C. R. Mecanique 338 (2010) 305–310

Contents lists available at ScienceDirect

Comptes Rendus Mecanique

www.sciencedirect.com

Data mining techniques for numerical approximations analysis: A test case of asymptotic solutions to the Vlasov–Maxwell equations

Méthodes de data mining pour l'analyse d'approximations numériques : Le cas de solutions asymptotiques des équations de Vlasov–Maxwell

Franck Assous [a,b], Joel Chaskalovic [a,c,*]

[a] *Ariel University Center, 40700 Ariel, Israel*
[b] *Bar-Ilan University, 52900 Ramat-Gan, Israel*
[c] *IJLRDA, University Pierre and Marie Curie, 4, place Jussieu, 75252 Paris cedex 05, France*

ARTICLE INFO

Article history:
Received 4 June 2010
Accepted after revision 6 July 2010
Available online 21 July 2010

Keywords:
Computer science
Data mining
Vlasov–Maxwell equations

Mots-clés :
Informatique, algorithmique
Data mining
Équations de Vlasov–Maxwell

ABSTRACT

We propose a novel approach that consists in using data mining techniques to perform a sensitivity analysis of approximate models. We give here the example of asymptotic solutions to Vlasov–Maxwell equations, obtained with a paraxial model for relativistic short beams. This new heuristic approach offers new potential applications to treat numerical solutions to mathematical models.

RÉSUMÉ

On propose une nouvelle approche mettant en oeuvre des techniques de data mining, pour effectuer une étude de sensiblité de modèles approchées. L'exemple présenté dans cette Note traite d'une approximation paraxiale des équations de Vlasov–Maxwell, valable pour des faisceaux courts relativistes. Cette nouvelle méthodologie ouvre de nouvelles perspectives pour l'analyse d'approximations numériques des modèles mathématiques appliqués aux sciences de l'ingénieur.

Version française abrégée

Les nombres entre parenthèses renvoient à la version anglaise. L'analyse classique des résultats de méthodes numériques se limite souvent à la description d'isovaleurs ou de toute autre forme de graphiques. Cette exploitation « de bas niveau » résulte de la masse importante des données à analyser. Ceci est d'autant plus vrai pour les problèmes instationnaires et tridimensionnels, où, à chaque noeud du maillage, sont calculées les approximations de champs scalaires, vectoriels ou tensoriels.

Dans cette Note, on propose une méthodologie originale d'analyse des données à l'aide de techniques de data mining, qui ont fait leur preuves dans les domaines du marketing et de la communication ou de la biologie, tous producteurs de volumétries « boolimiques » de données.

* Corresponding author.
E-mail addresses: franckassous@netscape.net (F. Assous), j.chaskalovic@free.fr (J. Chaskalovic).

1631-0721/$ – see front matter
doi:10.1016/j.crme.2010.07.001

On se limite ici à comparer les solutions asymptotiques du premier et du second ordre ((1)–(2)) (notées $\mathcal{M}_1$ et $\mathcal{M}_2$), des équations de Vlasov–Maxwell modélisant un faisceau court de particules relativistes.

Le principe de la méthode repose sur la constitution d'une base de données regroupant l'ensemble des données numériques obtenus à chaque pas de temps et pour chaque noeud du maillage, soit plus de 125 000 enregistrements et 30 variables à analyser en colonne.

On introduit les variables $\omega_{1,2}$ et $\omega_{1,2}^{(3CLS)}$, où $\omega_{1,2}$ décrit la quote-part du modèle $\mathcal{M}_1$ au sein du modèle $\mathcal{M}_2$ relativement à la quantité $\delta v_r^{(i)}$ $(i = 1, 2)$ définie à la fin de la Section 2. $\omega_{1,2}^{(3CLS)}$ représente la variable catégorielle définie par trois classes équiréparties (« Low », « Medium » et « High ») définie à partir de la variable continue $\omega_{1,2}$.

Ces trois classes permettent de qualifier les éléments de la base de données caractéristiques d'une « forte » variation du modèle $\mathcal{M}_1$ par rapport au modèle $\mathcal{M}_2$ ou, *a contrario*, ceux qui correspondent aux enregistrements présentant un ordre de grandeur équivalent entre les deux modèles.

Cette caractérisation est effectuée par segmentation à l'aide d'un arbre de décision réalisé sous IBM SPSS Modeler.

Le principe de la segmentation est de constituer des sous-groupes au sein de la population initiale de la base de données afin de minimiser l'écart-type de la variable cible $\omega_{1,2}^{(3CLS)}$.

