FPGA ACCELERATION OF BAYESIAN MODEL BASED ANALYSIS FOR TIME-INDEPENDENT PROBLEMS

H. Trimino Mora, S. Bozhenkov, J. Knauer, P. Kornejew, S. Kwak, O. Ford, G. Fuchert,
E. Pasch, J. Svensson, A. Werner, R.C. Wolf[1] and D. Timmermann[2]

[1]Max-Plank-Institut für Plasmaphysik
[2]Institut für Angewandte Mikroelektronik und Datentechnik, Rostock University, Rostock, Germany

Abstract—Inverse problems in plasma physics commonly face a trade-off between approximated real-time schemes or post-processed rigorous uncertainty handling.
Bayesian analysis allows parameter and uncertainty estimation as well as joint analysis of multiple diagnostics in a strict mathematical way. It also improves the inference from correlated measurements but with long processing times.
For linear and non-linear problems, many optimal and sub-optimal Bayesian online algorithms are available but generally targeted at dynamic systems and introducing some level of approximation.
Given plasma physics time-independent non-linear inverse problems, several Wendelstein 7-X diagnostics use a Bayesian inference framework. This research focuses on accelerating this type of mathematically intense standard Bayesian analysis for such inverse problems. We show a significant acceleration for the estimation of electron density and temperature profiles. The approach maintains a floating-point double precision while reducing the processing time useful in applications where a reliable error estimation is required together with a fast processing time.

Keywords: Bayesian analysis, data fusion, hardware acceleration, inverse problems.

I. INTRODUCTION

When dealing with inverse problems in modern scientific research experiments it is common to have a problem of dual nature. In the case of parameter estimation there is a need for a fast processing of specific parameters in the studied phenomena for control and safety purposes. Nevertheless the typical aim of the experiment is to perform scientific inference in order to measure and understand phenomena that is not yet completely understood. This introduces the need for an analysis with the least amounts of approximations to understand and observe the behavior of the parameter of interest.

Current parameter estimation has been widely developed to suit specific needs of timing, time dependence and uncertainty estimation. One of the most transcendent current techniques is Bayesian analysis. It provides a rigorous handling of the uncertainty and the simplicity of introducing our knowledge or lack thereof, into the analysis.

For this paper the analysis will be applied to magnetic confinement devices. These devices are used to confine high temperature plasmas by magnetic fields. They are used to study and understand nuclear fusion as a new primary energy source.

Plasma diagnostics are sometimes used only for scientific inference in order to study the physical processes occurring in the plasma and estimate the most important plasma parameters [16]. In other cases they are used to protect the device from damages, control positioning of the plasma and other purposes requiring a real-time system [8].

For the special case of the magnetic confinement device of the stellarator type, its design provides inherent plasma stability that makes rigorous inference more important than fast control systems. The evolution of the plasma parameters and processes tends to be highly non-linear with complex models as well as extensive analytic expressions. It is also the case that variables in these models are also not known with certainty. This makes the niche of plasma physics in a stellarator favorable for standard Bayesian analysis where the least amount of biasing and approximations is preferred [17]. Nevertheless, fast parameter estimation and control for plasma stability is also important for machine protection. This drives the need to have a system that allows unbiased Bayesian inference in a faster time scale than the typical post-processing approach.

Bayesian inference for Wendelstein 7-X (W7-X) diagnostics is done with Minerva, a Bayesian modelling framework that generalises model description and inference [18]. The framework has for example the ability to use a set of optimizers and samplers, reusable for any given model. The framework has the flexibility of changing parameters from "known" to "free parameter". Another key advantage is the joint analysis of multiple diagnostics that are measuring one plasma parameter in common [15]. Nevertheless, Minerva is not designed for real-time capabilities and, therefore, has processing times that are suitable for post processing only.

Finally, it is also important to state that given the lack of knowledge of the dynamic behavior in many plasma processes, this Bayesian analysis is typically done for single observations and not for time-dependent models.

Typically, these types of inverse problems for complex physics models with several data sources and free parameters require a big amount of computation time and processing power. Model dimensionality can also grow to a point where the full posterior distribution is intractable and not of a standard analytic form. For this, various iterative sampling algorithms such as Markov Chain Monte Carlo (MCMC) are used to find proper (generally non-Gaussian) parameter distributions of interest, often taking minutes or hours on regular CPUs.

With this motivation, the question arises of whether it is possible to devise an accelerated form of this standard time-independent non-linear Bayesian analysis. This would bring parameter estimation to a faster time scheme while maintaining the rigor of the analysis. Maintaining Minerva's advantages like its modularity, reusability of optimizers and flexibility of redefining parameters as free is also desired.

