MASS SPECTRA SEPARATION FOR EXPLOSIVES DETECTION BY USING PROBABILISTIC LATENT COMPONENT ANALYSIS

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

Central Research Laboratory, Hitachi, Ltd.
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. In mass spectra, all components have no negative values. However, conventional factor analyses for basis decomposition use no constraints of non-negativity, and we can not apply these methods to mass spectra. The proposed method is based on probabilistic latent component analysis (PLCA). The constraints of non-negativity always hold in PLCA, so that the method is effective for mass spectra. In addition, PLCA is defined in a statistical framework, thus PLCA makes it possible to utilize additional a priori information. Therefore, we introduce sparseness assumptions in the domain of mass spectrometry to PLCA in order to estimate the components more accurately. Experimental results indicate that the proposed method outperforms existing methods.

Index Terms— Mass spectrum analysis, Blind source separation, Probabilistic latent component analysis, Sparseness assumption, Non-negativity

1. INTRODUCTION

The threat of improvised explosive devices has become a serious problem for all countries because the procedures and recipes for making these bombs 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 the explosives detection system, explosives compounds, other chemical compounds and the chemical background are mixed with each other in the mass spectra. It is necessary to separate the spectral components of each compound from the mass spectra. There are many researches that employ conventional factor analyses for basis decomposition such as Principal Component Analysis (PCA) [2] and Independent Component Analysis (ICA) [3, 4]. The components estimated by PCA may contain negative values because PCA uses no constraints of non-negativity. However, the components of the real mass spectra are non-negative, so that PCA is not effective for mass spectra. To make matters worse, PCA imposes the orthogonality constraint. In the explosives detection, some components of different compounds are mixed into the same *m/z* of mass spectra; i.e., the orthogonality does not hold in general, and so the performance of PCA is more degraded. ICA does not make use of constraints of the orthogonality, but it uses no constraints of non-negativity. Thus, ICA also suffers from performance degradation in mass spectra. Recently, new approaches called Non-negative Matrix Factorization (NMF) and Non-negative Independent Component Analysis (NICA) have been applied to the area of mass spectral imaging [5]. These approaches have the desirable property that the estimated components are guaranteed to be non-negative, and the approaches have the advantage of no distortion caused by the negative values. However, it is difficult for these methods to utilize other a priori information of the domain of mass spectrometry because the methods are not defined in a probabilistic framework.

In this paper, we propose a new mass spectra separation method for explosives detection. The proposed method uses probabilistic latent component analysis (PLCA) [6, 7, 8]. By using PLCA, the proposed method can separate the non-negative components without assumptions about the orthogonality. Moreover, PLCA is NMF redefined in a probabilistic framework, thus PLCA makes it possible to utilize additional a priori information. Therefore, in order to estimate the components accurately, we introduce sparseness assumptions of mass spectra to PLCA. Experimental results indicate that the proposed method outperforms existing methods.

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,

$$x(t, m) = \sum_{k=1}^{K} c_t(k) s_k(m), \tag{1}$$

where k is the index of a compound basis, K is the number of the kinds of the compounds in the air, $c_t(k)$ is the intensity of the k-th compound in the time index t, and $s_k(m)$ is the time-invariant spectral basis component for the k-th compound. Here, we normalize $s_k(m)$ such that $\sum_{m=1}^{M} |s_k(m)|^2 = 1$.

In this paper, we estimate the unknown variables $c_t(k)$ and $s_k(m)$ from the known variables $x(t, m)$. This problem equals to the blind source separation problem. We must add the following three constraints of the explosives detection system to this problem. The first is that $s_k(m)$ is non-negative for all compound and *m/z* because mass spectra represent the number of ions for each *m/z*. The

second is that we can not assume the orthogonality between different basis component $s_k(m)$ because different components are mixed into the same *m/z* in real environments. The third is that 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. MASS SPECTRA SEPARATION BY USING PLCA

3.1. PLCA model

In this section, we introduce PLCA model. PLCA model consists of a pair of probabilistic trials:

1. In each time t, the k-th compound is selected in a probability $P_t(k)$.
2. For the selected compound k, the m-th index of *m/z* is selected in a probability distribution $P(m|k)$, and a small positive constant value Δ is voted to the spectral bin (t, m) corresponded to $x(t, m)$.

