Automated analysis for detecting beams in laser wakefield simulations

Daniela M. Ushizima, Oliver Rübel, Prabhat, Gunther H. Weber, E. Wes Bethel, Cecilia R. Aragon
Cameron G.R. Geddes, Estelle Cormier-Michel, Bernd Hamann
Lawrence Berkeley National Laboratory,
Berkeley, CA 94720, USA
{dushizima, oruebel, prabhat, ghweber, ewbethel, craragon, cgrgeddes, emichel}@lbl.gov
hamann@cs.ucdavis.edu

Peter Messmer
Tech-X Corporation,
5621 Arapahoe Ave. Suite A, Boulder, CO 80303
messmer@txcorp.com

Hans Hagen
International Research Training Group "Visualization of Large and Unstructured Data Sets"
University of Kaiserslautern, Germany
hagen@informatik.uni-kl.de

Abstract

Laser wakefield particle accelerators have shown the potential to generate electric fields thousands of times higher than those of conventional accelerators. The resulting extremely short particle acceleration distance could yield a potential new compact source of energetic electrons and radiation, with wide applications from medicine to physics. Physicists investigate laser-plasma internal dynamics by running particle-in-cell simulations; however, this generates a large dataset that requires time-consuming, manual inspection by experts in order to detect key features such as beam formation. This paper describes a framework to automate the data analysis and classification of simulation data. First, we propose a new method to identify locations with high density of particles in the space-time domain, based on maximum extremum point detection on the particle distribution. We analyze high density electron regions using a lifetime diagram by organizing and pruning the maximum extrema as nodes in a minimum spanning tree. Second, we partition the multivariate data using fuzzy clustering to detect time steps in a experiment that may contain a high quality electron beam. Finally, we combine results from fuzzy clustering and bunch lifetime analysis to estimate spatially confined beams. We demonstrate our algorithms successfully on four different simulation datasets.

1. Introduction

The radiation pressure of an intense laser pulse fired into plasma allows laser wakefield accelerators (LWFAs) to generate self-trapping and acceleration of particles to relativistic speed in plasma density wakes. LWFAs are of interest because they are able to achieve very high particle energies within a relatively short distance when compared to traditional electromagnetic accelerators. The VORPAL [13] simulation code is used to model experiments such as those performed at the LOASIS facility at LBNL [8, 7, 11], and is useful in helping to gain deeper understanding of the phenomena observed in experiments, as well as to help formulate and optimize methodologies.

Particle-in-cell (PIC) simulation codes describe the physics of a new operating regime, where an initial trapped bunch of electrons loads the wake, forming a bunch of electrons isolated in phase space. At the dephasing point, as the bunch begins to outrun the wake, the particles are then concentrated near a single energy [8], when a high quality beam can form. PIC codes model the dynamics of particles in a simulation window that travels at approximately the speed of light, showing the position and velocities of the particles as well as fields at specified time intervals.

Identifying beam formation and quality are key problems in the analysis of laser wakefield simulation data. Currently, physicists must manually inspect 2D plots of the entire

dataset, visually determine adequate parameters to select a subset of particles (corresponding to the beam) and further analyze this subset. The procedure requires physicists to laboriously examine massive data over many time steps in different plots, which is a time-consuming process. Current simulation datasets are typically between 1GB and 100GB in size, and it is anticipated that future datasets will be of the order of TBs. As the number of datasets and dataset sizes increase, there is an emerging need for routines to automatically mine datasets for interesting structures such as beams.

Development and application of machine learning methods for scientific data mining is a growing field; however, few publications address similar efforts to the proposed framework on laser wakefield simulations. Bagherjeiran et al.[1] presented a comprehensive report on applying graph-based techniques for orbit classification in plasma simulations. They use the KAM classifier [15] to label points and components in single and multiple orbits. Love et al.[12] conduct an image space analysis of coherent structures in plasma simulations. They use a number of segmentation and region-growing techniques to isolate regions of interest in orbit plots. Both approaches analyze particle accelerator data, targeting the system dynamics in terms of particle orbits. However, their techniques do not address particle dynamics as a function of time or inspect the behavior of bunches of particles.

