# Integrating diagnostic data analysis for W7-AS using Bayesian graphical models

J Svensson, A Dinklage, J Geiger, and A Werner

Max-Planck-Institut für Plasmaphysik, EURATOM Association, Wendelsteinstr. 1,

17493 Greifswald, GERMANY

R Fischer

Centre for Interdisciplinary Plasma Science, Boltzmannstr. 2, 85748 Garching,

GERMANY

Analysis of diagnostic data in fusion experiments is usually dealt with separately for each diagnostic, in spite of the existence of a large number of interdependencies between global physics parameters and measurements from different diagnostics. In this paper, we demonstrate an integrated data analysis model, applied to the W7-AS stellarator, where diagnostic interdependencies have been modeled in a novel way by using so called Bayesian graphical models. A Thomson scattering system, interferometer, diamagnetic loop and neutral particle analyzer are combined with an equilibrium reconstruction, forming together one single model for the determination of quantities such as density and temperature profiles, directly in magnetic coordinates. The magnetic coordinate transformation is itself inferred from the measurements. Influence of both statistical and systematic uncertainties on quantities from equilibrium calculations, such as position of flux surfaces, can therefore be readily estimated together with uncertainties of profile estimates. The model allows for modular addition of further diagnostics. A software architecture for such integrated analysis where possibly large number of diagnostic and theoretical codes need to be combined, will also be discussed.

07.05.-t, 07.05.Kf, 07.05.Mh, 29.85.+c, 28.52.-s

## Introduction

A large fusion experiment typically utilises on the order of 100 diagnostics for measuring different aspects of the plasma. Usually, the data from each diagnostic is analyzed separately, in spite of the fact that dependencies between diagnostics are plenty and could be utilised to increase accuracy and alleviate consistency problems. One such set of dependencies comes from the direct usage, in the analysis of one diagnostics data, of physics parameters emerging from previous analysis of data from other diagnostics. Since such parameters are statistical quantities, they will deviate (sometimes considerably) from their true values, and will in such cases lead to erroneous results in such subsequent analysis. Another type of interdependency is the relation between quantities derived from different diagnostics through a physics model, e.g. a measurement of thermal energy (from a diamagnetic loop) should equal the volume integrated isotropic pressure inferred from profile diagnostics, ‘within error bars’. Further interdependencies can also enter in the form of constraints known *a priori*, such as pressure profiles being monotonically decreasing if flux surfaces are assumed, or volume integrated power transfer between ions and electrons being less than total input power.

Combining measurements from different diagnostics will not be generally feasible without the possibility of modeling systematic errors at each diagnostic. In Bayesian theory, this is accomplished by associating a systematic uncertainty with a probability density function (pdf), and then averaging over these ‘nuisance’ parameters to arrive at the marginal pdf of the physics quantities of interest.

In this paper, we model and visualize this web of probabilistic and physics interdependencies by so called Bayesian graphical models [1], in which each single measurement or free parameter/nuisance parameter is represented by a node, and edges between nodes represent a conditional dependency (probabilistic or deterministic) on the parent nodes (figure 1). Inference in such a system can then be done using Markov Chain Monte Carlo [2] techniques.

## Bayesian Graphical Models

In the Bayesian view of inference, all quantities, measurements and unknowns, are represented by probabilities or pdf:s. The associated pdf for a measurement is the distribution from which the measurement is believed to have been sampled, and the pdf for an unknown parameter represents the possible values this unknown can take. For inference of the unknowns from the measurements to be possible there must exist dependencies between them, captured by the joint density of measurements ($D$) and unknown parameters ($A$):

$$p(A, D) \qquad (1)$$

This joint pdf now represents everything we need for inference about the unknown parameters. For example, by using standard laws of probability, the conditional pdf for the unknowns $A$ given a specific observation of measurements $D$ is given by:

