The concept of Integrated Data Analysis of complementary experiments

R. Fischer* and A. Dinklage†

*Max-Planck-Institut für Plasmaphysik, EURATOM Association,
Boltzmannstr. 2, D-85748 Garching, Germany
†Greifswald Branch, Wendelsteinstr. 1, D-17493 Greifswald, Germany

Abstract. The Integrated Data Analysis (IDA) concept allows one to combine data from different experiments to obtain improved results. Heterogeneous and complementary experimental data as well as various kinds of physical prior information are easily integrated employing Bayesian probability theory. The concepts of IDA are compared to the traditional approach for data analysis where sequential analysis and iterative schemes are usually found. In contrast to classical inversion techniques IDA needs only forward modeling and a thorough error assessment: The ingredients are given by a model linking the physical quantities of interest to the measured data, a statistical description of the measurements, and a probabilistic description of all nuisance model parameters suffering from uncertainties. In practice, the probabilistic description of systematic measurement and model uncertainties are of major importance to resolve data inconsistencies. Complex error propagation is obtained automatically combining data in a concise probabilistic one-step analysis.

Key Words: Integrated Data Analysis (IDA), Bayesian probability theory, data consistency

INTRODUCTION

A major step in the analysis of experimental data from nuclear fusion is the coherent combination of measurements from different diagnostics. The goal is to replace the usual combination of results from the analysis of individual experimental data by a combination of the measured data sets for a one-step analysis of pooled data. The analysis of the pooled data allows one to obtain a coherent and unique result from exploiting all information/measurements available. Integrating heterogeneous diagnostics by combining measured data instead of combining inferred results automatically considers all correlations involved in the parameters to be inferred. It is the use of these correlations which allows one to extract more information from given data compared to sequential analysis.

Integrated Data Analysis (IDA) in the framework of Bayesian probability theory offers a unified way of combining all available information. The advantages from an integration of the measured data are manyfold: Physical interdependencies of heterogeneous diagnostics are considered from the beginning and no iterative procedure is necessary [1]. The interdependencies also imply the proper treatment of complex error propagation [2]. A quantitative framework for data validation and consistency checks is provided [3]. A measure for signal credibility or if a measurement should be regarded to be faulty can be provided. An one-step analysis allows to build automated procedures for next generation fusion devices which huge amount of data being analyzed automatically [4]. IDA

provides off-line to real-time analysis approaches on different time scales for different purposes [5].

The problems arising from the inversion of ill-posed problems from noisy data are mitigated by providing more data and using only forward modeling of the data. Since the probabilistic approach compels one to make quantitative and testable statements about every piece of information entering the analysis, a full documentation of the analysis process is provided. This is the basis for effective maintenance or revisions of data analysis tools. Consequently, the discussion about the validity of arguments or the credibility of uncertainty measures is based on a quantitative formulation [6, 7]. Analyzing measured data to obtain first-interest quantities is conceptually easy to couple with theory codes, e.g. for the evaluation of transport mechanism in plasmas [8]. In addition to the analysis of measured data from a running experiment, IDA provides in the framework of Bayesian experimental design an approach to optimize future experiments and combination of experiments with respect to physical goals already in the construction phase [9, 10].

The effort for the implementation of IDA consists in a thorough assessment and quantification of all sources of data, additional information, and errors and uncertainties in the measured data as well as in the modeling of the data. The probabilistic formulation of the inference problem in the Bayesian framework is straightforward, but one has to be aware that the necessary elaborate description of the different experiments poses a major effort for the physicists in charge. Quantification of the errors in all measured data and quantification of uncertainties in all model nuisance parameters is often a non-trivial task but is of vital importance for a comparable analysis of heterogeneous diagnostics.

IDA is of great value if large amount of data, additional information and interdependencies exist. This is the case in nuclear fusion but can be extended to other experiments or complex systems for which heterogeneous information is available (data from different measurements, model parameters, physical constraints, etc.). The present work compares the concept of IDA with the traditional data analysis scheme exemplified at a typical use case in fusion.