We assume that $x(t, m)$ is generated by voting in these probabilistic trials $S(t)$ times for each time index t, where $S(t) = \sum_{m=1}^{M} x(t, m)$. Therefore, we can formulate the probability distribution that $x(t, m)$ is generated as follows:

$$P(x(t,m)\forall t,m) = \prod_{t=1}^{T}\prod_{m=1}^{M}\left\{\sum_{k=1}^{K} P_t(k)P(m|k)\right\}^{x(t,m)} \tag{2}$$

leading to the log-likelihood

$$\log P(x(t,m)\forall t,m) = \sum_{t=1}^{T}\sum_{m=1}^{M} x(t,m)\log\sum_{k=1}^{K} P_t(k)P(m|k). \tag{3}$$

In PLCA model, we can interpret (1) as a event of trials in the probability distribution of (3). $P_t(k)$ corresponds to $c_t(k)$ in (1), and we call $P_t(k)$ the probabilistic activities. Also, $P(m|k)$ corresponds to $s_k(m)$ in (1), and we call $P(m|k)$ the probabilistic spectral basis components. Therefore, at first, the estimation process calculates $P_t(k)$ and $P(m|k)$ that maximize (3), next it calculates $c_t(k)$ and $s_k(m)$ from these probabilistic distributions easily. However, we can not estimate both of $P_t(k)$ and $P(m|k)$ directly because $P_t(k)$ is needed to estimate $P(m|k)$. We can employ the Expectation-Maximization (EM) algorithm for maximum likelihood estimation with missing data such as this problem, as we explain in Section 3.2. This PLCA-based method solves the problem of non-negativity because both $P_t(k)$ and $P(m|k)$ are probability variables, and these are always non-negative. Also, we do not any assumption about the orthogonality in (3), thus we can estimate the spectral basis components that are not orthogonal to each other.

3.2. Solution of PLCA by using sparseness assumptions

Returning back to the constraint that K is an unknown parameter, we must set K in the estimation process to a sufficiently large number. However, in this case, the estimation process can lead to an incorrect solution such that one compound is scattered into $P_t(k)$ and $P(m|k)$ for multiple bases, consequently, it degrades the accuracy of $P_t(k)$ and that of $P(m|k)$. In order to estimate the solution accurately, we introduce three sparseness assumptions of mass spectra as follows:

Sparsity of $P_t(k)$ Only a few compounds k are active in the same time.

Sparsity of $P(m|k)$ Each spectral basis component consists of only a few bins on the spectrum.

Sparsity between spectral basis components Spectral basis components $P(m|k)$ for different compounds k do not similar to each other.

In the area of audio and image signal processing, there are some approaches that use sparseness assumptions very similar to our assumptions of mass spectrometry. The approaches introduce the assumptions as "entropic priors" [6, 8] to PLCA. PLCA has the advantage that it enables to use a priori knowledge of domains like these methods. We also introduce the assumptions of mass spectra as entropic priors, and solve the estimation problem. We define the cost function by adding the term of "entropic priors" to (3) as follows:

$$\begin{aligned} & J\left(\{P_t(k)\},\{P(m|k)\}\right) \\ = & \sum_{t=1}^{T}\sum_{m=1}^{M} x(t,m)\log\sum_{k=1}^{K} P_t(k)P(m|k) \\ & -\beta_{\mathsf{a}}\sum_{t=1}^{T} H(\{P_t(k)\}_k) - \beta_{\mathsf{b}}\sum_{k=1}^{K} H(\{P(m|k)\}_m) \\ & -\beta_{\mathsf{c}}\sum_{k,k'|k\neq k'} H(\{P(m|k)\}_m, \{P(m|k')\}_m), \end{aligned} \tag{4}$$

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 $H(\{P_i\}_i, \{Q_i\}_i) = -\sum_i P_i \log Q_i - \sum_i Q_i \log P_i$. In (4), the second term corresponds to the sparsity of $P_t(k)$, third term corresponds to the sparsity of $P(m|k)$, and forth term corresponds to the sparsity between bases.