This work intends to show how this type of generic Bayesian analysis for the described variety of inverse problems can benefit from the development of FPGA architecture and its advances in resource availability as well as double precision floating point architecture. An accelerated version of the forward modeling combined with modern FPGA implementations of inversion algorithms can bring many of these inverse problems calculations closer to a real-time scheme. This would not only serve as a faster tool for interpretation of diagnostic data but also as a signal for control systems that do not require a sub-microsecond time resolution.

For this paper, an FPGA post-implementation simulated design

for an accelerated plasma electron density and temperature profile estimation is presented. The profile is determined by the joint analysis of two ubiquitous diagnostics in magnetic confinement device: the Dispersion Interferometer (DI) and the Thomson Scattering (TS) diagnostics. The acceleration of the DI model alone was demonstrated already in [11].

II. MODELING

Among the plasma parameters that are typically measured, electron density and electron temperature tend to be important for several reasons going from quality of a plasma to safety limits of the machine. Examples of this are plasma positioning or determination of a time window to stop heating.

To estimate both of these parameters the two physical phenomena measured are refraction index and Thomson scattering. In the case of W7-X, these are measured with a Dispersion Interferometer and a Thomson Scattering diagnostic which have a coinciding Line of Sight (LoS). From the latter we can infer both parameters with some limitations. If we introduce the DI to the analysis, we can improve our inference for cases where TS under-performs.

A. Thomson Scattering Diagnostic at W7-X

The Thomson Scattering is a typical and well developed diagnostic used in many fusion devices. In the case of the W7-X, it measures light from a Nd:YAG laser that is scattered by the electrons in the plasma in order to determine electron density and temperature. The laser, while crossing the plasma, will accelerate free electrons in the plasma making the electron emit a scattered photon at a Doppler shifted frequency. This scattered wave's spectrum is proportional to the electron velocity distribution, that is its temperature, and its intensity is proportional to the electron density. The scattered light is picked up by two sets of lenses that observe the whole plasma cross-section. The light is collected and transported by fiber bundles to polychromators with 5 filters. These measure spectral intervals with different width due to signal to noise considerations. The light intensity is then detected by avalanche photodiodes. For the scattered power P_s the analytical expression for incident and observed light polarized perpendicular to the scattering plane is [12]:

$$\frac{d^3 P_s}{d\epsilon d\Omega dr} = r_e^2 n_e \langle S_l \rangle S(\epsilon, \theta, \alpha) \tag{1}$$

where r_e is the classical electron radius, n_e the electron density, S_l the Poynting vector for the incident laser light beam, θ the scattering angle, ϵ the normalized wavelength shift $(\lambda_s - \lambda_l)/\lambda_l$ and α the normalized inverse temperature $m_e c^2/(2T_e)$.

For each polychromator, the ith spectral interval signal measured by the avalanche diode can be modeled as

$$\left(\int s dt \right)_i = n_e E_l \frac{r_e^2}{hc} \delta\Omega g_0 \int \int S(\epsilon, \theta, \alpha) \frac{\lambda}{\lambda_l} \frac{g_i(\lambda)}{g_0} dL d\lambda \tag{2}$$

where E_l is the laser energy, $S(\epsilon, \theta, \alpha)$ the spectral density function, $\delta\Omega$ is the solid angle of the system, $g_i(\lambda)$ the absolute sensitivity, the left hand side a time integration and the right hand side a length of observation volume and a wavelength integration respectively [S. Bozhenkov, to be published]. In this case, where plasma parameters don't change significantly with respect to diagnostic resolution, the length integration can be simplified to a multiplication with the volume length.

A majority of the calculation of $S(\epsilon, \theta, \alpha)$ is stated below in order to represent the complexity of the forward model and show how modern FPGA have enough resources to handle such models. The rest can be found in [12].