We apply signal processing and machine learning techniques to a very different problem: we are interested in searching for electrons with high acceleration and with spatial coherence in a time-dependent, large and complex scientific data set created by a numerical simulation of a laser wakefield particle accelerator. The high-quality beam must be picked out from a large field of high-energy particles.

We describe our approach and implementation details in Section 2. Section 3 presents the results obtained with our integrated approach of combining data visualization and analysis with classification of electrons from simulation time series. We conclude with discussions and future directions in Section 4.

2 Methods

We address particle dynamics as a function of time, inspecting the behavior of bunches of particles across the simulation for later combination with a clustering algorithm, so that we can estimate which are the accelerated particles that form the electron beam. Our focus is on designing a framework to aid physicists in detecting beam formation and characterizing beams. These beams are collimated groups of particles exhibiting high momentum along the x-axis, parallel to the propagation direction of the laser pulse and having a small spread in the spatial-energy dimensions.

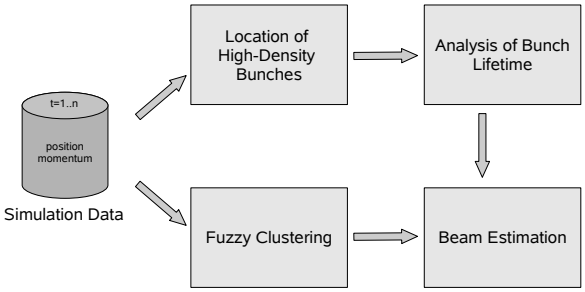


Figure 1. Processing pipeline

Figure 1 illustrates our processing pipeline, explained in the following subsections. First, we compute the location of high-density bunches of high-energy particles for each timestep via stationary point detection on probability density functions (pdf) of the particles along the laser propagation axis (see Section 2.1). Detected points are organized in a graph to characterize the lifetime of the detected high-energy bunches (see Section 2.2). In Section 2.3, we perform fuzzy clustering for each timestep to detect particles forming a potential beam. Based on the information from the bunch lifetime analysis and the fuzzy clustering, we can estimate the most energetic beam (see Section 2.3).

We illustrate our methods using 2D datasets, containing 5 variables: 2 spatial variables (x, y), 2 momentum variables (px, py) and a unique particle identifier. Table 1 presents details of the datasets used in our tests.

Dataset	Particles (10^3)	Timesteps	Total Size (Gb)
A	0.4	37	1.3
B	1.6	35	4.5
C	0.4	37	1.3
D	3.2	45	11

Table 1. Tested simulation Datasets.

2.1. Locating High-Density Bunches

The beams of interest are characterized by high density of high-energy particles in small spatial regions. Therefore, our objective is to identify potential groups of particles which exhibit this desirable property. In order to do this, for each recorded time step of the simulation we compute a histogram of particle distribution in the x-direction. Since energy is proportional to the momentum in the x-direction (px), we do not use the entire particle dataset, but instead choose a subset such that $px > 1e10$, which eliminates low energy particles without compromising the beam detection

procedure. Given this subset, we estimate the pdf, $f(x)$, of particles according to their position in the simulation window. We detect the maximal extremum (maximal turning point or relative maximum) at a point $x = X_o$, by calculating $df(x)/dx$, for $f(x)$ changing from positive to negative [10]. Differentiability is guaranteed by using a Gaussian smoothing kernel, where we use a rule of thumb to choose the bandwidth: 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power, as suggested by [14]. The maximal extremum points represent potential beam-point candidates for the next step in our analysis.

2.2. Analysis of Bunch Lifetime

Once potential beam-point candidates have been identified for each time step, we represent each maximum extremum point as a node n in a graph. First, we construct an incidence matrix for a graph by discretizing the spatial axis into bins and noting a non-zero entry where a candidate (n) was found. We repeat this process for different time steps, stacking up each time step as rows of a matrix. Note that we use the relative position of the particles in the moving simulation window to align multiple time steps.

Then we run a minimum spanning tree (MST) algorithm [2], that penalizes connections among nodes in the same time step. Next, we prune the graph by eliminating candidates that are connected by long arcs, since we do not expect high density regions to move erratically or by large amounts. This is scientifically motivated, as the high energy particles travel at approximately the speed of light, which is also the speed of the simulation window. Assuming that the time lag between recorded time steps is small enough to capture physical phenomena such as bunching and dephasing, we eliminate disconnected nodes.

