Neural, Parallel, and Scientific Computations 20 (2012) 283-290

A NEW METHOD TO EVALUATE ASYMPTOTIC NUMERICAL MODELS BY DATA MINING TECHNIQUES

JOËL CHASKALOVIC‡ AND FRANCK ASSOUS*

‡Dalembert, University Pierre and Marie Curie,
4 place Jussieu, 75252 Paris Cedex 5, France
*Department of Computer Science & Mathematics,
Ariel University Center of Samaria, 40700 Ariel, Israël

ABSTRACT. This paper is devoted to a new approach based on data mining to evaluate the efficiency of numerical asymptotic models. We first propose an asymptotic paraxial approximations to model ultrarelativistic particles. Then, we use data mining methods that directly deal with numerical results of simulations, to understand what each order of the asymptotic expansion brings to the simulation results. This new approach offers the possibility to understand, on the numerical results themselves, the efficiency of an asymptotic model, or to compare different asymptotic models, one to each other.

1. Introduction

The aim of this paper is to present a new approach based on data mining techniques to evaluate the efficiency of numerical asymptotic models. Indeed, data mining techniques could help scientific computing, as they have proved to be efficient in other contexts, like in biology [8], medicine [11, 12], marketing [10], advertising and communications [6, 7].

We focus our presentation on asymptotic paraxial approximations to model charged particle beams. Indeed, solving the time-dependent Vlasov Maxwell equations, which is one of the most complete mathematical model for collisionless plasma or non-collisional beams, can lead to very expensive computations especially in a three-dimensional domain. Therefore, whenever possible, it is worthwhile to take into account the particularities of the physical problem to derive approximate asymptotic models leading to cheaper simulations.

However, despite some theoretical convergence results, it is not always easy to choose between two different approximate models, which sometimes can have the same accuracy, or to determine which terms to retain in the asymptotic expansion to get a sufficiently precise but not too expensive model. This paper propose a new approach, based on data mining techniques, to answer to this question.

Received June 4, 2012

2. **The Paraxial Model**

To solve charged particle beams or plasma physics problems for collisionless plasma or non-collisional beams, one of the most complete mathematical models is the time-dependent Vlasov-Maxwell system of equations (cf. [5]). However, the numerical solution of such a model requires a large computational effort. Therefore, whenever possible, it is worth to take into account the particularities of the physical problem to derive approximate models leading to cheaper simulations, see for instance [13]. We recall here the paraxial model we consider and explain how to derive it. Details can be found in [4].

Let us consider a beam of charged particles with a mass m and a charge q which moves inside a perfectly conducting cylindrical tube, the z-axis being the axis of the tube and the optical axis of the beam. Since the domain under consideration is a bounded axisymmetric three-dimensional domain, we will therefore use the cylindrical coordinates (r, θ, z). We denote by Ω the transverse section of the tube of radius R, by Γ its boundary, so that $\Gamma = \{(r, \theta, z);\ r = R\}$, and by $\boldsymbol{\nu}$ the unit outward normal to Γ. For the sake of simplicity, we assume here that there is no external fields.

Each particle of the beam can be characterized by its position $\mathbf{X} = (r, \theta, z)$ and its velocity $\mathbf{V} = (v_r, v_\theta, v_z)$ in the phase space $(\mathbf{X}, \mathbf{V})$. Assume that the beam is relativistic and non collisional, we introduce the momentum $\mathbf{P} = (p_r, p_\theta, p_z)$. Hence, the motion of these particles can be described in terms of particle distribution function $f(\mathbf{X}, \mathbf{P}, t)$ by the relativistic Vlasov equation. The quantity $\mathbf{F} = (F_r, F_\theta, F_z)$ denotes the electromagnetic Lorentz force given by $\mathbf{F} = q(\mathbf{E} + \mathbf{V} \times \mathbf{B})$, that describes how an electromagnetic field $\mathbf{E} = (E_r, E_\theta, E_z)$ and $\mathbf{B} = (B_r, B_\theta, B_z)$ acts on a particle with a given velocity. This electromagnetic field satisfies the axisymmetric Maxwell equations in the vacuum.

One then exploits the physical/geometrical properties of the problem to derive paraxial asymptotic models, which approximate the Vlasov-Maxwell system with a known accuracy. For high energy short beams, a paraxial relativistic model has been derived (cf. [9], [4]) based on the following assumptions:

- The beam is highly relativistic i.e., satisfies $\gamma \gg 1$,
- The dimensions of the beam are small compared to the longitudinal length of the device,
- The longitudinal particle velocities v_z are close to the light velocity c,
- The transverse particle velocities $(v_r^2 + v_\theta^2)^{1/2}$ are small compared to c.