By maximizing the cost function $J\left(P_t(k), P(m|k)\right)$ in (4), we obtain the following EM algorithm to estimate $P_t(k)$ and $P(m|k)$:

E step:

$$P_t(k|m) = \frac{P_t(k)P(m|k)}{\sum_{k'=1}^{K} P_t(k')P(m|k')}, \tag{5}$$

M step:

$$\hat{c}_t(k) = \sum_{m=1}^{M} x(t,m)P_t(k|m), \tag{6}$$

$$P_t(k) = \begin{cases} \frac{1}{1+\sum_{k'\neq 1} g(\beta_{\mathsf{a}}, \hat{c}_t(k'))} & \text{if } k = 1, \\ \frac{g(\beta_{\mathsf{a}}, \hat{c}_t(k))}{1+\sum_{k'\neq 1} g(\beta_{\mathsf{a}}, \hat{c}_t(k'))} & \text{otherwise,} \end{cases} \tag{7}$$

$$r(m|k) = \sum_{t=1}^{T} x(t,m)P_t(k|m) - \beta_{\mathsf{c}}\sum_{k'\neq k} P(m|k'), \tag{8}$$

$$P(m|k) = g(\beta_{\mathsf{b}}, r(m|k)), \tag{9}$$

where $g(\beta, \gamma_i)$ is the α-order Renyi's entropic prior, which can be calculated by an iteration process as follows:

1. $h(i) = \beta\gamma_i + \frac{\alpha}{\alpha-1}\frac{g(\beta,\gamma_i)^{\alpha}}{\sum_{i'=1}^{K} g(\beta,\gamma_{i'})^{\alpha}}$
2. $g(\beta, \gamma_i) = \frac{h(i)}{\sum_{i'=1}^{K} h(i')}$
3. Return to 1 until convergence.

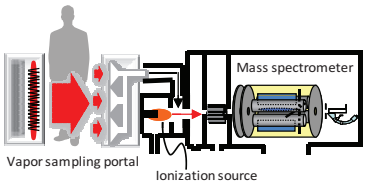
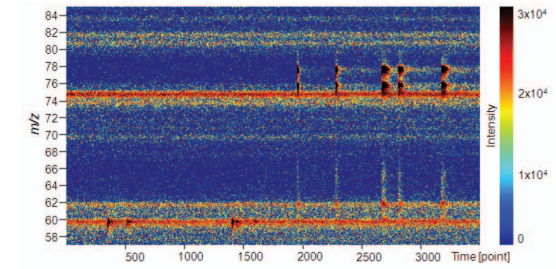
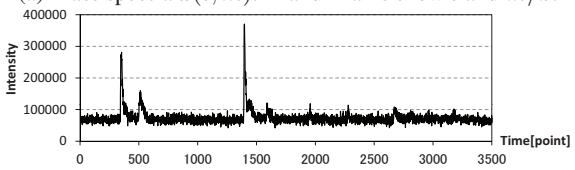


Fig. 1. Explosives detector.

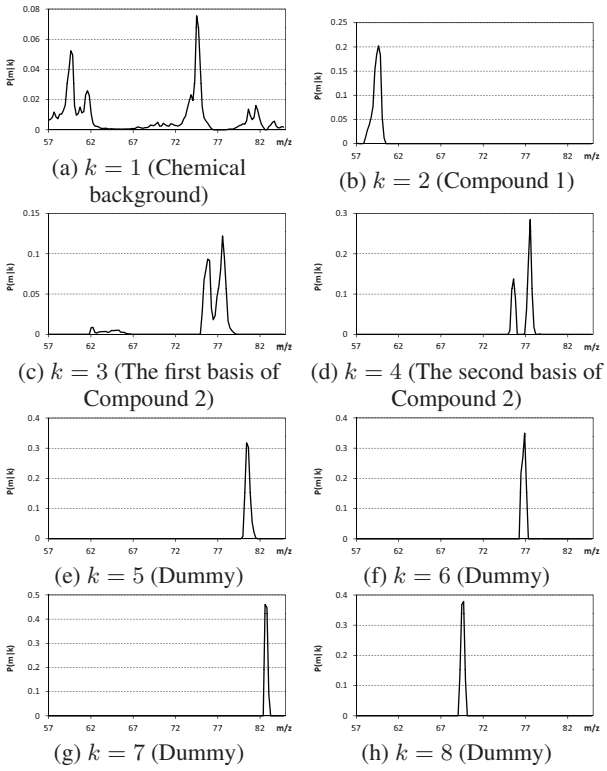


(a) Mass spectra $x(t, m)$. X and Y axis show t and m/z.



