ICA-BASED ACCELERATION OF PROBABILISTIC LATENT COMPONENT ANALYSIS FOR MASS SPECTROMETRY-BASED EXPLOSIVES DETECTION

Yohei Kawaguchi, Masahito Togami, Hisashi Nagano, Yuichiro Hashimoto,
Masuyuki Sugiyama, and Yasuaki Takada

Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
yohei.kawaguchi.xk@hitachi.com

ABSTRACT

We propose a new method to separate mass spectra into components of each chemical compound for explosives detection. The conventional method based on probabilistic latent component analysis (PLCA) is effective because the method can solve the problems of non-negativity and non-orthogonality by using sparsity of the domain of explosives detection. However, the convergence of the method is slow, and the calculation time is long. In order to solve this problem, the proposed method makes use of independent component analysis (ICA) in the initialization process. Experimental results indicate that the convergence of the proposed method is accelerated, and total calculation time is decreased.

Index Terms— Mass spectrometry, Blind source separation, Probabilistic latent component analysis (PLCA), Independent component analysis (ICA), Sparsity

1. INTRODUCTION

The threat of improvised explosive devices has become a serious problem for all countries because the procedures and recipes for making them are freely available on the Internet. To prevent terrorist attacks, we have developed a walkthrough portal explosives detector that consists of a high-throughput vapor sampling portal, a high-sensitivity atmospheric pressure chemical ionization source, and a high-selectivity linear ion trap mass spectrometer [1]. The mass spectrometer measures the intensity corresponded to the number of ions for each mass-to-charge ratio (*m/z*). The *m/z* series of intensities are called the mass spectrum. The detector observes the time series of the mass spectra continuously, and it detects characteristic patterns of explosives traces from the mass spectra data.

In mass spectra of the explosives detection system, explosives compounds, other chemical compounds, and the chemical background are mixed with each other. Thus, it is necessary to separate the mass spectra into the different compounds. The system does not know what kind of chemical compounds can be measured in advance, and so the task of the system is a blind source separation (BSS) problem. There are many researches that employ BSS for mass spectra separation, such as principal component analysis (PCA) [2] and independent component analysis (ICA) [3, 4]. Because PCA and ICA impose the orthogonality and the independence respectively without constraints of non-negativity, and so these methods are not fit to mass spectrometry domain. Thus these methods suffer from performance degradation. Recently, there have been several researches that apply non-negative matrix factorization (NMF) [5, 6] and probabilistic latent component analysis (PLCA) [7] to the area of mass spectrometry. These approaches have the desirable feature that the estimated components are guaranteed to be non-negative, and the approaches have the advantage that distortion is not caused by negative values. Furthermore, the conventional method based on PLCA [7] can solve the uncertainty problem of the number of compounds by using statistical knowledge as sparsity priors. However, the convergence of the method is slow, and the total calculation time is long. Thus, the method can not run in real time, and it is difficult to apply the method to the explosives detection system in practice.

In this paper, we propose a acceleration method for PLCA. We focus on that ICA can stably obtain a solution near the correct solution, and its speed is fast. Thus, the proposed method makes use of ICA in the initialization process of PLCA. Experimental results indicate that the convergence of the proposed method is accelerated, and total calculation time is decreased.

2. PROBLEM STATEMENT

The input signal is the time series of mass spectra $x(t,m)$, where t is the index of a time, and m is the index of *m/z*. T is the number of the time index, and M is the number of the index of *m/z*. $x(t,m)$ is modeled as follows,

$$x(t,m) = \sum_k c(k|t)s(m|k), \tag{1}$$

where k is the index of a compound basis, K is the number of the kinds of the compounds in the air, $c(k|t)$ is the intensity of the k-th compound in the time index t, and $s(m|k)$ is the time-invariant spectral basis component of the k-th compound.

In this paper, we estimate the unknown variables $c(k|t)$ and $s(m|k)$ from the known variables $x(t,m)$. This problem equals to the blind source separation problem. In addition, we consider the following three conditions of the explosives detection system. First, $s(m|k)$ is non-negative for all compounds and *m/z* because mass spectra represent the number of ions for each *m/z*; second, we can not assume the orthogonality between different basis component $s(m|k)$ because different components are mixed into the same *m/z* in real environments; third, the number of compounds in the air K is unknown because suspected chemical compounds and the chemical background change depending on the environment at the time and place.