Since $v_z \simeq c$ for any particle in the beam, we rewrite the Vlasov-Maxwell equations in the beam frame, which moves along the z-axis with the light velocity c. Hence we set $\zeta = ct - z, \quad v_\zeta = c - v_z$. As a consequence, the bunch of particles is evolving slowly in this frame. We denote by $\overline{v}$ the transverse characteristic velocity of the particles. Then, introduce a small parameter η defined by $\eta = \frac{\overline{v}}{c} \ll 1$. The paraxial model is derived by retaining the first four terms in the asymptotic expansion of the distribution function and the electromagnetic fields with respect to η, see [9], [4] for more details.

2.1. **The approximate models $\mathcal{M}_1$ and $\mathcal{M}_2$.** Using the asymptotic expansion described above, one can derive several "nested" approximate models of the Vlasov-Maxwell equations. Indeed, a paraxial model is derived by retaining the first terms in the asymptotic expansion of the distribution function and the electromagnetic fields with respect to η. Hence, one can consider the model denoted $\mathcal{M}_i$ in which the asymptotic function f is approximated by the i^{th} order expansion $f^{(0)} + \eta f^{(1)} + \cdots + \eta^i f^{(i)}$.

In this paper, our aim is to illustrate the possibility of data mining techniques applied on scientific computing (see also [1], [2]). Hence we will only consider and compare the 2 first models $\mathcal{M}_1$ and $\mathcal{M}_2$. Let us now expose them.

Following [4], [9] one can show that the i^{th} order asymptotic expansion of f (here $i = 1, 2$) is entirely determined from the knowledge of the $(i-1)^{\text{th}}$ order expansion of the electromagnetic Lorentz force $(F_r^{(i-1)}, F_\theta^{(i-1)}, F_z^{(i-1)})$. One thus obtain the following two models.

1. The model $\mathcal{M}_1$:
   In this model, the asymptotic expansion $f^{(0)} + \eta f^{(1)}$ is entirely determined from the zero order expansion $(F_r^{(0)}, F_\theta^{(0)}, F_z^{(0)})$ of the electromagnetic force. To compute them, it is sufficient to know the principal part of the transverse electromagnetic fields, which satisfies following [9], [4]:

$$\left\{ \begin{aligned} E_r^{(1)} &= cB_\theta^{(1)} = \frac{1}{\varepsilon_0\, r} \int_0^r \rho^{(1)} s\, ds\,, \\ E_\theta^{(1)} &= B_r^{(1)} = 0, \end{aligned} \right. \tag{2.1}$$

   whereas the corresponding forces have the following expression

$$F_r^{(0)} = q v_\zeta^{(1)} B_\theta^{(1)}\,, \qquad F_\theta^{(0)} = 0\,, \qquad F_z^{(0)} = q v_r^{(1)} B_\theta^{(1)}\,. \tag{2.2}$$

   Note that in this model, the longitudinal fields $E_z^{(1)}, B_z^{(1)}$ are identically zero.
2. The model $\mathcal{M}_2$:
   We also consider the model $\mathcal{M}_2$, in which the expansion $f^{(0)} + \eta f^{(1)} + \eta^2 f^{(2)}$ is entirely determined from the first order expansion $(F_r^{(1)}, F_\theta^{(1)}, F_z^{(1)})$ of the electromagnetic force. To characterize them, it is proved ([9], [4]) that the transverse electromagnetic fields have to verified the same equations as (2.1) for

the transverse fields, but at the order 2, namely

$$\left\{\begin{aligned} & E_r^{(2)} = cB_\theta^{(2)} = \frac{1}{\varepsilon_0\, r}\int_0^r \rho^{(2)} s\, ds\,, \\ & \\ & E_\theta^{(2)} = B_r^{(2)} = 0, \end{aligned}\right. \tag{2.3}$$

supplemented with, for the longitudinal fields:

$$\left\{\begin{aligned} & \frac{\partial E_z^{(2)}}{\partial r} = \frac{\partial B_\theta^{(2)}}{\partial t}\,, \\ & \\ & E_z^{(2)}(r=R) = 0\,, \end{aligned}\right. \quad \text{and} \quad \left\{\begin{aligned} & \frac{\partial B_z^{(2)}}{\partial r} = \mu_0 J_\theta^{(2)}\,, \\ & \\ & \int_0^R B_z^{(2)} r dr = 0\,. \end{aligned}\right. \tag{2.4}$$