$$S(\epsilon, \theta, \alpha) = S_z(\epsilon, \theta, \alpha) q(\epsilon, \theta, \alpha) \tag{3}$$

where we know that $S_z(\epsilon, \theta, \alpha)$ and $q(\epsilon, \theta, \alpha)$ are

$$S_z(\epsilon, \theta, \alpha) = \frac{e^{-2\alpha(x-1)}}{2K_2^*(1+\epsilon)^3 \sqrt{2(1-\cos(\theta))((1+\epsilon)) + \epsilon^3}} \tag{4}$$

$$K_2^*(2\alpha) \approx \sqrt{\frac{\pi}{2(2\alpha)}} \left(1 + \frac{15}{8(2\alpha)} + \frac{105}{128(2\alpha)} + \frac{315}{1024(2\alpha)} \right) \tag{5}$$

$$q(\epsilon, \theta, \alpha) = 1 - 4\eta\zeta \left(\frac{p_0(\zeta)\eta + p_1(\zeta)\eta + p_2(\zeta)\eta}{q_0(\zeta)\eta + q_1(\zeta)\eta + q_2(\zeta)\eta} \right) \tag{6}$$

$$u = \frac{\sin\theta}{1 - \cos\theta}, \qquad x = \sqrt{1 - \frac{\epsilon^2}{2(1 - \cos(\theta))(1+\epsilon)}} \tag{7}$$

$$y = \frac{1}{\sqrt{x^2 + u^2}}, \quad \eta = \frac{y}{2\alpha}, \quad \zeta = xy \tag{8}$$

B. The Dispersion Interferometer

The Dispersion Interferometer is a diagnostic that, in the case of the W7-X, allows the indirect measurement of line integrated electron density due to its proportionality with the phase shift introduced by the plasma refraction index [5], [7].

The DI works by frequency doubling and polarizing an amount of the incident 10.6 μm wavelength CO_2 laser light. This beam crosses the plasma with two different frequency and polarization components, each receiving a different phase shift. After crossing the plasma twice, the base frequency component gets doubled and shifted before filtering the base-band frequency and leaving just the interference pattern. This is then measured by a photo-diode. The recorded signal can be modeled as:

$$V = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(m\pi \sin(\omega t) + \Delta\varphi), \tag{9}$$

where m is the modulation depth set by a photoelastic modulator at $\omega = 50 kHz$, $I_{1,2}$ are typically constant along period of the signal and $\Delta\varphi$ is the phase difference between the two wave components in the beam path. The phase difference is proportional to the line integrated electron density $\int n_e dl$ of the plasma along the crossed LoS.

C. Bayesian Model

In order to tackle this problem with Bayesian analysis we can formulate Bayes' theorem with the parameters of interest. In this case, the parameters of interest are the density and temperature values of each TS channel. The integration of these points along the plasma cross-section joins the TS points to the DI's line integrated electron density measurement and can be expressed as,

$$p(\vec{n_e}, \vec{T_e} \mid \vec{D_{TS}}, D_{DI}) = \frac{p(\vec{D_{TS}}, D_{DI} \mid \vec{n_e}, \vec{T_e}) p(\vec{n_e}) p(\vec{T_e})}{p(\vec{D_{TS}}, D_{DI})}, \tag{10}$$

where $\vec{D_{TS}}$ is the data vector from all TS channels, D_{DI}, the data from the interferometer, $p(\vec{n_e}, \vec{T_e} \mid \vec{D_{TS}}, D_{DI})$, is the posterior probability density function (pdf). The likelihood, $p(\vec{D_{TS}}, D_{DI} \mid \vec{n_e}, \vec{T_e})$, is defined by a normal distribution of the data set around the predicted forward modeled values. Finally $p(\vec{n_e})$ and $p(\vec{T_e})$ are the priors of the selected parameters and $p(\vec{D_{TS}}, D_{DI})$ is the evidence. For this parameter estimation the evidence factor can be disregarded given that it is constant for every given data set. For the prior, given the natural limits of temperature and density, values lower than 0

and an upper limit are not expected. This allows a truncated normal distribution to be selected and can be expressed for density as:

$$p(n_{ei}) = \begin{cases} \frac{1}{A_i \sigma_{ne_i}\sqrt{2\pi}} exp(-\frac{1}{2}\left(\frac{n_{e_i}-\mu_{nei}}{\sigma_{ne_i}}\right)^2) & 0 \le n_{ei} \le b \\ 0 & otherwise \end{cases}$$
(11)

where i is the channel number of the TS diagnostic, A is the normalization value for the truncated normal and μ_{ne_i} the mean value for the density value of that specific channel.

For the likelihood with data coming from different sources, we can write it as [14]:

$$p(\vec{D_{TS}}, D_{DI} \mid \vec{n_e}, \vec{T_e}) = \mathcal{N}(\vec{D}_{TS}; \vec{V}_{TS}, \vec{\sigma}_{TS}^2)\mathcal{N}(D_{DI}; V_{DI}, \sigma_{DI}^2).$$
(12)

Here, the normal distributions $\mathcal{N}$ of the data $D_{TS/DI}$ over the predicted forward modeled value V are multiplied. $\sigma_{TS/DI}^2$ represents the covariance matrix and variance value of the TS and DI data respectively, representing the noise level of the data. The resulting posterior pdf is 20 dimensional.