3. PLCA-BASED MASS SPECTRA SEPARATION

In this section, we explain about the conventional mass spectra separaion method based on PLCA [7]. The PLCA model considers that $x(t,m)$ is propotional to the probability distribution that generates $x(t,m)$ as follows:

$$x(t,m) \propto P(t,m) = P(t) \sum_{k} P(k|t) P(m|k) \tag{2}$$

PLCA estimates the unknown parameters $P(k|t)$ and $P(m|k)$ from the input signal $x(t,m)$. $P(k|t)$ corresponds to $c(k|t)$ in (1), and we call $P(k|t)$ the probabilistic activation. $P(m|k)$ corresponds to $s(m|k)$ in (1), and we call $P(m|k)$ the probabilistic basis component. Also, in order to solve the under-determined problem that the number of compounds in the air K is unknown, the conventional method makes use of sparsity in the activations, sparsity in the basis components, and sparsity among the basis components. Thus, the method maximizes the following objective function with entropic priors:

$$\begin{aligned} & J\left(\{P(k|t)\}, \{P(m|k)\}\right) \\ = & \sum_{t}\sum_{m} x(t,m) \log \sum_{k} P(k|t) P(m|k) \\ & -\beta_{\mathrm{a}} \sum_{t} H(\{P(k|t)\}_{k}) - \beta_{\mathrm{b}} \sum_{k} H(\{P(m|k)\}_{m}) \\ & -\beta_{\mathrm{c}} \sum_{k,k'|k\neq k'} H(\{P(m|k)\}_{m}, \{P(m|k')\}_{m}), \end{aligned} \tag{3}$$

where β_{a} is the parameter of the sparsity of $P_t(k)$, β_{b} is the parameter of the sparsity of $P(m|k)$, β_{c} is the parameter of the sparsity between bases, $H(\{P_i\}_i)$ is the α-th order Renyi's entropy defined as $H(\{P_i\}_i) = \frac{1}{1-\alpha} \log \sum_i {P_i}^{\alpha}$, and $H(\{P_i\}_i, \{Q_i\}_i)$ is the cross entropy defined as $H(\{P_i\}_i, \{Q_i\}_i) = -\sum_i P_i \log Q_i - \sum_i Q_i \log P_i$.

By maximizing $J\left(P(k|t), P(m|k)\right)$, we can obtain **PLCA** algorithm (Algorithm 1) to estimate $P(k|t)$ and $P(m|k)$. After the algorithm converges, finally, we can calculate the estimate $\hat{c}(k|t)$ of $c_(k|t)$ from (5), and also the estimate $\hat{s}(m|k)$ of $s(m|k)$ from (7). The conventional method achieve the correct solution in many cases. However, the speed of convergence is slow, and the total calculation time is long.

4. PROPOSED METHOD

We assume that the reason why the speed of convergence is slow is that the initial solution is not adequate. As our conventional work [7], in mass spectra separation domain, the correct solution is likely to a sparse solution in terms of both time direction and *m/z* direction. However, the solution initialized by random values tends to be far from a sparse solution. Thus, we need to think of a method of initialization by a solution near the correct solution.

We focus on ICA for initialization. Similarly to PLCA, ICA is an blind source separation method, so that ICA is available for initialization. ICA does not impose non-negativity to the solution. However, ICA imposes independence that is assumed also in PLCA, and so the solution of ICA is near that of PLCA. Also, fast algorithms of ICA are commonly known, for example, Fast ICA [9] and the Natural Gradient algorithm [10]. So that, by comparison with the calculation time of PLCA, that of ICA is extremely short. Thus, by initializing the unknown parameters by ICA and reducing the number of iterations of PLCA, we aim to shorten the total calculation time. By converting **PLCA**, we can achieve **ICA-PLCA** algorithm (Algorithm 2). Similarly to the conventional method [7], in order to concentrate the stationary chemical background on the first basis, i.e. k = 1, we set $P(m|k=1)$ to the uniform distribution in (13), and set $P(k=1|t)$ to the higher value than $P(k \neq 1|t)$ in (14).