$$p(A \mid D) = \frac{p(A,D)}{p(D)} = \frac{p(D \mid A)\,p(A)}{p(D)} \qquad (2)$$

which is Bayes formula. The conditional pdf for *A* on the left is referred to as the *posterior* pdf and the unconditional pdf for *A* on the right, the *prior*. If we are interested in inference about only a subset of the parameters in *A*, we integrate out of the posterior those parameters we are not interested in ($A=\{a_1,a_2,a_3..a_N\}$):

$$p(a_1, a_2 \mid D) = \iint_{\ldots} p(a_1, a_2, a_3 \ldots a_N \mid D)\, da_3 \ldots da_N \qquad (3)$$

referred to as *marginalization* of the posterior.

The joint pdf (1) for a system consisting of many diagnostics, with dependencies of the types described in section 1, will be very complex, and is therefore here built up using so called Bayesian graphical models, which are graphs where each node represents a conditional distribution or a deterministic function of the parent nodes to which it is directly connected (figure 1). There is more than one possible graph for a given pdf, but by building the graph along ‘causal’ directions of influence (here, from physics parameters to measurements), an intuitive representation of the complex model can be built up. The joint pdf of all measurements and free parameters is then given directly from the graph by the product of the individual conditional distribution nodes given their parents:

$$p(U) = \prod p(u_i \mid par(u_i)) \qquad (4)$$

Here $U=\{A,D\}$ and $par(u_i)$ is the set of parent nodes to node $u_i$, which can be empty. In this way complex dependencies can be built up and further nodes added to an existing pdf represented as a Bayesian graphical model. For example, the sub-graphs at the bottom of figure 1 are graph representations of participating diagnostics of the integrated system described here, including an outline of the YAG Thomson scattering diagnostic model defined in [3]

## Integrated Data Analysis Model for W7-AS

The idea of integrating a Thomson scattering diagnostic, diamagnetic loop measurement, interferometer, and results from a neutral particle analyzer (with the possibility of modular addition of further diagnostics) with an equilibrium calculation, is the direct inference of joint physics quantities directly in magnetic coordinates, where all uncertainties (statistical and systematic) in the whole system influence all calculated quantities, including quantities from the equilibrium calculation itself. The magnetic coordinate system is determined directly from inference on the (uncertain) pressure profile inferred from the measurements from participating diagnostics. This self-consistent inclusion of an equilibrium calculation is necessary in order not to introduce systematic errors from a ‘fix’ magnetic coordinate system. The most important free parameters of the model are the parameterized electron density, electron temperature and ion pressure as a function of toroidal magnetic flux (upper left of figure 1), from which the magnetic coordinates for the participating diagnostics are calculated through a fast equilibrium calculation (using Function Parameterization (FP) [4]). The parameterised form of these profiles are (as a function of the normalised toroidal flux, *s*):

$$f(s) = a_0(1-s^2)^\alpha \exp(a_1 s + a_2 s^2 + a_3 s^3) \qquad (5)$$

where $\{a_i\}$ are the free parameters and α=1 for temperature and density profiles, α=2 for pressure profiles. This form is chosen because it is identical to the profile shapes used for the FP calculations [4] and for its similarity with routinely used profile forms for fitting experimental profiles at W7-AS. Also needed for the equilibrium calculation is an extra free parameter related to the distance between the LCFS and the nearest in-vessel component. Twelve (one per channel) uncertainties related to the calibration of the YAG Thomson system [3] are also included, treated here as nuisance parameters to be marginalized out (3) from the joint posterior. This gives a total of 25 free parameters for the problem.

## Results

Inference of specific parameters in the joint model, such as a parameterized profile, is done using Markov Chain Monte Carlo (MCMC) techniques. MCMC is a method for drawing samples from unnormalized pdf:s. By taking successive samples from the joint posterior distribution (after a number of ‘burn-in’ steps of the Markov chain), the marginalization procedure in (3) can be carried out by simply collecting those elements from the sampled vector belonging to the free parameters whose marginal posterior is sought. Pdf:s for functions of the free parameters are easily calculated from repeated function evaluations from such samples.