Given that this analysis was partially selected due to the proper handling of the uncertainty, besides the most likely value an uncertainty estimation is desired. Thus the posterior distribution, which is the full description of the parameters given the data, needs to be consequently explored. The Metropolis-Hastings Markov chain Monte Carlo (MCMC) method was selected which samples the posterior iteratively.

The iterative nature of MCMC makes it hard for a real-time solution since it depends on the value of the previous sample point during iterations. The MCMC requires a *burn-in* period of the chain to enter a high probability region representative of the distribution, plus a number of representative samples of the distribution around this point.

In order to limit the number of *burn-in* iterations, a good initial point for each free parameter is required. This will assure a faster convergence of the chain and therefore requiring less iterations. Amongst the many optimizer algorithms Hooke and Jeeves was selected for this work to find a good initial point of the Maximum a Posteriori (MAP).

The analysis was done with real data from the first operation campaign of W7-X and validated against analyzed data of the TS diagnostic.

After testing several possible combinations, a minimum of 10 rounds of Hooke and Jeeves and 40000 rounds of MCMC were required. When compared to the reference analyzed data, Figure 1 was generated and describes the desired density and temperature profile to be achieved in a real-time frame.

III. ACCELERATION WITH FPGA ARCHITECTURE

Previous works in the field of Bayesian filtering has been very successfull in bringing Bayesian analysis to a real-time scheme with different filtering techniques. The majority are for analysis of time-dependent models where the state-space approach and Bayesian filtering and smoothing are used for control systems [13]. Linear problems with Gaussian noise can be solved with the Kalman filter, which is achieved by reformulating Bayes' theorem [3]. Many others have also been developed to deal with non-linear problems on time-dependent models that reduce the volume of calculations through sub-optimal algorithms and linearizations while following a Bayesian approach. Some of these are: particle filters, extended Kalman filters (EKF), grid based algorithms as well as variational Bayes techniques.
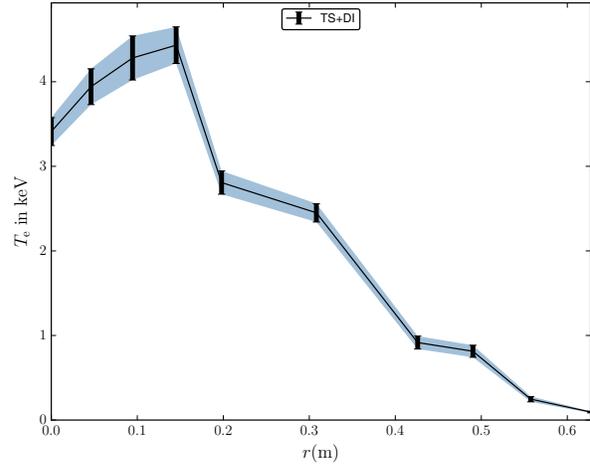
Fig. 1. Temperature Profile. *40,000 MCMC iterations, 10 MAP iterations.*

Most of these are Bayesian solutions to time-varying inverse problems formulated as stochastic state-space models [1], [2], [4], [6], [9]. For many cases, this is the best way to reach real-time solution in a control theory frame and dynamic model state-space approach. Unfortunately, for the discussed scenario they do not completely satisfy the aforementioned requirements. Therefore an FPGA design was selected to accelerate this type of analysis.

To find possibilities for acceleration, the analysis can be divided into tasks. These are: the inversion algorithm, the calculation of Bayes' theorem and the forward model. The results of this paper focuses on the latter and how the processing time of models with complex mathematical operations can be reduced using a dedicated FPGA architecture.

A. Architecture Design

The first acceleration possibility of an FPGA architecture is parallelism. For this inverse problem, as well as many others with several data sources, this is an immediate advantage. The forward model of each channel as well as the spectral integration in (2) for each of the 5 filters can be parallelized. Besides this, (3) shows that $S_z(\eta, \theta, \alpha)$ and $q(\eta, \theta, \alpha)$ can also be parallelized.

To meet the required precision, double precision floating point architecture was selected. Previous FPGA generations would have presented resource difficulties to implement a model of this size with the aforementioned precision. Nevertheless, current resource availability in modern FPGA and optimized reconfigurable IP Cores makes this no longer an issue.