TRADITIONAL DATA ANALYSIS SCHEME

The left panel of figure 1 depicts a typical flow-chart of a traditional approach for data analysis. Different measurement techniques based on different physical effects were applied to the same experiment in order to estimate the same physical parameters (quantities of interest). Usually, the measurements were analyzed separately although an overlap of the quantities of interest exists. In the present case a Thomson scattering measurement providing electron temperature T_e and density n_e profiles of a plasma and an electron cyclotron emission (ECE) measurement providing T_e only were analyzed individually. Both experimental techniques have their advantages and disadvantages such that they complement each other. It should be noted that the independent analysis of heterogeneous measurements generally found in scientific research is traditionally owing to the personalization of hardware and software developments. De-personalization of software, as routinely realized in industry, is of minor importance in science but has

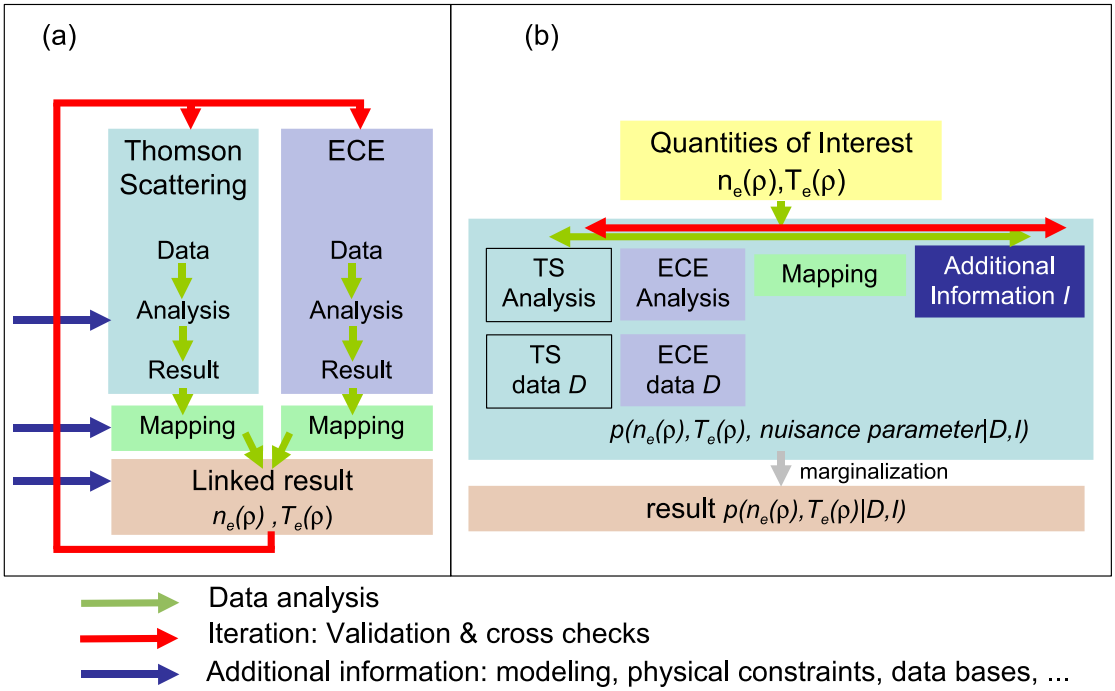

FIGURE 1. Simplified flow-charts for typical data analysis steps inferring electron temperature and density profiles for magnetic confinement fusion experiments from the Thomson scattering and electron cyclotron emission (ECE) diagnostics in (a) the traditional approach and (b) within the IDA concept.

to become more important as the scientific devices increase in complexity.

Since both measurements provide information about the same subset of parameters a combination of the results, e.g. different T_e-profiles, to obtain a unique profile is performed traditionally in a second step. A difficulty of a straightforward combination of the results is given by the fact that the measurements are not performed exactly at the same time and spatial coordinates. Time dependencies or measurements at different spatial positions might corrupt the assumption of having the same values for the quantities of interest. In magnetic confinement fusion devices, measurements at different spatial positions are mapped to a common coordinate system of so called magnetic flux surfaces which are constant pressure surfaces in ideal magneto hydrodynamics [11]. This is depicted in figure 1 as a block which maps the laboratory coordinates $\vec{x}$ of the different measurements on a common magnetic coordinate system $\rho(\vec{x})$. The mapping procedure itself usually suffers from uncertainties since the flux surfaces depend on the plasma pressure and there are usually modeling simplifications [12]. The situation complicates since some of the input quantities of the mapping procedure are the quantities of interest T_e and n_e. T_e and n_e profiles influence the plasma equilibrium and, therefore, the mapping from laboratory to magnetic coordinates. In the traditional approach the inter-

dependencies of the mapping procedure and the different experimental data are solved iteratively. An iterative solution becomes a tedious task if two or more measurements have to be combined with additional information from physical considerations and (uncertain) physical data from other data bases. An automized procedure of the full iteration, e.g. necessary for next generation steady-state fusion devices like W-7X or ITER, appears to be barely feasible.