Consequently, we now have a representation of the temporal history of the high-density particle regions, where a node n_i is connected to n_j if they are from different time steps. The distance between nodes is smaller than a value $d = 2\mu$m, the approximate size of a beam, as suggested by the physicists, so that the lifetime diagram represent temporally stable high density regions.

2.3. Fuzzy Clustering and Beam Estimation

As discussed in Sec. 2.1, a high particle density is not sufficient to identify a high quality beam - both the spatial and momentum features play an important role in classifying a bunch of particles as a beam of interest.

Unsupervised algorithms are appropriate for "data mining" applications, where the information content of a large database is not known beforehand, but can emerge during the partitioning process. Without supervision, nonhierarchical clustering methods can use an optimization model to classify inter-point distances and dissimilarity data. The objective is to minimize total dissimilarity amongst all objects and the corresponding most representative objects.

In this context, our approach focuses on searching for the primary beam particles: the beam is confined to a small spatial region, having high energy in the px-direction. Classical clustering algorithms would try to assign each data point to exactly one cluster [5], but our problem requires relaxing this condition so that each particle has some graded or fuzzy membership in each cluster.

We use the R package `cluster`, which contains the algorithm `fanny` for fuzzy analysis clustering, to identify two groups of particles in a multivariate space, defined by x, y, px, py. First, we normalize all the spatial and momentum variables before calculating dissimilarities: the most frequent normalization strategy consists in the transformation of the original data such that the new feature set is now guaranteed to have zero mean and unit standard deviation [6]. The algorithm performs fuzzy C-means clustering, regarding dissimilarities between observations using the squared Euclidean distances and then minimizes the objective function

$$F = \sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^{m} u_{jv}^{m} d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^{m}} \qquad (1)$$

where v is a cluster, n is the number of observations, k is the number of clusters, m is the membership exponent and $d(i,j)$ is the dissimilarity between observations i and j, u_{iv} is the membership of observation i to cluster v [9]. The dissimilarity measure appears as an $L1$ norm; it finds "medoids" (median-based centroids) instead of ordinary centroids. The minimization algorithm is based on direct application of the Lagrange multiplier approach with Kuhn-Tucker conditions [4].

Finally, we compare the estimated-beam cluster, calculated for each time step independently, with the beam-point candidates, i.e. the points represented by nodes that remained after the pruning process (Sec. 2.2), for the respective time step. If the most accelerated group of particles, expected to be at the rightmost cluster, and a beam-point overlap, we conclude that we have located the beam and return the particles in the current cluster assignment.

3 Results

Physicists need auxiliary mechanisms to identify high quality beams and track their accelerated particles along simulations. We propose an automated framework to detect time steps in a simulation that may contain a high quality electron beam. We test the methods discussed in the Sec. 2 on different particle simulation datasets, illustrating

each step of the pipeline for the simulation dataset A. Similar analysis was conducted for three other datasets: B, C and D, then we only present their plots for the final beam segmentation result.

The proposed framework contains data processing and machine learning algorithms from R project [3]. The R project, or simply R, is a free multi-platform software environment for statistical computing, containing useful packages for data analysis, visualization and machine learning. We perform our computations on a Dell Optiplex 755 Intel Core Duo 3GHz; the processing took ≈1-4 minutes for each time step using 2 Gb RAM memory.

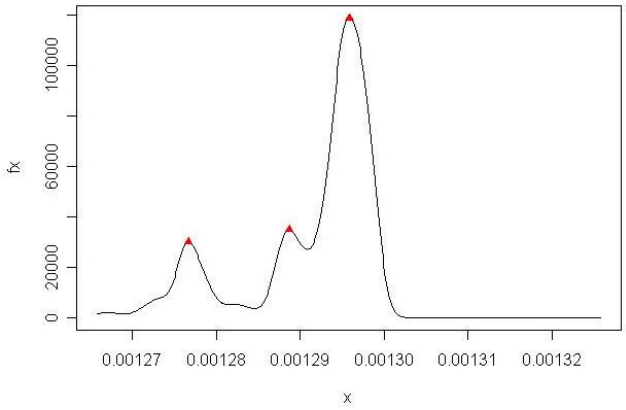


Figure 2. Locating high-density bunches for Dataset A at a single time step: detection of three beam-point candidates (peaks).

