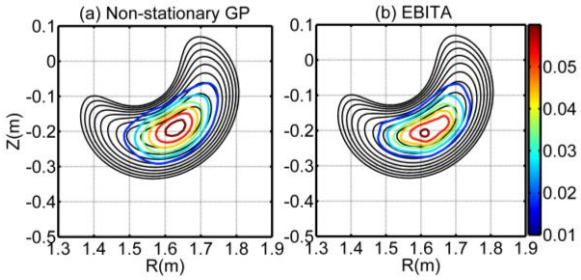

FIG.17: Comparison of the reconstructions by the non-stationary GP and EBITA methods using the experimental data from shot number 18272 in TJ-II. They both have similar shape and location. Black contours show the flux surface derived from the equilibrium calculation of the vacuum configuration.

IV. Discussion

The difference between the non-stationary GP and MaxEnt methods mainly arises from the different ways to impose a prior regularization. The non-stationary GP imposes a prior regularization directly and naturally on the underlying function by defining the correlations between any pair of function values, which is a natural way of describing a diffusion process. In contrast, MaxEnt method uses an entropic prior which regularizes the underlying function by maximizing the total entropy of a number of discrete function values based on the statistical description of the collective and random behavior of many particles, which apparently has different characteristics to a diffusion process. Particularly, the Gaussian Process prior is improved by using a non-stationary covariance function to make the regularization flexible and can be adjusted at different locations by the locally adaptive length scales which determine the extent of correlation depending on the distance. The smooth edge region is accordingly assigned larger length scales and the center is assigned smaller length scales to help recover fine structures.

Additionally, MaxEnt method specifies the underlying function in an exponential from as a default model and successfully ensures the positivity of its reconstructions, here it is found this may distort the gradient of the reconstructed profiles and thus degrade the precision. Bayesian Gaussian Process method of tomographic reconstruction not only seems to be a more suitable model for the soft X-ray inversion problem, but also has the advantages of making the posterior mean analytically without nonlinear iterations, and also provides proper uncertainties on the solution.

V. Summary

The purpose of this work is to develop a method to reconstruct a most probable emissivity distribution with its uncertainties, from a number of noisy chord measurements. Through comparisons with different inversion methods using both simulated and experimental data, our non-stationary GP method produces convincing reconstructions, which is further confirmed by a good agreement between reconstructions from other methods, and also good correspondence with equilibrium flux surfaces. As can be seen from inversions using simulated data, with added noise, the non-stationary GP outperforms MaxEnt and shows a better resistance to severely noisy data. The regularization imposed by a Gaussian Process prior, expressed in correlation length scales, is possibly a more natural way to describe prior assumptions on diffusion processes, than the entropy prior of MaxEnt. The application of a non-stationary GP improves the precision of the reconstructions by using locally adaptive length scales to adapt to the varying smoothness in the emissivity distribution. Additionally, this approach is fast enough for real time applications since it does not involve an iterative or nonlinear computation and does not rely on

complementary information from additional calculations. Finally, the posterior MAP and covariance are both analytically available, the latter giving full uncertainties of reconstructions and predicted reconstructed data.

Acknowledgement

The authors would like to thank Dr. Oliver.P.Ford for his helpful suggestions.

Reference

[1]K. Bockasten, J. Opt. Soc. Am. **51**,943 (1961).
[2]R. Decoste, Rev.Sci.Instrum. **56**, 807(1985).
[3]A.P. Navarro, V.K. Pare and J.L. Dunlap, Rev. Sci. Instrum. **52**, 1634(1981).
[4]R.S. Granetz, J.F. Camacho, Nucl. Fusion **25**, 727(1985).
[5]J.F. Camacho, R.S. Granetz, Rev. Sci. Instrum. **57**, 417(1986).
[6]D. Stratakis,et al. Phys. Plasmas **14**, 120703 (2007)
[7]A.P. Navarro, M.A. Ochando, and A. Weller, IEEE Tans. Plasma Sci. **19**, 569 (1991).
[8]K. Ertl, W.Von der Linden, V. Dose and A. Weller, Nuclear Fusion **36**, 1477 (1996).
[9]J. Svensson, "Non-parametric Tomography Using Gaussian Processes", submitted to IEEE Transactions on imagine processing.
[10]N. Choudhuri, S. Ghosal and A. Roy, Statistical Methodology **4** (2007) 227–243.
[11]C.E.Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning*, (MIT Press, 2006), p.79.
[12]M. Hirsch et al. Plasma Phys. Control. Fusion **50** (2008) 053001 (204pp).
[13]A. Weller, C. Görner and D. Gonda, Rev.Sci.Instrum. **70**, 484 (1998).
[14]C. Görner, Ph.D. thesis, Technische Universität, 1998.
[15]A. Tarantola, *Inverse Problem Theory*, (SIAM, 2005), p.202.
[16]Carl Edward Rasmussen and Christopher K. I. Williams, Gaussian *Processes for Machine Learning*, MIT Press, 2006.
[17]Michael L. Stein, *Statistical Interpolation of Spatial Data: Some Theory for Kriging*, Springer, 1999.
[18]C. Plagemann, K. Kersting, and W. Burgard, '2008 Proceedings of the European conference on Machine Learning and Knowledge in Databases', p.204-219.
[19]C.J. Paciorek and M.J. Scherbish, Advances in Neural Information Processing Systems 16. The MIT Press, 2004.
[20]E. Ebden, "Gaussian Processes for Regression: A Quick Introduction," 2008.
[21]C.E. Rasmussen, *Gaussian Processes in Machine Learning*, (Springer-Verlag, Heidelberg, 2004).
[22]D.S. Sivia, J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, 2006), p.6.
[23]A.Gelman, J.B.Carlin, H.S. Stern, D.B. Rubin and A. Gelman, *Bayesian Data Analysis* (Chapman & Hall/CRC, 2004), p.7.
[24]J. Svessen, A. Werner, Plasma Phys. Control. Fusion **50**, 085002(2008).
[25]F.Medina, L.Rodríguez-Rodrigo, J. Encabo-Fernández, A. López-Sánchez, P. Rodríguez, and C. Rueda, Rev. Sci. Instrum. **70**, 642 (1999).
[26]Joachim E. Geiger, Arthur Weller, et al. Fusion Science and Technology, **46**, P13-23.
[27]A. Weller, et al. Plasma Physics and Controlled Fusion 45 (2003) A285–A308.
[28]S.P. Hirshman, W.I. van RIJ, Computer Physics Communications **43** (1986) 143-155, North-Holland, Amsterdam.