(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.



(a) $k = 1$ (Chemical background)
(b) $k = 2$ (Compound 1)
(c) $k = 3$ (The first basis of Compound 2)
(d) $k = 4$ (The second basis of Compound 2)
(e) $k = 5$ (Dummy)
(f) $k = 6$ (Dummy)
(g) $k = 7$ (Dummy)
(h) $k = 8$ (Dummy)

Fig. 3. Estimates of the probabilistic spectral basis components $P(m|k)$ for each k. X and Y show *m/z* and $P(m|k)$.

We set $P_t(k = 1)$ to the higher value than $P_t(k \neq 1)$ in (7) to handle the component of the stationary chemical background as one basis. Finally, after the algorithm converges, we can calculate the estimate $\hat{c}_t(k)$ of $c_t(k)$ from (6), and also the estimate $\hat{s}_k(m)$ of $s_k(m)$ can be calculated by $\hat{s}_k(m) = \frac{P(m|k)}{\sum_{m=1}^{M} |P(m|k)|^2}$.

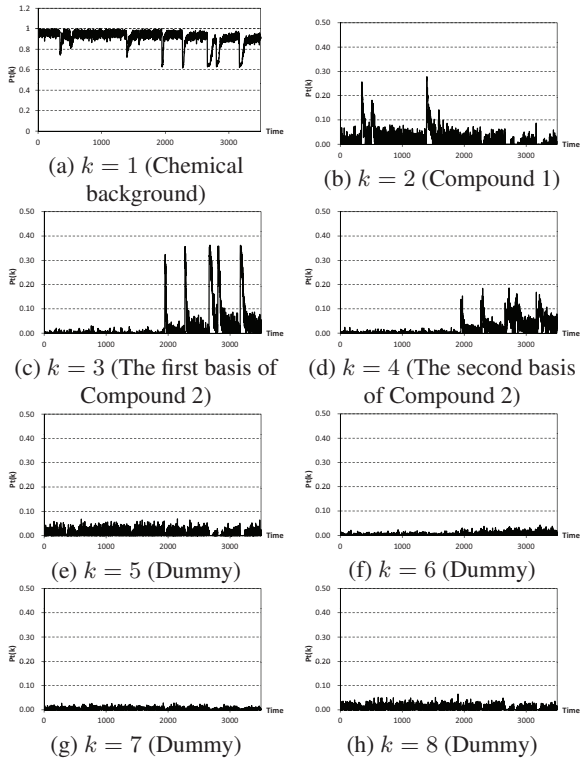
4. EXPERIMENTAL RESULTS

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 environments. 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 256. 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) 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) 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. $t = 1950$). From Fig. 2, $P(m|z)$ for each compound k are not orthogonal to each other obviously. We set the number of bases K in the estimation process at eight. β_{a} was 0.1, β_{b} was 0.9, β_{c} was 0.5, and the order of Renyi's entropy α was 1.2.

Figure 3 and 4 show the estimates of $P(m|k)$ and $P_t(k)$. As Fig. 3 shows, all correct main peaks were estimated for the chemical background, Compound 1 and Compound 2. Also, as Fig. 4 shows, the peaks exist at the correct times when Compound 1 and Compound 2 were passed. In particular, for Compound 1, the fourth peak of $P_t(k)$ was obviously higher than $P_t(k)$ of the latter half of the time (Fig. 4 (b)). These results indicate that the proposed method works well. However, Compound 2 was separated into two different bases $k = 3$ and 4. The sidelobes of the spectral basis component $P(m|k = 3)$ (Fig. 3 (c)) were broader than those of $P(m|k = 4)$ (Fig. 3 (d)). According to this feature, we can think that $k = 3$ and 4 correspond to the saturation state and the non-saturation one. In this experiment, saturation actually occurred in the mass spectrometer because the amount of Compound 2 was too large. The proposed method can not handle multiple states such as these two states as one compound because the method models only one mass spectrum for each time index, so that the method does not model the time series structures of mass spectra. This is a future work.