The calculation complexity of PLCA is $O(LTKM)$, where L is the number of iterations. In contrast the calculation complexity of the above initialization process is $O(LTK^2)$. So that the initialization process is faster than PLCA in the case of $K < M$. The proposed method make use of this feature, and it can reduce the total calclation time by increasing the number of iterations of the initialization process and decreasing that of PLCA.

5. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed method. We used the device of the walk-through portal explosives detector [1] to record the input mass spectra. Some of the authors had developed a prototype device as supported by Ministry of Education, Culture, Sports, Science and Technology, Japan for three years since 2007. Based on this prototype device, the device of this experiment was developed. Figure 1 shows a model of the device. We recorded the mass spectra in a real station to measure the chemical background of real envi-

Algorithm 1 PLCA [7]

1. Initialization process
Set all the unknown parameters to random values.
2. Iteration process
Iterate the following E step and M step.
E step:

$$P(k|t,m) = \frac{P(k|t)P(m|k)}{\sum_{k'} P(k'|t)P(m|k')}, \tag{4}$$

M step:

$$\hat{c}(k|t) = \sum_{m} x(t,m) P(k|t,m), \tag{5}$$

$$P(k|t) = \begin{cases} \frac{1}{1+\sum_{k'\neq 1} g(\beta_{\mathrm{a}}, \{\hat{c}(k'|t)\}_{k})} & \text{if } k=1, \\ \frac{g(\beta_{\mathrm{a}}, \hat{c}(k|t))}{1+\sum_{k'\neq 1} g(\beta_{\mathrm{a}}, \{\hat{c}(k'|t)\}_{k})} & \text{otherwise}, \end{cases} \tag{6}$$

$$\hat{s}(m|k) = \sum_{t} x(t,m) P(k|t,m) - \beta_{\mathrm{c}} \sum_{k'\neq k} P(m|k'), \tag{7}$$

$$P(m|k) = g(\beta_{\mathrm{b}}, \{\hat{s}(m|k)\}_{m,\tau}), \tag{8}$$

where $g(\beta, \{\gamma_i\}_i)$ is the entropic prior of Grindlay and Ellis [8]: $g(\beta, \{\gamma_i\}_i) = \frac{{\gamma_i}^{\beta}}{\sum_i {\gamma_i}^{\beta}}$.

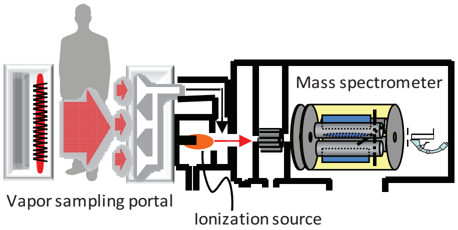
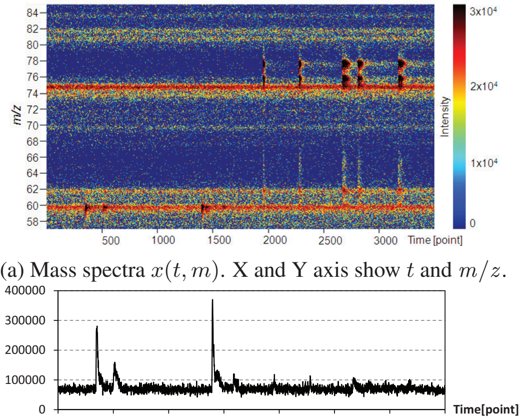


Fig. 1. Explosives detector.