As discussed in Sec.2.1, we locate high-density particle regions for each time step. Figure 2 presents x-position of candidates to beam locations for the dataset A. Note that the smoothed pdf presents distinct beam-point candidates that correspond to high-density regions, collected for the lifetime diagram analysis. Both Figure 2 and 4 illustrate the same time step in database A, where the beam is most visible.

We implement a novel procedure to analyze the history of particles (Sec2.2): we start with a tree representation of high density points in the whole time series, followed by calculating the minimum spanning tree to connect the nodes. We then use a pruning procedure, which guarantees no node connection within the same time step and a distance greater than $d = 2\mu$m. Figure 3 shows the temporal tree-based representation of the particle history after pruning ("lifetime"), conveying only the most likely candidates.

The tree representation combined with the clustering output yields two important pieces of information: (a) detection of the time step containing high energy particles by checking for overlapping; (b) estimation of a beam containing particles that behave similarly, according to their spatial

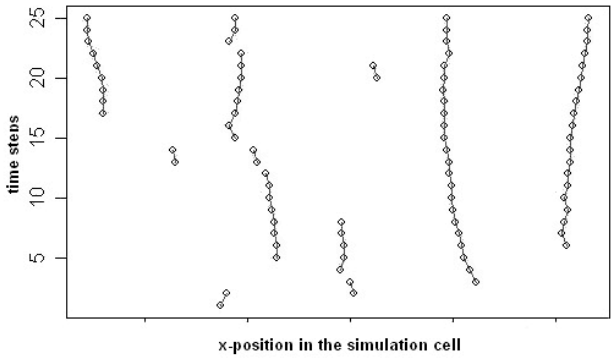


Figure 3. Life time representation for Dataset A: particle history as a pruned MST with likely branches and connected nodes.

coordinates and energy attributes.

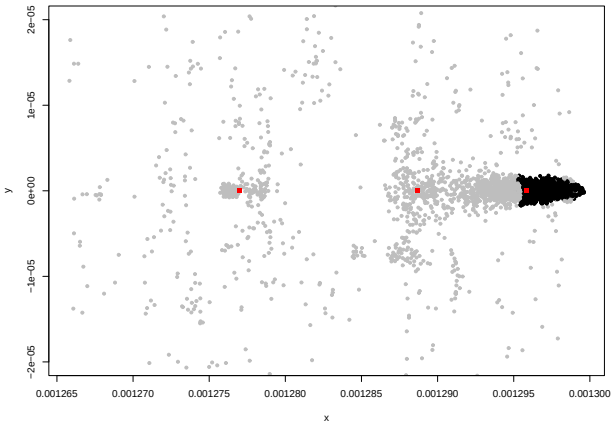
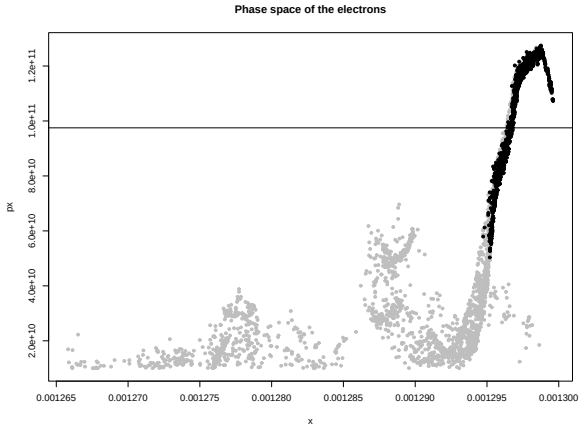


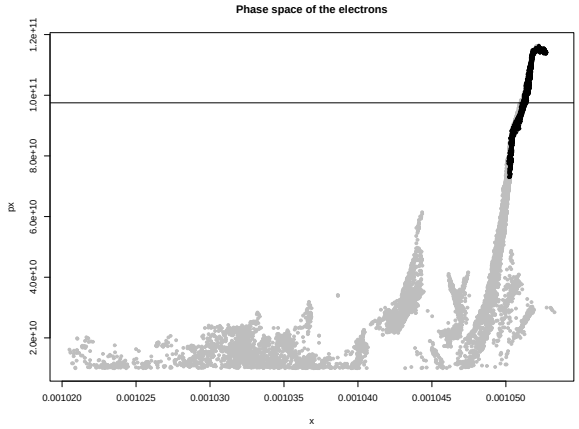
Figure 4. Beam-point candidates (Fig. 2) combined with fuzzy clustering: beam-points (squares), beam particles (in black) and non-beam particles (in gray).

