Separation of Photoelectrons via Multivariate Maxwellian Mixture Model

Genta Ueno[1], Nagatomo Nakamura[2], and Tomoyuki Higuchi[1]

[1] The Institute of Statistical Mathematics, Tokyo 106-8569, Japan
gen@ism.ac.jp
[2] Sapporo Gakuin University, Hokkaido 069-8555, Japan

Abstract. Electron velocity distribution obtained by direct spacecraft observation in space is contaminated by photoelectrons. The photoelectrons are generated due to the solar ultraviolet ray, and are regarded as artificial noise from a viewpoint of scientific research. We propose a method for separating photoelectron component from ambient electron component. Our method uses multivariate normal mixture model, whose parameters are determined via the Expectation-Maximization (EM) algorithm. Initial parameters of the EM algorithm are computed through the classification of the velocity space by a spherical surface of some arbitrary radius.

1 Introduction

The process of knowledge discovery begins with data acquisition and ends with identification of a new pattern in data. Between the start and the goal, the process includes various steps what we roughly call "data analysis." In the scheme proposed by Fayyad et al. [2], the process consists of (1) data selection, (2) data preprocessing, (3) data transformation, (4) data mining (hypothesis generation), and (5) hypothesis interpretation / evaluation. Creation and development of the computational strategy for such steps enable us to reduce the time in achieving the knowledge discovery, and are indispensable in dealing with large database.

We have demonstrated that the multivariate normal mixture model is an effective tool for characterizing an observation of three-dimensional space plasma velocity distributions [6, 7]. The normal distribution is called as the Maxwellian distribution in the plasma physical field. That is, when multiple peaks exist in the observed velocity distribution, these peaks can be well represented by the multiple Maxwellian distributions that compose the mixture model [4]. We applied the mixture model to the ion velocity distribution and determined the parameters of the model through the Expectation-Maximization (EM) algorithm [1, 3]. This procedure is regarded as a step of "data mining" in the scheme of Fayyad et al.[2].

In this paper, we present that the similar procedure can be applied to the "preprocessing step" of the analysis of electron velocity distributions with minor modification. Since a spacecraft in sunlight is irradiated by the solar ultraviolet ray, photoelectrons are produced from illuminated surface material. The spacecraft is then charged to positive potential relative to the ambient plasma, which attracts the photoelectrons emitted from the surface. When the electron measurement is carried out in such an environment, returned photoelectrons are detected together with the ambient natural electrons what is originally expected to be observed. Since the amount of those photoelectrons is comparable or even larger than that of ambient electrons, it is difficult to obtain the real information from the data. This difficulty have prevented the progress of the quantitative study of electron dynamics.

However, we found that the two-component Maxwellian mixture model can represent the photoelectron and the ambient electron by the two component mixture model. While the algorithm used is similar to that in the previous work [6], a new algorithm has been developed in the part of setting the initial parameters.

2 Data

We used electron velocity distribution obtained by the Low Energy Particle Energy-per-charge Analyzer (LEP-EA) on board the Geotail spacecraft. LEP-EA measured three-dimensional velocity distributions by classifying the velocity space into 32 for the magnitude of the velocity, 7 for elevation angles, and 16 for azimuthal sectors (Figure 1).

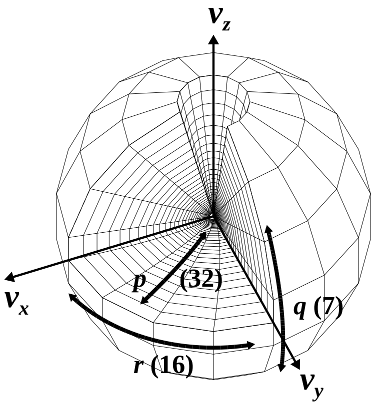

Fig. 1. Classes for observation of an electron velocity distribution with LEP-EA (RAM B mode)