A severe issue for the combination of measurements is the lack of standardization of error interpretation and treatment which hampers the comparability of different experimental data and results. Statements about estimation uncertainties are at best based on Gaussian error propagation. Additionally, it is often difficult to obtain uncertainties on model parameters given in literature or data bases, e.g. cross sections or atomic data. In subsequent analyses such values are treated as being exactly known although in many cases they provide the leading role in estimation errors. A general agreement about interpretation, quantification and use of errors is still lacking.

INTEGRATED DATA ANALYSIS

The Bayesian approach of IDA provides an alternative scheme for integrating any kind of (uncertain) information. The right panel of figure 1 shows the corresponding flowchart for the two measurements described above. The basic idea is simple: IDA aims to determine the probability of the quantities of interest, given all data and physics assumptions. IDA starts with the quantities of interest, e.g. T_e and n_e, as a function of the relevant coordinates, e.g. the magnetic coordinates ρ. Due to the large number of diagnostics routinely applied to fusion machines the list of physical parameters of interest might become long. Additional parameters of interest are given by, e.g., plasma impurities, particle and energy transport mechanism, the interaction of the hot plasma with the surrounding walls and heating scenarios. Plasma modeling provides additional information which links various of those parameters [8].

Modeling individual diagnostics

With the corresponding subset of the parameter list the measured data of an individual diagnostics is modeled. Modeling of the diagnostics data is usually done independent of IDA and it is often straightforward to implement the already present data descriptive model into the Bayesian framework. First the canonical coordinates ρ have to be mapped on laboratory coordinates of the individual diagnostics $\vec{x}(\rho)$. The mapping procedure which has to be applied here is inverse to the traditional mapping procedure since the point $\vec{x}$ of a diagnostic line of sight needs to be determined for any magnetic coordinate ρ. It is assumed that all parameters necessary for the mapping procedure are provided in the list of quantities of interest. Hence, no iteration is involved in order to obtain a consistent mapping between coordinate systems.

In contrast to classical inversion techniques IDA needs forward modeling only. For a given set of values for the quantities of interest, e.g. T_e and n_e, the calculation of the

ideal measured data is usually much simpler to be provided than the solution of the inversion problem. The inverse problem is often ill-posed due to the inevitable noise in the measured data. The forward modeling only has to provide ideal data which means data which would be measured if there would be no statistical measurement noise. The measurement noise enters the probabilistic description with the likelihood probability distribution function (pdf) which quantifies the probability of measuring the actual data given the modeled (ideal) data. The maximum likelihood (ML) principle in orthodox statistics exploits this statistical interpretation and minimizes the misfit between the measured and modeled data with respect to the parameters of interest. Here, the likelihood describes the uncertainty (reliability) of the measured data. It quantifies the plausibility of measuring the data set given the values for the parameters.

An important ingredient of the IDA approach is an elaborate assessment of all uncertainties of the measurement systems. This is necessary to allow for a reliable combination of the heterogeneous diagnostics. The uncertainties arise usually from statistical fluctuations but are often also given by systematic uncertainties. Statistical fluctuations appear in the measured data to be analyzed as well as in data recorded for relative and absolute calibration [2]. Systematic uncertainties may arise due to incomplete/simplified modeling of the physics, due to uncertainties in model (data base) quantities such as cross sections, due to mis-specification of the measurement system or due to non-stationary measurement conditions such as darkening of windows or degeneration of glass fibers. All statistical and systematic uncertainties have to be incorporated in the likelihood pdf or have to be described using prior pdfs with subsequent marginalization of the corresponding nuisance parameters [3]. Another systematic uncertainty can arise from the different measurement techniques of the same physical quantity. An example is given by the different measurement techniques of TS and ECE for determining T_e. TS measures the electron energy distribution in the scattering geometry whereas ECE measures the radiation temperature perpendicular to the magnetic field. So far it is assumed that both temperatures are identical. Furthermore the assumption of thermal equilibrium might be too optimistic resulting in deviations from the Maxwellian velocity distribution.

Additional information

The probabilistic description of the heterogeneous experiments and the mapping procedure have to be complemented by additional information which can easily be provided using prior probability distribution. Examples for additional information arise from simple constraints such as positivity constraints for T_e and n_e up to quite complex constraints with interdependencies of various parameters. In plasma physics monotonicity on the electron pressure $p_\mathrm{e} \propto n_\mathrm{e} T_\mathrm{e}$ is often assumed where the assumption can be relaxed using appropriate prior distributions. Another example is given by the energy balance and particle transport equations which link n_e, T_e and the ion temperature T_i [13].