Algorithm 2 ICA-PLCA

```
In PLCA, replace the initialization process with the following equations:
1. By the whitening matrix $\boldsymbol{W}$, prewhiten $\boldsymbol{x}(t) = [x(t,1),\cdots,x(t,M)]^T$ and reduce the number of dimensions:
    $$\boldsymbol{z}(t) = [z(t,1),\cdots,z(t,K)]^T = \boldsymbol{W}\boldsymbol{x}(t). \quad (9)$$
2. Compute the separated signals:
    $$\boldsymbol{y}(t) = [y(t,1),\cdots,y(t,K)]^T = \boldsymbol{V}\boldsymbol{z}(t). \quad (10)$$
3. Compute the natural gradient;
    $$\boldsymbol{V} = \boldsymbol{V} + \eta\left[\boldsymbol{I} - \frac{1}{T}\tanh(\boldsymbol{y}(t))\boldsymbol{y}(t)\right]\boldsymbol{V}. \quad (11)$$
4. Return to 2. until convergence.
5. Convert $\boldsymbol{V}$ into a basis matrix $\boldsymbol{S}$ on the m/z space:
    $$\boldsymbol{S} = \boldsymbol{V}\boldsymbol{W}. \quad (12)$$
6. By normalizing $\boldsymbol{S}$, initialize $P(m|k)$:
    $$P(m|k) = \begin{cases} \frac{1}{M} & \text{if } k=1, \\ \frac{|S_{m,k-1}|}{\sum_m |S_{m,k-1}|} & \text{otherwise.} \end{cases} \quad (13)$$
7. By normalizing $\boldsymbol{y}(t)$, initialize $P(k|t)$:
    $$P(k|t) = \begin{cases} \frac{1}{1+\sum_k' {y'}_k(t)} & \text{if } k=1, \\ \frac{{y'}_{k-1}(t)}{1+\sum_k' {y'}_k(t)} & \text{otherwise.} \end{cases} \quad (14)$$
   where $y'_k(t) = \frac{|y_k(t)|}{\sum_k |y_k(t)|}$.
8. Run the iteration process of PLCA.
```



(a) Mass spectra $x(t,m)$. X and Y axis show t and m/z.

(b) Chromatogram (time profile) of around *m/z* 59. X and Y axis show t and the intensity $I(t) = \sum_{m\in[m/z\ 58,\ m/z\ 60]} x(t,m)$.

Fig. 2. Input signal.

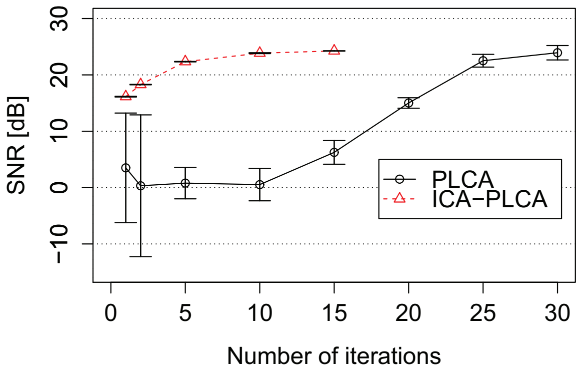


Fig. 3. SNR for each method. X and Y show the number of iterations and SNR [dB]. Error bars represent 95% confidence intervals.

ronments. We used 3500 mass spectra of about five minutes from the whole recorded data; i.e., $T = 3500$, and the number of the *m/z* index M was 512. Figure 2 (a) shows the input mass spectra, and Fig. 2 (b) is the chromatogram (time profile) of around *m/z* 59. The chemical background components have stationary peaks at *m/z* 59, *m/z* 62 and *m/z* 75 (Fig. 2 (a)). In this experiment, an experimenter passed through the device with Compound 1 (*m/z* 59), i.e. k=2, four times in the former half of the time, and with Compound 2 (*m/z* 59, *m/z* 62, *m/z* 76 and *m/z* 77), i.e. k=3, five times in the latter half of the time. As Fig. 2 (b) shows, the fourth peak of Compound 1 ($t = 1600$) was small and it had the same level as those of when Compound 2 was passed (e.g. $t = 1950$).