Pour des arbres de décision binaires, l'algorithme de segmentation identifie, à chaque niveau de l'arbre de décision, la variable explicative *var* la plus discriminante (celle qui permet de produire à ce niveau de l'arbre des sous-groupes tel que l'écart-type de la variable à expliquer soit minimal), assortie d'un seuil τ tels que les sous-groupes correspondants à $var \leqslant \tau$ et $var > \tau$ présentent une homogénéité supérieure (donc, d'un écart-type moindre), par rapport à celle du noeud situé en amont. L'analyse de l'arbre de décision obtenu a conduit aux résultats suivants :

- La composante $E_z^{(2)}$ est identifiée comme la variable explicative la plus discriminante permettant de différencier les « Low » $\omega_{1,2}$ des « High » $\omega_{1,2}$ (Fig. 1) : L'amélioration produite par le modèle $\mathcal{M}_2$ est principalement dûe à la présence de la composante $E_z^{(2)}$, alors que celle-ci n'apparait pas dans le modèle $\mathcal{M}_1$.
- La deuxième variable explicative la plus discrimante est la composante $E_r^{(2)}$. Cette propriété est, *a contrario* de la première, inattendue dans la mesure où le développement asymptotique de la solution $\mathcal{M}_1$ présente une composante $E_r^{(1)}$.
- La composante $B_z^{(1)}$ n'apporte aucune contribution significative dans l'amélioration de la solution du modèle $\mathcal{M}_2$, bien que n'apparaissant pas dans le modèle $\mathcal{M}_1$.

Ces résultats illustrent l'importance effective des différentes variables dans le développement asymptotique de la solution du problème (1)–(2). Cette méthodologie d'exploration des données par des méthodes de data mining devrait s'avérer féconde pour l'analyse d'approximations numériques de modèles mathématiques appliqués aux sciences de l'ingénieur.

1. Introduction

Solving a problem described by partial differential equations, as for instance the time-dependent Vlasov–Maxwell equations, can lead to very expensive computations. Therefore, whenever possible, one takes 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 determine which terms to retain in the asymptotic expansion to get a sufficiently precise but not too expensive model. In other terms, the asymptotic models are often difficult to compare directly one to the other. In this Note, we propose to use data mining techniques to perform a sensitivity analysis of approximate models.

As an application of this novel approach, we give here the example of asymptotic solutions to Vlasov–Maxwell equations, obtained with a paraxial model for relativistic short beams [1]. Our method directly deals with numerical results of simulations and try to understand what each order of the asymptotic expansion brings to the simulation results over what could be obtained by other lower-order or less accurate means?

For this purpose, we consider a second-order accurate formulation (model $\mathcal{M}_2$) and a first-order one (model $\mathcal{M}_1$) derived from the same asymptotic expansion (see [1,2]). Beyond the mathematical analysis or theoretical results of approximation ($\mathcal{M}_2$ is more "accurate" than $\mathcal{M}_1$!), we propose here to check a given order accuracy *on the numerical results themselves*.

One could assess this accuracy with a series of numerical test cases where the parameters of the asymptotic expansion vary. But this can lead to tedious computations. First, since there are often several parameters, so that one of them varies whereas other aspects are fixed.

Then, we often get a huge quantity of results to compare. Moreover, contrary to the data mining techniques, such sensitivity studies are not always supported by a theory.

We propose here an alternative based on the use of data mining techniques. In addition, the numerical results are generally only partially exploited. Typically, one exploits the time evolution of a variable, a snapshot of a given quantity at a fixed time, or some specific diagnostics.

For instance for the Vlasov–Maxwell simulation, one depicts the position–velocity phase space at a given time. In the approach we propose here, the numerical results considered as a "database" are entirely used: the data mining techniques allow us to exploit all the time steps at all the nodes of the mesh.

Beyond this particular study, we think that this new heuristic approach can be useful in a lot of other domains of applications.

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. However, the numerical solution of such a model requires a large computational effort. Therefore, whenever possible, we take into account the particularities of the physical problem to derive approximate models leading to cheaper simulations.

In previous works, an asymptotic paraxial model has been introduced [1] and numerically solved [2] for high energy short beams. We recall it briefly here. Consider a beam of charged particles with a mass m and a charge q which moves inside a perfectly conducting cylindrical tube of radius R, the z-axis being the axis of the tube. As the domain we consider is axisymmetric, we will use the cylindrical coordinates (r, θ, z).