Figure 2 shows a large number of posterior samples for the inferred electron density profile. What is important about this inferred profile is that it is a direct function of a magnetic coordinate, where uncertainty in the magnetic coordinate system itself, as well as other uncertainties in the system, statistical and systematic, are represented by the profile width at any given point.

Figure 3 shows a similar graph for flux surfaces for the plane including the YAG laser line of sight. The uncertainty for each flux surface is plotted as the width of the surface. Generally, each node in figure 1 corresponding to a free or derived parameter has a corresponding posterior pdf calculated from a MCMC run.

## Software Architecture

The problem of integrating a number of diagnostic analysis codes, theory codes (such as equilibrium calculations) etc. necessary for the type of integrated data analysis described in this paper, could easily prevent any such analysis from happening. Codes are written in different languages and run on different platforms. The model described here is dependent on a combination of a FORTRAN code running on a Linux platform, a C++ program running on Windows NT and a platform independent Java-program. The addition of further diagnostics or theory codes will increase this list. We have therefore defined the interfacing between modules in a principled way based on newly emerged so called web service standards [5], where message passing between different software modules is done using language and platform independent XML messages. The definition of methods and types used by a specific module is defined in a WSDL (Web Service Definition Language) file from which clients for any language or platform can then be automatically generated. This definition of standard interfaces to all participating codes makes the framework flexible, modular, and easily expandable.

## References


[1] S L Lauritzen, *Graphical Models* (Clarendon Press, 1996)

[2] D Gamerman, *Markov Chain Monte Carlo, Stochastic Simulation for Bayesian Inference* (Chapman & Hall, 1997)

[3] R Fischer, A Dinklage, and E Pasch, Plasma Phys. Contr. Fusion **45**, 1095 (2003).

[4] H Callaghan, P Mc Carthy, and J Geiger, Plasma Phys. Contr. Fusion **42**, 1013 (2000).