Please note that such additional informations should enter only once which is not the case if the individual experiments are analyzed separately and the results are combined afterwards. In the present example prior information on the parameters, e.g. T_e, would

be used twice if used in the traditional data analysis scheme.

It can be shown that a sequential analysis of data is fully equivalent to a one-step analysis if the full probabilistic approach is applied and the dependencies are considered correctly. Since this is usually not the case for the traditional sequential approach a one-step analysis is to be preferred.

Bayes theorem and marginalization

After a thorough assessment and quantification of all known sources of uncertainties, the likelihood pdfs for the measured data, the prior pdfs for the nuisance parameters and the prior pdfs for additional testable information are multiplied according to Bayes theorem. The advantage of the Bayesian formulation is that any kind of information can be combined since the probabilistic formulation allows any kind of functional form to be multiplied. For example a Gaussian likelihood can be combined with a Poissonian likelihood from another experiment, or a Gamma distribution quantifying prior information about a model parameter, e.g. a cross section.

Subsequent marginalization of all nuisance parameters provide the final result, namely the marginal posterior probability distribution with respect to the quantities of interest which were defined at the beginning. The value of the posterior pdf for the chosen values of the quantities of interest evaluates the probability (plausibility) of the parameter values given the data and information at hand. From this generally multi-dimensional posterior distribution estimates and estimate uncertainties for the quantities of interest can be derived by maximizing the posterior pdf or taking moments of the distribution.

Data consistency and sensitivity study

In a next step the posterior pdf and parameter estimates can be used to check for a consistent description of the measured data. The difference of the measured data point and the modeled data using the estimated parameters weighted with the data uncertainties (residues) indicate if the modeled data and uncertainties are correct. If a consistent model of all data within the quantified uncertainties is not obtained an important ingredient in either the physical model or the uncertainties is missing.

A major step towards a consistent and reliable description of the individual experiments is possible only on the basis of a thorough quantification of all uncertainties known so far. It provides a framework for a quantitative improvement for single measurements. The frequently observed blaming of the *other* experiment for being responsible for systematic deviances is then replaced by a quantitative approach which helps to improve debate culture.

The Bayesian framework does not provide an answer what is missing or wrong in the physical description of the experiments but the probabilistic formulation allows to identify the most crucial issues in the description. Within a sensitivity study of the most important uncertainties one can identify the most promising experimental improvements [3]. Reasons for inconsistent data or wrong assumptions can be identified by appropriate

case studies where the description is hypothetically modified. In such a virtual diagnostic the changes in the result can then be checked for being a possible candidate for the inconsistency.

In conclusion, a probabilistic description of systematic measurement and model uncertainties are of major importance to resolve data inconsistencies.

Complex error propagation

Another advantage of the IDA approach is the intrinsic property of complex error propagation. The combination of likelihood and prior probability distributions describing different experiments and additional information contain all interdependencies of parameters both of interest and of nuisance. Marginalization of nuisance parameters results in a propagation of their uncertainties to the quantities of interest. If the interest is in a subset of the original list of quantities, marginalization of the complementary subset yields a propagation of their uncertainties to the subset of quantities of interest. There is no need for applying the familiar Gaussian error propagation laws which assumptions are frequently not fulfilled since the assumption of existing second moments as well as the assumption of an uni-modal probability distribution might fail. Marginalization of uncertain parameters works under any assumption and provides complex error propagation automatically when combining data in a concise probabilistic one-step analysis.

Bayesian probability theory provides thereby a standardization of error interpretation, quantification and use. Statistical and systematic uncertainty are described using probability distributions as a measure for credibility of the information. The χ^2-misfit of data fitting is equivalent to a special case of normally distributed errors.

At first view the complex error propagation produces parameter estimates with larger uncertainties than for the traditional analysis approach. The larger uncertainties result from inclusion of all uncertainties involved. But the uncertainties of the parameter estimates can be even smaller than in the classical approach. This was shown for an example of Thomson scattering data and supplemented electron temperature T_e data from soft X-ray measurements [2]. The accuracy of the electron density estimate n_e increases by 30% although the additional information does not contain information about the density n_e at all. The reason for the increase of the accuracy of n_e is the correlation between T_e and n_e. More information about T_e implies a reduction of uncertainty on n_e. This synergistic effect employing the full correlations due to T_e-n_e interdependencies of the measurements is one of the most convincing arguments in favor of a Bayesian approach for an integrated data analysis although the effect is easily understood *a-posteriori*.