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 phase space.

Assume that the beam is relativistic and non-collisional and introduce the momentum $\mathbf{P} = (p_r, p_\theta, p_z) = \gamma m(v_r, v_\theta, v_z)$, where $\gamma = (1 - \mathbf{V}^2/c^2)^{-1/2}$. 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.

In addition, the particle distribution function, and the data are assume to be independent of θ. Hence, the function f satisfies the axisymmetric Vlasov equation, in which the force $\mathbf{F} = (F_r, F_\theta, F_z)$ denotes the electromagnetic Lorentz force 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. The charge and the current densities ρ and $\mathbf{J} = (J_r, J_\theta, J_z)$ are classically obtained as the zeroth and the first moments of the distribution function f.

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, the model has been derived 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, the Vlasov–Maxwell equations are rewritten in the beam frame, which moves along the z-axis with the light velocity c. Hence we set $\zeta = ct - z$, $v_\zeta = c - v_z$.

As a consequence, the bunch of particles is evolving slowly in this frame. We denote by $\bar{v}$ the transverse characteristic velocity of the particles. Then we introduce a small parameter η defined by $\eta = \frac{\bar{v}}{c} \ll 1$.

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.

In this Note, we will restrict ourselves to the comparison of the two first asymptotic models. Let us denote by $\mathcal{M}_1$ the model in which the asymptotic expansion $f^0 + \eta f^1$ is entirely determined from the zeroth-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

$$E_r^{(1)} = cB_\theta^{(1)} = \frac{1}{\varepsilon_0 r}\int_0^r \rho^{(1)} s\,\mathrm{d}s, \qquad E_\theta^{(1)} = B_r^{(1)} = 0, \quad \text{with}$$
$$F_r^{(0)} = q v_\zeta^{(1)} B_\theta^{(1)}, \qquad F_\theta^{(0)} = 0, \qquad F_z^{(0)} = q v_r^{(1)} B_\theta^{(1)} \tag{1}$$

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, the transverse electromagnetic fields have to verify Eqs. (1) (changing the $^{(1)}$ in $^{(2)}$) supplemented with, for the longitudinal fields

$$\frac{\partial E_z^{(2)}}{\partial r} = \frac{\partial B_\theta^{(2)}}{\partial t}, \qquad \frac{\partial B_z^{(2)}}{\partial r} = \mu_0 J_\theta^{(2)}, \quad \text{with}$$
$$F_r^{(1)} = q\big(v_\theta^{(2)} B_z^{(2)} + v_\zeta^{(2)} B_\theta^{(2)}\big), \qquad F_\theta^{(1)} = -q v_r^{(2)} B_z^{(2)}, \qquad F_z^{(1)} = q\big(E_z^{(2)} + v_r^{(2)} B_\theta^{(2)}\big) \tag{2}$$

The aim is now 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 (here the radial velocity v_r) as characteristic variables in the data mining analysis.

Following [2], we introduce for each model $\mathcal{M}_i$ $(i = 1, 2)$, the variables $\delta v_r^{(i)} := \gamma(v_r^{(i)} - v_{r,aver}^{(i)})$ and $\delta v_z^{(i)} := \gamma(v_z^{(i)} - v_{z,aver}^{(i)})$, the index $_{aver}$ denoting the average velocity.

3. Data mining analysis

3.1. Data mining methods

Data mining goal is to discover hidden or *a priori* unknown facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results.

Decision trees [3] belong to the supervised data mining tools to process segmentation. The purpose of a segmentation is to constitute homogeneous subgroups inside a given population regarding a *target* variable which is to be explained versus *predictor* variables. This is processed by an algorithm of segmentation which is basically a minimization of the standard deviation for the concerned target variable.

In the case of the segmentation we considered in our study, the target variable is a categorial one; It describes the belonging to one of several classes which characterizes a given level (called "Low", "High" and so on), of a target variable to be explained.

A decision tree is then composed by different subgroups (called *nodes*) of the initial population (called *root node*). These *nodes* are obtained with the segmentation algorithm by identifying among the predictor variables the most discriminant one regarding the homogeneity degree of the resulting *nodes*.

Each split of the segmentation divides a given *node* into several *nodes* (here, in our study into two nodes which is the specific case of a binary decision trees), based on the most discriminant predictor variable *var* such that the left resulting *node* obeys to the inequality $var \leqslant \tau$ and the right one to $var > \tau$ (τ being a threshold optimally computed by the algorithm of segmentation).