We compared the performance of the proposed method with existing methods PCA and ICA. The measurements were SNR as follows:

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



(a) $k = 1$ (Chemical background) (b) $k = 2$ (Compound 1) (c) $k = 3$ (The first basis of Compound 2) (d) $k = 4$ (The second basis of Compound 2) (e) $k = 5$ (Dummy) (f) $k = 6$ (Dummy) (g) $k = 7$ (Dummy) (h) $k = 8$ (Dummy)

Fig. 4. Estimates of the probabilistic activities $P_t(k)$ for each k. X and Y show t and $P_t(k)$.

where $\mathcal{A}_{k,i}$ was the area around the i-th time when k is passed through the device, and $\mathcal{N}_k$ is the non-active time area; i.e., $\mathcal{N}_{k=1}$ was $[2000, 3500]$, and $\mathcal{N}_{k=2}$ was $[0, 1500]$. In Table 1, the performance of the proposed method was higher than that of the other methods. In relation to the results, we show the spectral basis components for Compound 1 estimated by PCA and ICA in Fig. 5. We compare Fig. 5 with Fig. 3 (b). The existing methods estimate the spectral basis components without constraints of non-negativity, and so the estimates of the spectral basis components have both a negative peak and a positive peak. In contrast, the proposed method can estimate the spectral bases accurately by using the constraint of non-negativity. From these results, we think that the reasons for the high performance of the proposed methods are the constraint of non-negativity and no assumptions about the orthogonality.

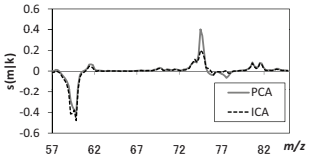


Fig. 5. Estimates of the spectral basis components $\hat{s}_k(m)$ for Compound 1 ($k = 2$) of PCA and ICA. X axis shows *m/z*. Y axis shows the estimated $s(m|k)$.

Table 1. $\mathrm{SNR}_{k,i}$[dB] for each method. "PLCA" means the proposed method.

Trial of pass-through		Method		
k	i	PLCA	PCA	ICA
$k = 2$	1	**19.4**	11.8	13.0
(Compound 1)	2	**15.0**	7.75	9.35
	3	**21.0**	13.0	14.4
	4	**13.2**	6.19	7.23
$k = 3$	1	**25.5**	8.05	16.0
(Compound 2)	2	**26.3**	8.35	16.5
	3	**27.2**	8.09	16.6
	4	**26.2**	7.97	16.3
	5	**26.9**	8.12	16.2

5. CONCLUSION

We proposed a new method to separate mass spectra into components of each chemical compound for explosives detection. The proposed method is based on PLCA. By using PLCA, the proposed method can separate the non-negative components without assumptions about the orthogonality. In addition, making use of the advantage that PLCA is defined in a probabilistic framework, we introduced sparseness assumptions in the domain of mass spectrometry to PLCA so as to estimate the solution more accurately. In the experiment using the data in a real environment, it was shown that the proposed method outperforms other conventional methods.

6. REFERENCES

[1] S. Kumano, M. Sugiyama, Y. Takada, H. Nagano, E. Nakajima, H. Hasegawa, Y. Hashimoto, and M. Sakairi, "Field test evaluation of a walkthrough portal detector of improvised explosive devices at a train station," in *ASMS 2010*, 2010.

[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] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *NIPS 2007*, 2007.

[7] B. Raj, R. Singh, M. Shashanka, and P. Smaragdis, "Bandwidth expansion with a Pólya urn model," in *ICASSP 2007*, 2007.

[8] P. Smaragdis, M. Shashanka, B. Raj, and G.J. Mysore, "Probabilistic factorization of non-negative data with entropic co-occurrence constraints," in *ICA 2009*, 2009.