We define a probability function of observed electron velocity $\boldsymbol{v}_{pqr}$ [m/s]:

$$f\left(\boldsymbol{v}_{pqr}\right) = \frac{f_0\left(\boldsymbol{v}_{pqr}\right) d\boldsymbol{v}_{pqr}}{\sum_{p,q,r} f_0\left(\boldsymbol{v}_{pqr}\right) d\boldsymbol{v}_{pqr}}, \tag{1}$$

where $f_0\left(\boldsymbol{v}_{pqr}\right)$ [s^3/m^6] is an observed electron velocity distribution function, and $d\boldsymbol{v}_{pqr}$ is the class interval whose class mark is $\boldsymbol{v}_{pqr}$. Subscription p, q and r

are indicators of the magnitude of the velocity, elevation angle, and azimuthal sector, and they take integers $p = 1, \cdots, 32$, $q = 1, \cdots, 7$, and $r = 1, \cdots, 16$.

3 Method

3.1 Multivariate Maxwellian Mixture Model and EM Algorithm

We approximate the probability function (1) by the mixture model composed of the sum of two multivariate Maxwellian distributions:

$$f\left(\boldsymbol{v}_{pqr}\right) \simeq \sum_{i=\mathrm{ph,am}} n_i g_i \left(\boldsymbol{v}_{pqr} | \boldsymbol{V}_i,\ \mathbf{T}_i\right), \tag{2}$$

where n_i is the mixing proportion of the Maxwellians ($\sum_{i=\mathrm{ph,am}} n_i = 1,\ \ 0 \leq n_i \leq 1$). Notations "ph" and "am" mean photoelectron and ambient electron, respectively. Each Maxwellian g_i is written as

$$\begin{aligned} & g_i \left(\boldsymbol{v}_{pqr} | \boldsymbol{V}_i,\ \mathbf{T}_i\right) \\ = & \left(\frac{m_e}{2\pi}\right)^{3/2} \frac{1}{\sqrt{|\mathbf{T}_i|}} \exp\left[-\frac{m_e}{2}\left(\boldsymbol{v}_{pqr} - \boldsymbol{V}_i\right)^T \mathbf{T}_i{}^{-1} \left(\boldsymbol{v}_{pqr} - \boldsymbol{V}_i\right)\right], \end{aligned} \tag{3}$$

where m_e [kg] is the electron mass, $\boldsymbol{V}_i$ [m/s] is the bulk velocity vector and $\mathbf{T}_i$ [J] is the temperature matrix of i-th Maxwellian. The log-likelihood of this mixture model becomes

$$l\left(\theta\right) = N \sum_{p,q,r} f\left(\boldsymbol{v}_{pqr}\right) \log \sum_{i=\mathrm{ph,am}} n_i g_i \left(\boldsymbol{v}_{pqr} | \boldsymbol{V}_i,\ \mathbf{T}_i\right), \tag{4}$$

where $\theta = \left(n_{\mathrm{ph}},\ \boldsymbol{V}_{\mathrm{ph}},\ \boldsymbol{V}_{\mathrm{am}},\ \mathbf{T}_{\mathrm{ph}},\ \mathbf{T}_{\mathrm{am}}\right)$ denotes the all unknown parameters, and N is the total number of the particle count.

Partially differentiate (4) with respect to $\boldsymbol{V}_i$ and $\mathbf{T}_i^{-1}$ $(i = \mathrm{ph}, \mathrm{am})$ and put them equal to zero, we obtain the equations that should be satisfied by the parameters as maximum likelihood estimators. Utilizing these equations, we estimate the unknown parameters through the iteration of the EM algorithm with regarding posterior probabilities as unmeasured data [5, 6]. We finish the iteration when the log-likelihood and unknown parameters become unchanged in the iteration.

3.2 Initial Parameters of EM Algorithm

To reduce an iteration of the EM algorithm, a proper setting for the initial parameters is desirable. We used the k-means algorithm for setting the initial parameters for the iteration of the EM algorithm in the previous work [6]. However, since the k-means algorithm is a clustering algorithm for an exclusive division, it is not applicable in setting initial parameters for a mixture model whose bulk velocities are close to each other.