Typically operations like divisions, exponentials and square root require more resources compared to the other arithmetic operations. These can be modified to reduce the resources used. Multiplications with the inverse of a constant replace more costly divisions over a constant. Also, for equations like (5), factors like $1/2\alpha$ can be calculated through one division and the rest of the equation with multiplications and sums. This reduces the sheer volume of the resources required for a forward model of this size.

The second acceleration possibility gained through dedicated hardware is pipelining. The main critical path is the calculation of the forward model for 3700 wavelength values before the spectral integration. This iterative process can benefit from the pipelining of all arithmetic operations. For this initial design, a test clock frequency

TABLE I
ACCELERATION VS CPU TIME $(0.8 \pm 0.1\mu s)$

CPU Freq.(MHz)	FPGA(ns)	Acceleration(N-fold)
100.0	10.48	81
200.0	5.24	162
300.0	3.49	244

TABLE II
RESOURCE CONSUMPTION FOR A SINGLE TS CHANNEL

Resource	Utilization	Available	Utilization %
LUT	61035	433200	14.09
FF	39269	866400	4.53
LUTRAM	3143	174200	1.8
DSP	721	3600	20.03

of 100 MHz was selected. To avoid FPGA timing violations for the design placing and routing, a minimum initial latency of 5 clock cycles for operations like sums and multiplications was applied. For slower and resource demanding operations like divisions, exponentials and square roots, a latency of 20 clock cycles was selected. This allows for several wavelengths be calculated simultaneously in the pipeline. Figure 2 describes this proposed architecture.
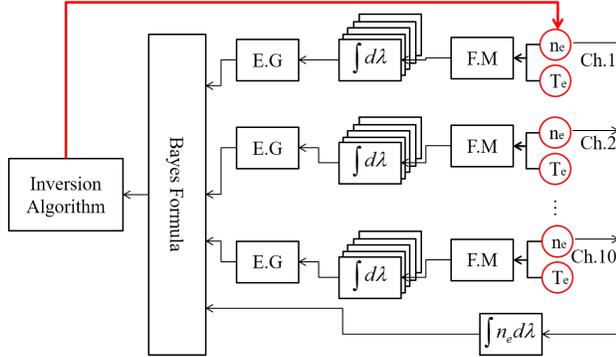
Fig. 2. FPGA Architecture. 10 n_e and T_e Channels.

IV. RESULTS AND ANALYSIS

As mentioned before, this paper focuses on the immediate advantages of FPGA dedicated hardware architecture for acceleration of the forward model.

To compare against FPGA performance, a Java profiling of the Minerva calculation of the forward model over 20 runs resulted on an average of $0.8 \pm 0.1\mu s$ duration for each value of wavelength. This was tested on a Intel Xeon E3-1505M CPU running at 2.8GHz with 16 GB of RAM memory. The critical path in this architecture belongs to the TS forward model, specifically the branch calculating (6). Post-implementation simulations of the TS forward model shows how the critical path requires 180 clock cycles for a single value of wavelength. If we consider a wavelength resolution of 3, 700 we can then calculate a duration for each wavelength and the acceleration compared to CPU code in Table I.

Here we can see the achieved ≈ 80 Fold acceleration for a single wavelength calculation as well as other frequencies well within the limits of operation of the IP Cores. Since the algorithm needs to calculate this for each channel, a parallelization of the forward model could result in another tenfold reduction of processing time without considering DI's forward model gain.

Profiling also showed that the forward model represents the bulk of the full analysis. Considering the achieved results, 40,000 iterations of this fully parallelized forward model would require $\sim 1.5s$ for the slowest clock frequency, which indeed brings the analysis closer to a real-time frame.

Regarding resource availability, for this first implementation the

Virtex 7 xc7vx690tffg1761-2 was selected. For a single channel of the TS, Table II shows that the most used resource is the DSP slices which, if necessary, can also be traded for logic within the IP Core configuration.

V. CONCLUSIONS

Compared to the number of calculations needed for each inversion iteration and application of Bayes' theorem, the forward model takes up most of the processing time. An 80-fold acceleration of the forward model was achieved with dedicated FPGA architecture, reuse of architecture through pipelining and a parallelism for each channel and spectral integration. This type of acceleration is ideal for inverse problems with several channels using the same forward model but not limited to those. Complex calculations can be branched, parallelized and gain from pipelining.