We applied **PLCA** and **ICA-PLCA** described in Section 4. In **PLCA**, all the unknown parameters were initialized by random values. On each condition, the estimation process was run 20 times. We set the number of bases K in the estimation process at eight. $\beta_{\rm a}$ was 1.02, $\beta_{\rm b}$ was 1.02, $\beta_{\rm c}$ was 0.4. The estimation process was run in C# on a PC with an Intel Core i7 3.3GHz CPU and 12GB of RAM. The measurements were $\mathrm{SNR}_{k,i,j}$ as follows:

$$\mathrm{SNR}_{k,i,j} = 10\log_{10}\frac{\max_{t\in\mathcal{A}_{k,i}}|\hat{c}(k|t)_j|}{\sqrt{\frac{1}{|\mathcal{N}_k|}\sum_{t\in\mathcal{N}_k}|\hat{c}(k|t)_j|^2}}\ [\mathrm{dB}]\,, \quad (15)$$

where $\mathcal{A}_{k,i}$ was the area around the i-th time when the k-th compound is passed through the device, $\mathcal{N}_k$ is the non-active time area; i.e., $\mathcal{N}_{k=2}$ was $[2000, 3500]$, and $\mathcal{N}_{k=3}$ was $[0, 1500]$, and j is the index of executions. Next, we defined SNR as an ensemble mean over k i, and j. In the case of the arithmetic mean, a peak $\mathrm{SNR}_{k,i,j}$ of which will be extremely high tends to cause SNR to be higher excessively. In order to make much account of worse $\mathrm{SNR}_{k,i,j}$, we defined SNR as the harmonic mean of $\mathrm{SNR}_{k,i,j}$ over k i, and j:

$$\mathrm{SNR} = \left\{\sum_{k=2,3}\sum_{i,j}\frac{1}{\mathrm{SNR}_{k,i,j}}\right\}^{-1} \quad (16)$$

As Fig. 3 shows, the larger the number of iterations was, mostly the higher the performance was. SNR of **ICA-PLCA** converged to about 22 dB at about 10 iterations. However, in the cases that the range of the number of iterations was 1 to 10, SNR of **PLCA** was about 0 dB. The convergence of **PLCA** was much slower than that of **ICA-PLCA**, and SNR of **ICA-PLCA** converged at about 30 iterations. These results indicates that the performance of **ICA-PLCA** with 10 iterations is comparable to that of **PLCA** with 30 iterations. In contrast, as Fig. 4 shows, the calculation time of **PLCA** with 30

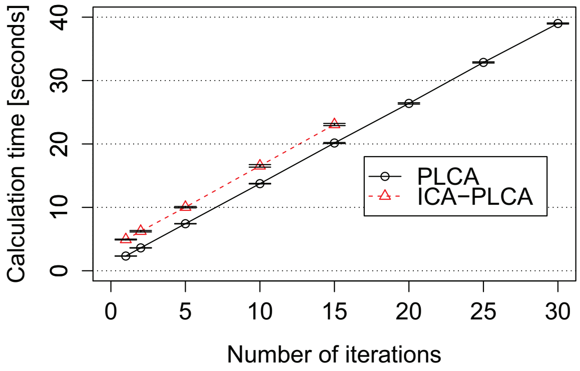


Fig. 4. The calculation time for each method. X and Y show the number of iterations and the calculation time [second]. Error bars represent 95% confidence intervals.

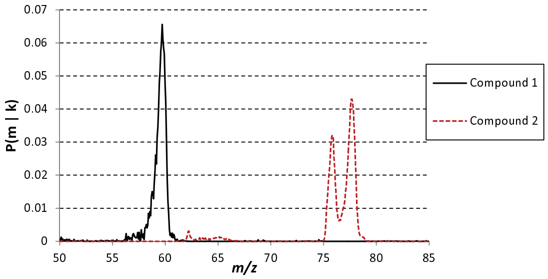


Fig. 5. Estimates of the probabilistic spectral basis components $P(m|k)$. X and Y show *m/z* and $P(m|k)$.

iterations was much longer than that of **ICA-PLCA** with 10 iterations. This indicates that the total calculation time can be reduced to about 1/3 without loss of performance by reducing the number of iterations of PLCA. Thus, the proposed method can reduce the total calculation time by using ICA.