This process stops when the splitting is not feasible: either any new subgroup can be found to be more homogeneous than the previous one or the resulting segmentation is composed by insignificant subgroups. The path from the *root node* to each *leaf* (a terminal node) defines a succession of inequalities on the predictor variables that characterizes the solutions belonging to the *leaf* with a certain risk which depends on the percentage of misclassified solutions in the *leaf*.

By choosing the *leaves* that predict the membership to the class of interest with the minimum risk, one is able to characterize this class with a set of "rules" at minimum risk.

The database we considered is composed by data computed with the help of finite differences method and described numerical approximations of problem (1)–(2) solutions. Then, at each time step and for each node of the concerned space grid, we get a set of variables which are:

$$v_r^{(i)}, v_\theta^{(i)}, v_\zeta^{(i)}, E_r^{(i)}, E_z^{(i)}, B_z^{(i)}, \rho^{(i)}, J_\theta^{(i)}, F_r^{(i-1)}, F_\theta^{(i-1)}, F_z^{(i-1)}, \delta v_r^{(i)}, \delta v_z^{(i)} \quad (i = 1, 2) \tag{3}$$

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.

Considering all the 100 time steps and the 1250 space nodes, the database we treated was composed by more than 125 000 rows and 30 variables to be analyzed.

Because our objectives are to appreciate the improvement of the results depending on the order of the asymptotic development of problem (1)–(2) solutions, we introduce the two following variables:

- $\omega_{1,2} = \frac{\delta v_r^{(1)}}{\delta v_r^{(2)}}$ which measures the weight of the model $\mathcal{M}_1$ in the model $\mathcal{M}_2$, regarding the variable δv_r, where $\delta v_r^{(i)}$ is defined above.
- $\omega_{1,2}^{(3CLS)}$ defines a ternary variable processed by binning the distribution of $\omega_{1,2}$ into three equal classes of $\omega_{1,2}$: Low, Medium and High.

Without *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 count of rows for each modality of the ternary variable.

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 we kept to explain the target variable $\omega_{1,2}^{(3CLS)}$ are the non-vanishing electromagnetic components $E_r^{(2)}$, $E_z^{(2)}$, $B_z^{(2)}$.

Moreover, to take into account the coupling with the Vlasov equation, we also kept the components of the Lorentz force $F_r^{(i)}$, $F_\theta^{(i)}$, $F_z^{(i)}$. Note that the other available variables mentioned in (3) could be considered in further developments.

Taking into account these choices, we performed the Kolmogorov–Smirnoff test to evaluate the differences, if any, between the distributions of the dependent variables which are suspected to be different between the models $\mathcal{M}_1$ and $\mathcal{M}_2$.

Therefore, the results show (see Table 1) that the corresponding asymptotic significance (2-tilde) is less than 5/1000. This clearly demonstrates a real and significant difference between the distributions for each variable computed in the

Table 1
Kolmogorov–Smirnov test results.

	$E_r^{(2)}$	$E_z^{(2)}$	$B_z^{(2)}$	$F_r^{(1)}$	$F_\theta^{(1)}$	$E_z^{(1)}$
Kolmogorov–Smirnov Z	7.145	10.925	3.398	7.145	5.123	10.925
Asymp. Sig. (2-tailed)	0.004	0.002	0.005	0.004	0.003	0.003

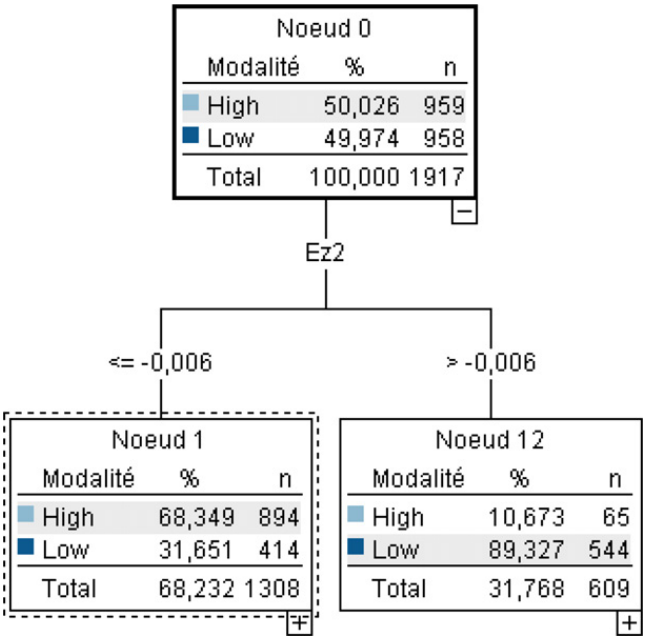

Fig. 1. First level of the decision tree.

two introduced classes ("Low" and "High"). Then, the variable $\omega_{1,2}^{(3CLS)}$ is analyzed by a decision tree whose purpose is to characterize each modality of the ternary variable.