Now we approximate an observation of electron distribution by photoelectrons and ambient electrons. The parameters are expected to satisfy

$$n_{\mathrm{ph}} > n_{\mathrm{am}}, \tag{5}$$

$$\boldsymbol{V}_{\mathrm{ph}} \simeq 0, \tag{6}$$

$$|\boldsymbol{V}_{\mathrm{am}}| < \sqrt{2\mathrm{tr}\,\mathbf{T}_{\mathrm{ph}}/3m_e} < \sqrt{2\mathrm{tr}\,\mathbf{T}_{\mathrm{am}}/3m_e}. \tag{7}$$

This is the case when the k-means algorithm does not work well. We then adopt the following method suitable for such a distribution.

1. Divide the 32 classes about the magnitude of velocity (radial direction in the velocity space) into two groups by a certain boundary of radius R ($R = 2, 3, \cdots, 31$).
2. Compute mixing proportion, bulk velocity vectors, and temperature matrices for both groups by usual moment calculation procedure.
3. Set these value as the initial parameters of the EM algorithm.

4 Application

The top panel of Figure 2 shows an observation of electron velocity distribution which was obtained in the time interval 1420:00–1420:12 on January 16, 1994. Displayed two lines are densities in the velocity space along the v_x and v_y axes. We find high density around the origin ($v_x = v_y = 0$), which corresponds to the photoelectrons.

When applying the two-component Maxwellian mixture model to the data, we obtained photoelectron component and ambient electron component separately as shown in the bottom-left and bottom-central panels of Figure 2, respectively. The estimated parameter are given in Table 1. The sum of the two components are also shown in the bottom-right panel.

Table 1. Estimated parameters for the two-Maxwellian mixture model in the time interval 1420:00–1420:12 on January 16, 1994. The value of n is the mixing proportion multiplied by the total number density

Electron	n [/cc]	V_x [km/s]	V_y	V_z	T_{xx} [eV]	T_{xy}	T_{xz}	T_{yy}	T_{yz}	T_{zz}
photo	3.819	−534	257	123	7	0	0	7	0	7
ambient	0.064	270	−100	−87	144	−4	1	131	1	122

In setting the initial parameters of the EM algorithm, it may matter how to select the cutting radius R which potentially classify the data into photoelectron and ambient electron components. We found, however, that the results after the iteration of the EM algorithm are the same in most R selection. In this example, we can obtain the same result for $R = 2$ to 25.