Finally, the corresponding forces are expressed:

$$F_r^{(1)} = q(v_\theta^{(2)} B_z^{(2)} + v_\zeta^{(2)} B_\theta^{(2)})\,, \quad F_\theta^{(1)} = -qv_r^{(2)} B_z^{(2)}\,, \quad F_z^{(1)} = q(E_z^{(2)} + v_r^{(2)} B_\theta^{(2)})\,. \tag{2.5}$$

In the next section, our aim is to perform a sensitivity analysis of these two models via data mining techniques; For instance to understand what the second order in the model $\mathcal{M}_2$ *practically brings to the simulation results* over what could be obtained by the model $\mathcal{M}_1$. In such Vlasov Maxwell simulations, one is often interested in the particle motion. For this reason, we will use the particle velocities as significant variables in the data mining analysis. Note that the choice of these variables can not be automatic: it will always depend on the human expertise that will decide what to be explored in the data. Following [3], [4], we introduce for each model $\mathcal{M}_i$, $(i = 1, 2)$, the variable

$$\delta v_z^{(i)} := \gamma |v_z^{(i)} - v_{z,aver}^{(i)}|$$

for the longitudinal velocity, where the index $_{aver}$ denotes in each case the average velocity.

3. Data Mining and Decision Trees

3.1. Segmentation by decision tree.

Data Mining goal is to discover hidden or *a priori* unknown facts contained in databases. Decision trees [14] belong to the supervised data mining tools to process the so-called *segmentation*, whose aim is to constitute homogeneous subgroups inside a given population. For this purpose, we select in a given database a variable y to be explained, named the *target variable*. Then, we assume a formal unknown relation $y = f(x_1, x_2, \ldots, x_n)$ between the target y and n other variables $x_1, x_2, \ldots, x_n$ of the database, called the *predictors*. Basically based on the minimization of the standard deviation of the target variable y, an algorithm of segmentation determines the resulting optimized homogeneous subgroups. This results to a decision tree.

A decision tree is then a tree composed by different subgroups (called *nodes*) of the initial population (called *root node*). At each level of the tree, these *nodes* are obtained by the segmentation algorithm, by identifying among the predictor variables $(x_1, x_2, \dots, x_n)$, the most discriminating one, regarding the *homogeneity degree* of the resulting *nodes*.

This process stops when the splitting is not feasible: either any new subgroup cannot be found to be more homogeneous than the previous one or the resulting segmentation is composed by insignificant subgroups, typically composed by a two low number of individuals.

3.2. **Application to our data.** In the database we considered, the data are computed by the help of finite differences method and described numerical approximations of problem (2.1-2.5) solutions. Then, at each time step and for each node of the concerned space grid, we get a set of variables which are:

$$v_r^{(i)}, v_\theta^{(i)}, v_\zeta^{(i)}, E_r^{(i)}, E_z^{(i)}, B_z^{(i)}, \rho^{(i)}, J_\theta^{(i)}, F_r^{(i-1)}, F_\theta^{(i-1)}, F_z^{(i-1)}, \delta v_r^{(i)}, \delta v_z^{(i)}, \quad (i = 1, 2)\,. \tag{3.1}$$

Therefore, we organize the data such that each row of the database (or ”individual”, the devoted terminology in database language) contains the information of the above variables for a given time step and for a space node. Because our objectives are to appreciate the improvement of the results depending on the order of the asymptotic development of problem (2.1-2.5) solutions, we introduce the two following variables to define an appropriate target variable:

- Let us denote by X a given variable to be computed by the two asymptotic models $\mathcal{M}_1$, $\mathcal{M}_2$. We set $X^{(1)}$ its value computed by the model $\mathcal{M}_1$ and $X^{(2)}$ its corresponding value from the model $\mathcal{M}_2$. The first variable we consider here, $\omega_{1,2}$, is defined by:

$$\omega_{1,2} = \left| \frac{X^{(1)}}{X^{(2)}} \right| . \tag{3.2}$$

  It measures the weight of the model $\mathcal{M}_1$ in the model $\mathcal{M}_2$, regarding the variable X.
- From the variable $\omega_{1,2}$, we are able to introduce our target variable $\omega_{1,2}^{(3CLS)}$, obtained by splitting the distribution of $\omega_{1,2}$ into three equal classes of individuals: Low, Medium and High.

Without any *a priori* on the meaning of Low or High contributions of the model $\mathcal{M}_1$ in the model $\mathcal{M}_2$, it is usual to define the categorial variables $\omega_{1,2}^{(3CLS)}$ as follows: the three classes of individuals are determined based on an equal number of individuals for each category, (Low, Medium and High).