CONCLUSION

The concepts of IDA are compared to the traditional approach for data analysis. IDA provides a framework for combining any kind of measured data and additional information. Heterogeneous and complementary experimental data as well as various kinds of physical prior information can be integrated employing Bayesian probability theory.

An elaborate error analysis of single measurements and modeling is important for a comprehensive analysis of the pooled data. In contrast to classical inversion techniques IDA needs forward modeling only. The probabilistic description of systematic measurement and model uncertainties are of major importance to resolve data inconsistencies. Complex error propagation is obtained automatically combining data in a concise probabilistic one-step analysis.

REFERENCES

1. R. Fischer and A. Dinklage. Integrated data analysis of fusion diagnostics by means of the Bayesian probability theory. *Rev. Sci. Instrum.*, 75:4237–4239, 2004.
2. R. Fischer, A. Dinklage, and E. Pasch. Bayesian modelling of fusion diagnostics. *Plasma Phys. Control. Fusion*, 45:1095–1111, 2003.
3. R. Fischer, C. Wendland, A. Dinklage, S. Gori, V. Dose, and the W7-AS team. Thomson scattering analysis with the bayesian probability theory. *Plasma Phys. Control. Fusion*, 44:1501–1519, 2002.
4. A. Dinklage, R. Fischer, and J. Svensson. Topics and methods for data validation by means of Bayesian probability theory. *Fusion Science and Technology*, 46:355–364, 2004.
5. A. Dinklage, R. Fischer, J. Geiger, G. Kühner, H. Maassberg, J. Svensson, and U. von Toussaint. From off-line to real-time analysis: Accelerating Bayesian analysis codes. In R. Koch and S. Lebedev, editors, *30th EPS Conference on Controlled Fusion and Plasma Physics*, volume ECA 27A, pages P–4.80. Europ. Phys. Soc., Geneva, 2003.
6. A. Dinklage, R. Fischer, M. Hirsch, E. Pasch, A. Weller, and the W7-AS team. Increasing the significance of Thomson scattering data by Bayesian modelling. In R. Koch and S. Lebedev, editors, *30th EPS Conference on Controlled Fusion and Plasma Physics*, volume ECA 27A, pages P–1.52. Europ. Phys. Soc., Geneva, 2003.
7. J. Svensson, A. Dinklage, J. Geiger, and R. Fischer. An integrated data analysis model for the W7-AS stellarator. In R. Koch and S. Lebedev, editors, *30th EPS Conference on Controlled Fusion and Plasma Physics*, volume ECA 27A, pages P–1.65. Europ. Phys. Soc., Geneva, 2003.
8. A. Dinklage, R. Fischer, H. Dreier, J. Svensson, and Yu. Turkin. Integrated approaches in fusion data analysis. In R. Fischer, R. Preuss, and U. von Toussaint, editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume Conf. Proc. 735, pages 43–51, Melville, NY, 2004. AIP.
9. R. Fischer, H. Dreier, A. Dinklage, B. Kurzan, and E. Pasch. Integrated Bayesian experimental design. In K.H. Knuth, A.E. Abbas, R.D. Morris, and J.P. Castle, editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume Conf. Proc. 803, pages 440–447, Melville, NY, 2005. AIP.
10. H. Dreier, A. Dinklage, R. Fischer, M. Hirsch, P. Kornejew, and E. Pasch. Bayesian design of diagnostics: Case studies for Wendelstein 7-X. *Fusion Science and Technology*, 50:262–267, 2006.
11. H. Zohm. Physics of hot plasmas. In A. Dinklage, T. Klinger, G. Marx, and L. Schweikhard, editors, *Plasma Physics - Confinement, Transport and Collective Effects*, volume Lecture Notes in Physics 670, pages 75–93. Springer, Berlin, 2005.
12. J. Svensson, A. Dinklage, J. Geiger, A. Werner, and R. Fischer. Integrating diagnostic data analysis for W7-AS using Bayesian graphical models. *Rev. Sci. Instrum.*, 75:4219–4221, 2004.
13. R. Fischer, A. Dinklage, and Y. Turkin. Non-parametric profile gradient estimation. In *33th EPS Conference on Controlled Fusion and Plasma Physics*. 2006.