[5] http://www.w3.org/2002/ws/

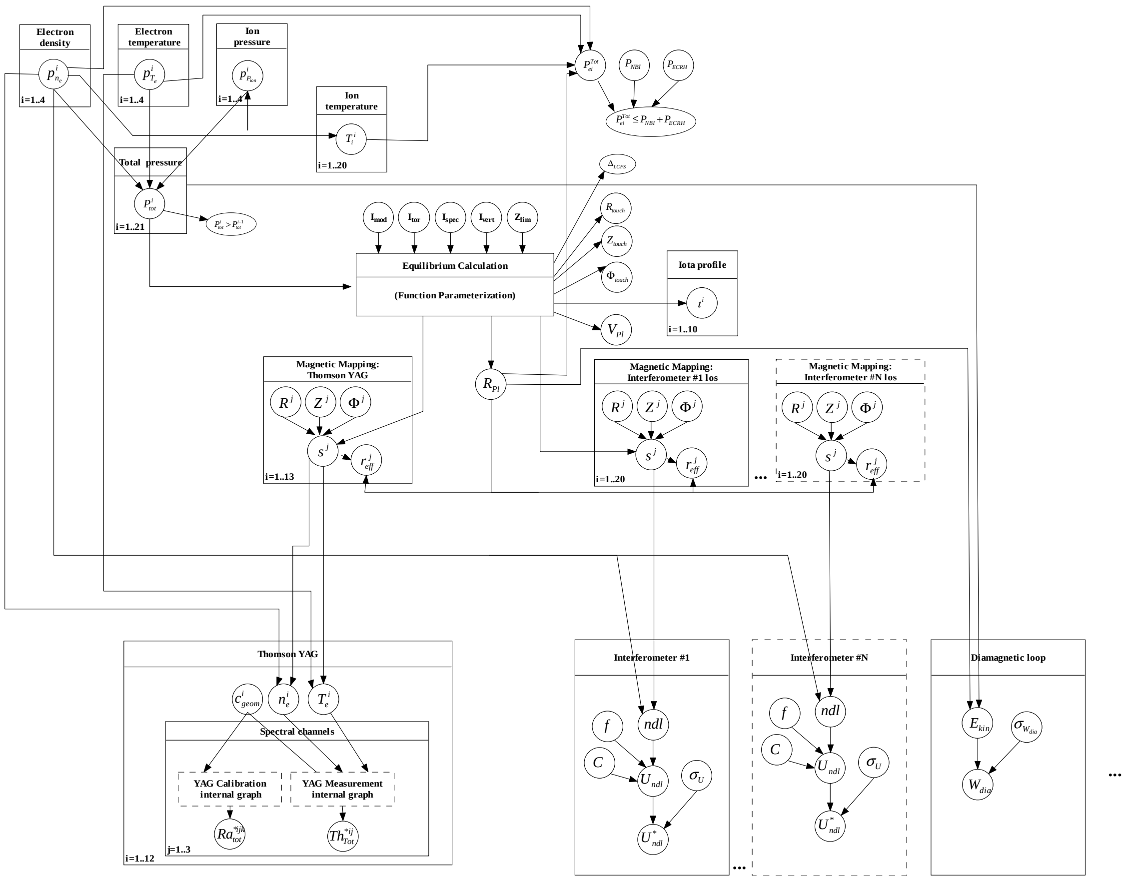


**Figure 1. Paper G01. Jakob Svensson**

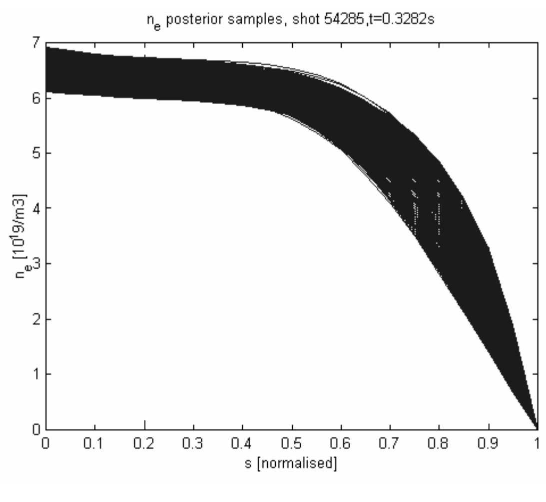


**Figure 2. Paper G01. Jakob Svensson**

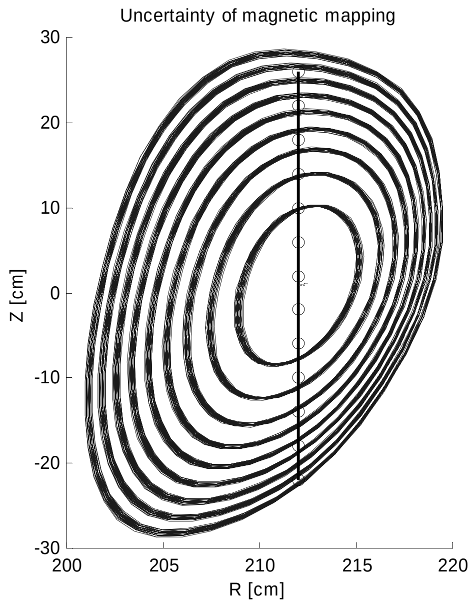


**Figure 3. Paper G01. Jakob Svensson**

**Figure 1. Main part of Bayesian Graphical Model for W7-AS. A sequence indicator in the lower left corner of boxes indicates that each node within the box has a multiplicity given by the sequence.**

**Figure 2. Posterior samples for $n_e$-profile**

**Figure 3. Uncertainty of flux surfaces**