The purpose of our analysis being to point out the role of the electromagnetic fields in the sensitivity between the models $\mathcal{M}_1$ and $\mathcal{M}_2$, the dependent variables - that is to say the predictors - we kept to explain the above two classes are the non vanishing electromagnetic components, the charge and current densities and the particles velocities computed by the model $\mathcal{M}_2$, namely

$$v_r^{(2)}, v_\theta^{(2)}, v_\zeta^{(2)}, E_r^{(2)}, E_z^{(2)}, B_z^{(2)}, \rho^{(2)}, J_\zeta^{(2)} \,. \tag{3.3}$$

As a complement to take into account the coupling with the Vlasov equation, we also add to the above list of predictors the components of the Lorentz force involved in the model $\mathcal{M}_2$, that is

$$F_r^{(1)}, F_\theta^{(1)}, F_z^{(1)} \,. \tag{3.4}$$

4. Result: Comparison Between the Model $\mathcal{M}_1$ and $\mathcal{M}_2$

As shown on Figure 1 , the precision of the decision tree given by the risk estimate is equal to 5.2 percent. So, 94.8 percent of data are correctly classified by the model of segmentation computed by the corresponding decision tree. One more time, the quality level of the decision tree is very high and it let us to use it with a high level of confidence.

The first segmentation which appears on the decision tree (Fig. 1) highlights the most discriminated predictor variable in the set of all the available potential predictors. More precisely, $F_z^{(1)}$ is detected as this predictor with a corresponding computed optimal threshold equal to 37.24.

Based on the information avalaible within the decision tree dedicated to δv_z, we get:

- $F_z^{(1)}$ is the most discriminate variable. This was an expected result, since $E_z^{(1)} = 0$ in the model $\mathcal{M}_1$, which implies that the main difference between $F_z^{(1)}$ and $F_z^{(0)}$ is essentialy due to $E_z^{(2)}$ (see Eq.(2.5)).
- The second most important predictor identified at the root of the decision tree is $E_r^{(2)}$ which an unexpected feature since the corresponding component $E_r^{(1)}$ is non zero.
- On the contrary, $B_z^{(2)}$ appears as a non significant predictor even if $B_z^{(1)}$ was null in the model $\mathcal{M}_1$.

5. Conclusion

In this paper, we have presented a new approach based on data mining techniques and statistical tools applied to scientific computing. We focused our study to the specific case of an asymptotic paraxial approximation to model ultrarelativistic particles. Our aim was to determine the role of the different powers in the asymptotic expansion, restricted to the models $\mathcal{M}_1$ and $\mathcal{M}_2$. As we have considered an approximate model of the Vlasov-Maxwell equations, we have chosen $\delta v_z^{(i)}$, $(i = 1, 2)$, as the main physical variable.

Beyond the particular case we treated in this paper, we suggest that data mining techniques can be applied to the analysis of any scientific computations as it is applied in a lot of other domains.