Figures 4 and 5 show classification results from clustering, narrowing the membership to 0.7, so that the cluster closer to the rightmost high-density point will contain only particles which have 70% probability to belong to this cluster. This idea is analogous to the search for the core particles in the multivariate distribution given by x, y, px, py for the primary beam.

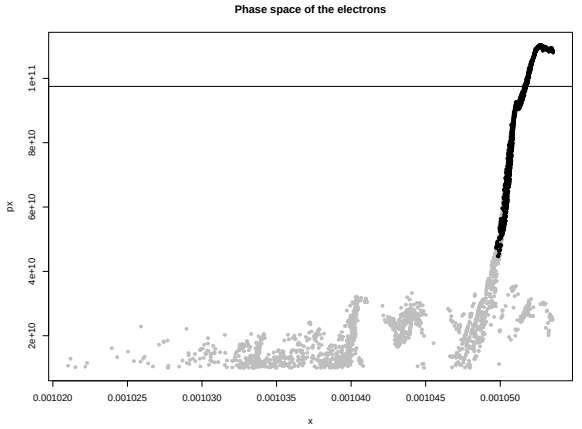
The beam cluster is labeled in black in Figure 4, showing the (x, y) position of the particles in the simulation window. Figure 5 illustrates the phase space for the different datasets in a time step where the beam is most visible. Physicists have noticed that the estimated beams completely enclose



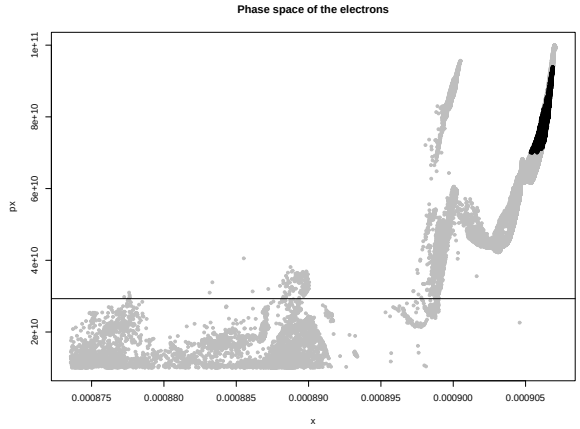
(a) Dataset A



(b) Dataset B



(c) Dataset C



(d) Dataset D

Figure 5. Clustering result in the phase space scatterplot for different datasets. All particles with $px > 1e10$ are shown in gray and the particles detected by our analysis in black.

the high energy bunch of particles.

3.1. Validation

Figure 5 shows the resulting phase space diagrams using the proposed methodology for four different 2D datasets. The datasets vary in size as well as overall behavior. In practice a physicist is interested in finding single, distinct, and compact particle bunches of high momentum and low energy spread in the data.

While no automated detection system was available, thresholding in px was a common practice to help identify high energy particles and the beam of interest. In addition, a researcher investigates movies of a variety of plots to determine an appropriate timestep and threshold values. As illustrated in Figures 5(a)- 5(c), for simulations A, B, C a researcher may choose, for example, a threshold of $px > 9.75e10$. In these examples only a single high energy beam is formed and a single threshold in px is sufficient to isolate the beam-particles. However, simulation D contains secondary structures, formed behind the beam and a single threshold in px is not sufficient to isolate the beam-particles (see Figure 5(d)). Therefore, a single threshold may result in a selection that is too large and only multiple thresholds in px, x and possibly y can isolate the beam of interest. One main deficiency of thresholding is that it is arbitrary and time consuming, requiring manual inspection of the dataset.