Particularly, we do want to identify the features of the "Low class" in opposition with the "High class" of $\omega_{1,2}^{(3CLS)}$. This is motivated by the fact that the "Low class" of $\omega_{1,2}^{(3CLS)}$ describes the elements of the database which are weakly characteristic of $\delta v_r^{(1)}$, and therefore, rather than specific of $\delta v_r^{(2)}$. Then, they correspond to those that bring a significant contribution by the model $\mathcal{M}_2$.

Finally, to obtain a significant description between the "Low" group to the "High" group, we eliminated the "Medium" group of $\omega_{1,2}^{(3CLS)}$ from the Database, to give to the future segmentation a better accuracy level.

3.2. Result

We analyze the decision tree we processed under IBM SPSS Modeler to model the ternary variable $\omega_{1,2}^{(3CLS)}$. Before giving the most important results we would like to emphasize that the exploration was processed on the all available time steps and for all the space nodes on the considered grid.

As one can see on the decision tree restricted to its first segmentation (see Fig. 1), the most discriminant predictor variable in the set of all the available predictors is $E_z^{(2)}$ with a corresponding identified threshold equal to -0.006. This means that the group of the "Low $\omega_{1,2}$" is mainly different from the group of the "High $\omega_{1,2}$", if one splits the whole involved population regarding the found threshold of $E_z^{(2)}$.

Hence, the model $\mathcal{M}_2$ brings a significant improvement in the computation of $\delta v_r^{(2)}$ primarily due to the presence of the variable $E_z^{(2)}$ in the asymptotic enriched model $\mathcal{M}_2$. This is an expected result since the E_z component is not present in the model $\mathcal{M}_1$, we mean $E_z^{(1)} = 0$.

On the other hand, the decision tree brings a classification of the predictors from the most discriminant to the less one. One gets that after the variable $E_z^{(2)}$, one must take into account with $E_r^{(2)}$ which participates to the relevant improvement of the asymptotic development modeling.

This feature is unexpected since the corresponding component $E_r^{(1)}$ was non-zero in the model $\mathcal{M}_1$. Because of the strong nonlinearity of the partial differential system, this was not a result that would be expected before this exploration.

Furthermore, the classification also pointed out that the variable $B_z^{(2)}$ does not appear as a predictor which brings a significant improvement of the asymptotic modeling. Contrarily to the component $E_r^{(2)}$ above, recall here that $B_z^{(1)}$ was null in the model $\mathcal{M}_1$. Nevertheless, it does not bring a significant improvement in the second-order model $\mathcal{M}_2$.

4. Conclusion

In this Note, we have proposed a new method applied to scientific computing and we implemented it on Vlasov–Maxwell equations. About this specific case, one can ask to himself, as a complement of this result, what the improvement of the third order would bring for all of the above features, or at which order would it be relevant to stop the asymptotic expansion?

In a more general way, we suggest that data mining techniques can be applied to the analysis of scientific computations as it is in a lot of other applications; This is already the case in marketing and communication or in biology.

This Note shows that it can be also done to investigate the relevance and/or the quality of numerical simulations, particularly when a large quantity of data – as for instance with massively parallel computers – is available.

Nevertheless, several problems remain open: up to now, this approach remains heuristic and mathematical analysis to asses the method would be useful. Moreover, data mining tools cannot (yet) be an automatic tool to analyze data. But, in anyway, the “expert” will always be necessary to “guide” the analysis and to explore the data.

References

[1] G. Laval, S. Mas-Gallic, P.-A. Raviart, Paraxial approximation of ultrarelativistic intense beams, Numer. Math. 69 (1) (1994) 33–60.
[2] F. Assous, F. Tsipis, Numerical paraxial approximation for highly relativistic beams, Comput. Phys. Comm. 180 (2009) 1086–1097.
[3] L. Rokach, O. Maimon, Data Mining with Decision Trees: Theory and Applications, World Scientific Publishing Company, 2001.