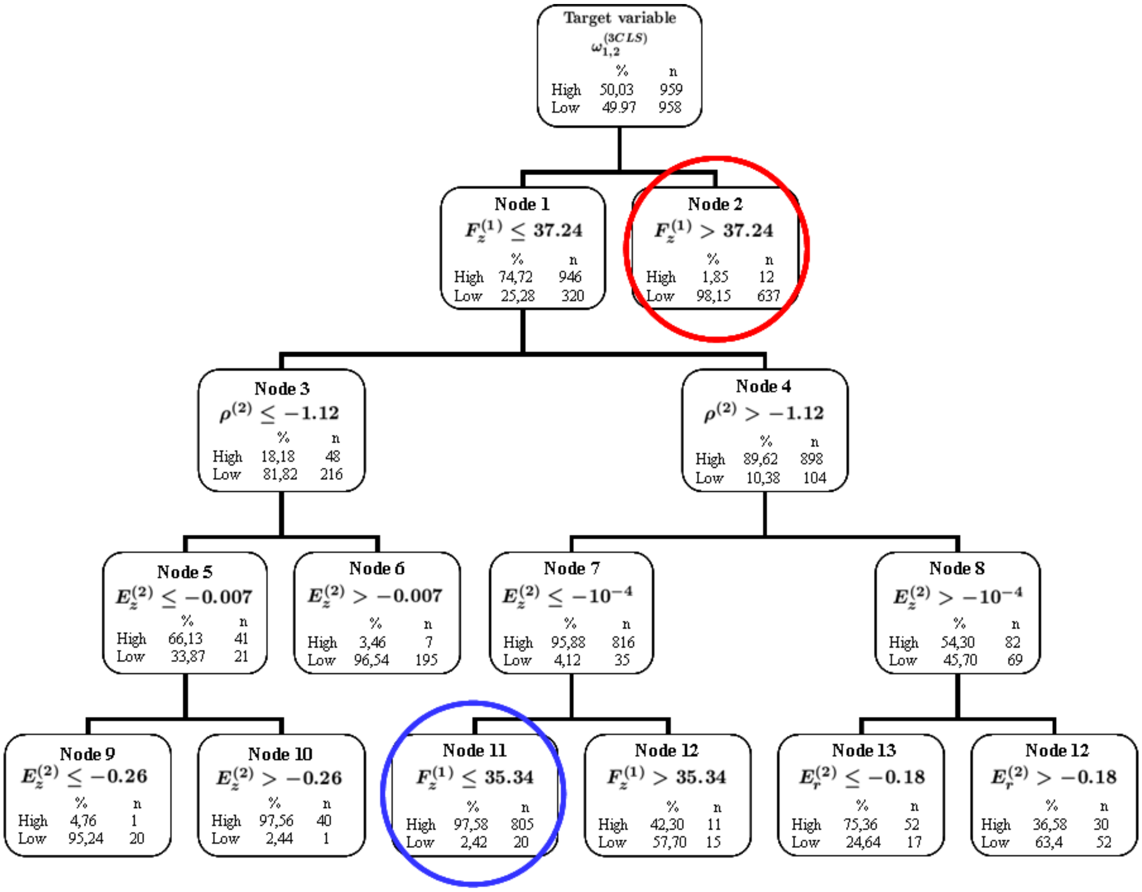

FIGURE 1. Decision Tree related to δv_z.

REFERENCES

[1] F. Assous, J. Chaskalovic, Data mining techniques for numerical approximations analysis: A test case of asymptotic solutions to Vlasov Maxwell equations, *Comptes Rendus Mécanique*, **338**, 305-310 (2010).

[2] F. Assous, J. Chaskalovic, Data mining techniques for scientific computing: application to asymptotic paraxial approximation to model ultrarelativistic particles, *J. Comput. Phys.*, **230**, 4811-4827 (2011).

[3] F. Assous, F. Tsipis, A PIC Method for Solving a Paraxial Model of Highly Relativistic Beams, *J. Comput. Appl. Math* **227-1**, pp. 136-146, 2009 (2008).

[4] F. Assous, F. Tsipis, Numerical paraxial approximation for highly relativistic beams, *Comput. Phys. Comm.* **180** 1086–1097, (2009).

[5] C.K. Birdsall and A.B. Langdon, *Plasmas Physics via Computer Simulation* (New York: Mac.Graw-Hill, 1985).

[6] J. Chaskalovic, A new approach in Media/Marketing Databases explorations for application in E-business, *National Congress of IREP*, Paris, 1999.

[7] J. Chaskalovic, A. Vanheuverzwyn, Innovation in estimations: A reliable approach for radio audience indicators, Proc. Esomar, WM^3 2007, Dublin, 306 juin 2007. bibitemHaTF09 T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics, 2nd Ed. 2009).

[8] O. Kulski, J. Chaskalovic *et al.*, Explicative factors for prognostics IIU: exploration on 2089 cycles done with statistical and data mining tools, *9th Meeting of the French Federation of the Reproduction Studies*, Palais des Congrés - Paris, 2004

[9] G. Laval, S. Mas-Gallic, P.-A. Raviart, Paraxial approximation of ultrarelativistic intense beams, *Numer. Math.*, **69(1)**, 33–60 (1994).

[10] R. Lefebure, G. Venturi, *Data Mining*, Eyrolles, 2001.

[11] XL Nguyên, J. Chaskalovic *et al.*, Residual subjective daytime sleepiness under CPAP treatment in initially somnolent apnea patients: a pilot study using data mining methods, *Sleep Med. 9 (5)*, pp. 511-516, 2007.

[12] XL Nguyên, J. Chaskalovic *et al.*, Insomnia symptoms and CPAP compliance in OSAS patients: A descriptive study using Data Mining methods, *Sleep Med. Jul. 2*, 2010.

[13] P.A. Raviart, E. Sonnendrucker, A hierarchy of approximate models for the Maxwell equations, *Numer. Math.*, **73(3)**, 329–372 (1996).

[14] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing Company , 2001.