Figure 5 demonstrates that our method detects a single, distinct, bunch of high energy particles in all four datasets. Both thresholding and the proposed algorithm depict particles of the beam for dataset A, B, and C, but our method is capable of identifying subsets of condensed particles among

the highly accelerated particles without requiring user interaction. For experiments where no high quality beam is formed, neither thresholding nor the proposed method can accurately detect beam particles as show for dataset D; the most condensed structure is located on the depression between the first and second peak ($px =\approx 4.5e10$), as shown in Figure 5(d). While our algorithm detects a single, spatially distinct particle bunch of high energy, it is driven toward finding distinct bunches of high energy and the clustering result encloses the second highest quality bunch.

The proposed approach presents promising results toward automated analysis of laser wakefield particle simulations. Our method is capable of extracting particle beams from the a large set and isolating time steps without user interaction, in contrast to manual thresholding wherein an expert is required to manually investigate the whole data in order to identify the beam of interest.

4 Conclusions and Future Work

In this paper, we propose a novel method for identifying and tracking density patterns in particle acceleration data. We generate a novel lifetime diagram of high-density bunches of electrons. We use a minimum spanning tree representation and prune that to recover high density peaks which are further combined with fuzzy clustering to segment the beam particles. Four different datasets illustrate our validating results by comparing to a manual selection by an expert. Also, we observe that low quality beams, not formed by the highest energy particles, may not be detected, hence algorithm improvements are necessary.

Future work must address the particle history for statistically significant centers using both spatial and energy components. In addition, beam-point candidates could be used by other clustering algorithms. We plan to quantify the beam quality in terms of intra-cluster measures and establish a relationship between the number of the particles in a beam given their membership. Correlating the beam formation to the underlying electromagnetic field remains an open problem. Analysis of 3D simulations have started, which includes spatial dimension z and its momentum pz into account. Besides the massive 2D simulation dataset generation, we expect to parallelize our computations to handle the significantly larger number of particles and overall size of the 3D datasets.

5 Acknowledgments

This work was supported by the Director, Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC) program's Visualization and Analytics Center for Enabling Technologies (VACET) and by the U.S. DOE Office of Science, Office of High Energy Physics, grant DE-FC02-07ER41499, through the COMPASS SciDAC project.

This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

We also thank the VORPAL development team for ongoing efforts in development and maintenance on a variety of supercomputing platforms, including those at NERSC.

References

[1] A. Bagherjeiran and C. Kamath. Graph-based methods for orbit classification. In *SDM*, 2006.

[2] C. F. Bazlamaçci and K. S. Hindi. Minimum-weight spanning tree algorithms a survey and empirical study. *Comput. Oper. Res.*, 28(8):767–785, 2001.

[3] M. J. Crawley. *The R Book*. John Wiley and Sons, Ltd, 2007.

[4] R. Dave and S. Sen. Robust fuzzy clustering of relational data. *IEEE Transactions on Fuzzy Systems*, 10(6):713–727, 2002.

[5] R. O. Duda, P. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, Ltd, 2001.

[6] C. L. F and M. J. R. C. *Shape analysis and classification:Theory and Practice*. CRC Press, Boca Raton, 2001.

[7] C. G. R. Geddes. *Plasma Channel Guided Laser Wakefield Accelerator*. PhD thesis, University of California, Berkeley, 2005.

[8] C. G. R. Geddes, C. Toth, J. van Tilborg, E. Esarey, C. Schroeder, D. Bruhwiler, C. Nieter, J. Cary, and W. Leemans. High-Quality Electron Beams from a Laser Wakefield Accelerator Using Plasma-Channel Guiding. *Nature*, 438:538–541, 2004. LBNL-55732.

[9] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, Ltd, 1990.

[10] E. Kreyszig. *Advanced Engineering Mathematics*. John Wiley and Sons, Ltd, 2006.

[11] LOASIS. http://loasis.lbl.gov/.

[12] N. S. Love and C. Kamath. Image analysis for the identification of coherent structures in plasma. In *Applications of Digital Image Processing. Edited by Tescher, Andrew G.. Proceedings of the SPIE*, volume 6696, 2007.

[13] C. Nieter and J. R. Cary. VORPAL: A Versatile Plasma Simulation Code. *J. Comput. Phys.*, 196(2):448–473, 2004.

[14] B. Silverman. *Density Estimation*. Chapman and Hall, 1986.

[15] K. M.-K. Yip. *KAM: A System for Intelligently Guided Numerical by Computer*. MIT Press, 1991.