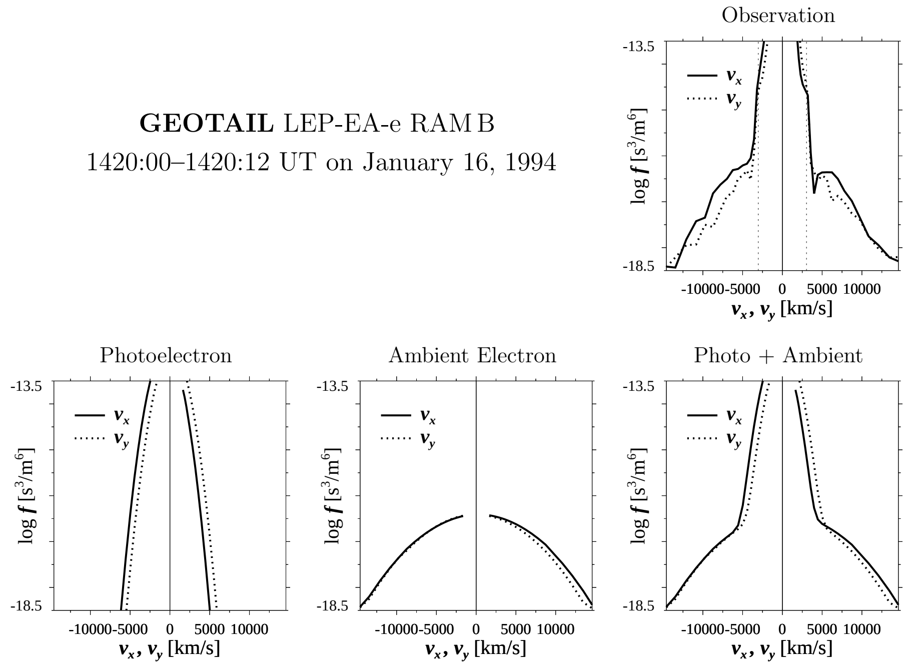

Fig. 2. Observation of electron velocity distribution along the v_x and v_y axes between 1440:00–1440:12 UT on January 16, 1994 (top). Bottom panels are estimated components for photoelectron (left), ambient electron (center), and the sum of the both (right). Two vertical broken lines in the top panel indicate electron velocities equivalent to the spacecraft potential at this time interval

5 Discussion

Traditionally, the spacecraft potential was utilized to decompose the electron data into photoelectron component and ambient electron component. When the spacecraft potential is $\phi\,[\mathrm{V}]$, the equivalent electron speed $v_e = \sqrt{2e\phi/m_e}\,[\mathrm{m/s}]$, where $e\,[\mathrm{C}]$ is the elementary electric charge. This means that a photoelectron particle whose speed is less than v_e cannot escape from the spacecraft and is pulled back to the spacecraft. Therefore, density of particles slower than v_e would be contributed by photoelectrons as well as ambient electrons. On the assumption that the photoelectrons were distributed within $|\boldsymbol{v}| \le v_e$, they thus did not use the density of the slow particles and interpolated the density of slow speed particles by the density of particles faster than v_e. However, the equivalent electron speed v_e is not so accurate as an indicator of the photoelectron distribution. In the same time interval (1420:00–1420:12 UT on January 16, 1994), the spacecraft potential $\phi = 26.09\,\mathrm{V}$ and then $v_e = 3029\,\mathrm{km/s}$, which is shown in the top panel of Figure 2 as $\pm v_e$ by two vertical broken lines. The two lines are located at smaller velocities than our expectation ($v_e \sim 4000\,\mathrm{km/s}$), and will give an inappropriate interpolation.

Since our method works automatically with less computational burden, it can compute the macroscopic quantity of the ambient electrons (n_{am}, $\boldsymbol{V}_{\mathrm{am}}$, and $\mathbf{T}_{\mathrm{am}}$) on board the spacecraft. It will be useful under the limited transmission resources from the spacecraft due to the telemetry constraint.

Acknowledgments. We would like to thank Prof. T. Mukai for providing us with Geotail/LEP data. This work was carried out under the auspices of JSPS Research Fellowships for Young Scientists.

References

1. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the *EM* Algorithm. J. Roy. Statist. Soc. B **39** (1977) 1–38
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine **17** (1996) 37–54
3. McLachlan, G. J., Krishnan, T.: The EM Algorithm and Extensions. Wiley Series in Probability and Statistics. John Wiley and Sons, New York (1997)
4. McLachlan, G. J., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. John Wiley and Sons, New York (2000)
5. Nakamura, N., Konishi, S., Ohsumi, N.: Classification of Remotely Sensed Images via Finite Mixture Distribution Models (in Japanese with English abstract). Proceedings of the Institute of Statistical Mathematics **41** (1993) 149–167
6. Ueno, G., Nakamura, N., Higuchi, T., Tsuchiya, T., Machida, S., Araki, T.: Application of Multivariate Maxwellian Mixture Model to Plasma Velocity Distribution Function. In: Arikawa, S., Morishita, S. (eds.): Discovery Science. Lecture Notes in Computer Science, Vol. 1967. Springer-Verlag, Berlin Heidelberg New York (2000) 197–211
7. Ueno, G., Nakamura, N., Higuchi, T., Tsuchiya, T., Machida, S., Araki, T., Saito, Y., Mukai, T.: Application of Multivariate Maxwellian Mixture Model to Plasma Velocity Distribution Function. to appear in J. Geophys. Res. (2001)