It is important to state that this paper only covers the part of the project that addresses the forward model, while inversion speed is also of importance and the current step being developed for the full solution. The field of statistical signal processing has also made significant advances in this area that can be used together with forward model acceleration. From the many optimizers and samplers accelerated on FPGA architecture, one good example is the acceleration of Parallel-Tempering MCMC [10]. This clearly shows how, by combining these solutions, there is enough development and algorithms to bring full Bayesian analysis closer to real-time frame using modern FPGAs. Besides this, reconfigurable hardware provides a suitable environment for the fast modification of these models, addition of new free parameters and exchange of models or optimizers to keep the flexibility of the analysis. If applied to time-independent inverse problems, as shown in this paper, a robust analysis tool like Bayesian analysis can be beneficial for scientific inference in modern physics experiments.

VI. ACKNOWLEDGMENTS

REFERENCES

[1] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A Tutorial on Particle Filters for Online Nonlinear / Non-Gaussian Bayesian Tracking. 50(2):174–188, 2002.
[2] Michael A Chappell, Adrian R Groves, Brandon Whitcher, and Mark W Woolrich. Variational Bayesian Inference for a Nonlinear Forward Model. 57(1):223–236, 2009.
[3] Z H E Chen. Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond. *Statistics*, 182(1):1–69, 2003.
[4] A Problem Formulation. Recursive Noise Adaptive Kalman Filtering by Variational Bayesian Approximations. 54(3):596–600, 2009.
[5] M Hirsch. Microwave and Interferometer Diagnostics prepared for first Plasma Operation of WENDELSTEIN 7-X. pages 1–8, 2015.

[6] Y. Ho and R. Lee. A Bayesian approach to problems in stochastic estimation and control. *IEEE Transactions on Automatic Control*, 9(4):333–339, 1964.

[7] P. Kornejew, H. Trimiño, S. Heinrich, and M Hirsch. Final Design of the Dispersion Interferometer for the Wendelstein7-X Stellarator. *40th EPS Conference on Plasma Physics*, (figure 1):10–13, 2012.

[8] M Lennholm, P S Beaumont, I S Carvalho, I T Chapman, R Felton, D Frigione, L Garzotti, A Goodyear, J Graves, D Grist, S Jachmich, P Lang, E Lerche, E De Luna, R Mooney, J Morris, M F F Nave, F Rimini, G Sips, E Solano, M Tsalas, and J E T Efda Contributors. ELM frequency feedback control on JET.

[9] Lifeng Miao, Student Member, Jun Jason Zhang, and Chaitali Chakrabarti. Ef fi cient Bayesian Tracking of Multiple Sources of Neural Activity : Algorithms and Real-Time FPGA Implementation. 61(3):633–647, 2013.

[10] Grigorios Mingas and Christos-savvas Bouganis. Parallel Tempering MCMC Acceleration Using Reconfigurable Hardware. pages 227–238, 2012.

[11] H. T. Mora, J. Svensson, O. J. Ford, A. Werner, R. Wolf, and D. Timmermann. Acceleration of bayesian model based data analysis. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, pages 523–529, Nov 2016.

[12] O Naito, H Yoshida, and T Matoba. Analytic formula for fully relativistic Thomson scattering spectrum. 4256(1993):10–13, 2015.

[13] S Sarkka. Bayesian Filtering and Smoothing. page 254, 2013.

[14] D.S. Sivia. *Data Analysis A Bayesian Tutorial*. Oxford University Press, 1996.

[15] J. Svensson, A. Dinklage, J. Geiger, A. Werner, and R. Fischer. Integrating diagnostic data analysis for W7-AS using Bayesian graphical models. *Review of Scientific Instruments*, 75(10 II):4219–4221, 2004.

[16] J. Svensson, O. Ford, D. C. Mcdonald, A. Meakins, A. Werner, M. Brix, A. Boboc, M. Beurskens, and Jet Efda Contributors. Modelling of JET Diagnostics Using Bayesian Graphical Models. *Contributions to Plasma Physics*, 51(2-3):152–157, 2011.

[17] J. Svensson, O Ford, a Werner, G Von Nessi, M Hole, D C Mcdonald, L Appel, M Beurskens, a Boboc, M Brix, J Howard, B D Blackwell, J Bertram, and D Pretty. Connecting physics models and diagnostic data using Bayesian Graphical Models. *37th EPS Conference on Plasma Physics*, (10):O4.117, 2010.

[18] J Svensson and A Werner. Large Scale Bayesian Data Analysis for Nuclear Fusion Experiments. In *IEEE International Symposium on Intelligent Signal Processing*, pages 1–6, 2007.