6. RELATION TO PRIOR WORK

The work presented here has focused on PLCA for mass spectra separation. As mentioned above, there have been recently several researches that apply NMF [5, 6] and PLCA [7]. In particular, PLCA has a feature that it is easy to use statistical knowledge as sparsity priors. However, the convergence of the method is slow, and the total calculation time is long. Thus, the method can not run in real time, and it is difficult to apply the method to the explosives detection system in practice. As far as we know, in the domain of mass spectrometry, there are no researches on acceleration of PLCA because, so far, it has not been necessary that signal separation is executed in real-time in the domain of mass spectrometry. There are several approaches on improving the initialization process of NMF [11, 12, 13], but these approaches have not been applied to mass spectrometry, and it is not obvious whether these approaches can be applied to PLCA. The proposed method makes use of ICA in the initialization process of PLCA, and reduces the total calculation time. This point was not considered in these earlier studies.

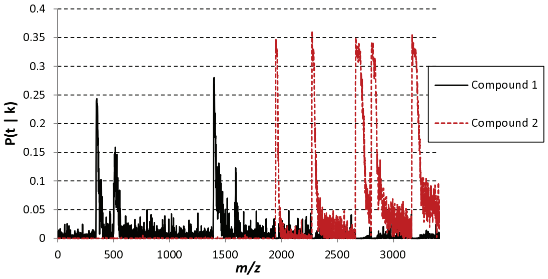


Fig. 6. Estimates of the probabilistic activations $P(k|t)$. X and Y show t and $P(k|t)$.

7. CONCLUSION

We proposed a new method to separate mass spectra into components of each chemical compound for explosives detection. In order to speed up the conventional method based on PLCA, the proposed method makes use of independent component analysis (ICA) in the initialization process. In the experiment using the data in a real environment, it was shown that the proposed method can reduce the total calculation time.

8. REFERENCES

[1] Y. Takada, H. Nagano, Y. Suzuki, M. Sugiyama, E. Nakajima, Y. Hashimoto, and M. Sakairi, "High-throughput walkthrough detection portal for counter terrorism: detection of triacetone triperoxide (tatp) vapor by atmospheric-pressure chemical ionization ion trap mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 25, pp. 2448–2452, 2011.

[2] Y.R. Lau, L. Weng, K. Ng, and C. Chan, "Time-of-flight-secondary ion mass spectrometry and principal component analysis: determination of structures of lamellar surfaces," *Analytical Chemistry*, vol. 82, pp. 2661–2667, 2010.

[3] M. Heikkinen, A. Sarpola, H. Hellman, J. Ramo, and Y. Hiltunen, "Independent component analysis to mass spectra of aluminium sulphate," *World Academy of Science, Engineering and Technology*, vol. 26, pp. 173–177, 2007.

[4] D. Mantini, F. Petrucci, P.D. Boccio, D. Pieragostino, M.D. Nicola, A. Lugaresi, G. Federici, P. Sacchetta, C.D. Ilio, and A. Urbani, "Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra," *Bioinfomatics*, vol. 1, pp. 63–70, 2008.

[5] P.W. Siy, R.A. Moffitt, R.M. Parry, Y. Chen, Y. Liu, M.C. Sullards, A.H. Merrill, and M.D. Wang, "Matrix factorization techniques for analysis of imaging mass spectrometry data," in *BIBE 2008*, 2008.

[6] R. Dubroca, C. Junot, and A. Souloumiac, "Weighted nmf for high-resolution mass spectrometry analysis," in *EUSIPCO 2012*, 2012.

[7] Y. Kawaguchi, M. Togami, H. Nagano, Y. Hashimoto, M. Sugiyama, and Y. Takada, "Mass spectra separation for explosives detection by using probabilistic latent component analysis," in *ICASSP 2012*, 2012.

[8] G. Grindlay and D.P.W. Ellis, "A probabilistic subspace model for multi-instrument polyphonic transcription," in *ISMIR 2010*, 2010.

[9] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[10] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *IEEE International workshop on wireless communication*, 1997.

[11] Y. Kim and S. Choi, "A method of initialization for nonnegative matrix factorization," in *ICASSP 2007*, 2007.

[12] C. Boutsidisa and E. Gallopoulosb, "Svd based initialization: Ahead start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[13] M. Rezaei, R. Boostani, and M. Rezaei, "An efficient initialization method for nonnegative matrix factorization," *Journal of Applied Sciences*, vol. 11, no. 2, pp. 354–359, 1999.