MASS SPECTRA SEPARATION FOR EXPLOSIVES DETECTION BY USING PLCA WITH AN ATTENUATION MODEL

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 individual chemical compounds for explosives detection. The conventional method based on probabilistic latent component analysis (PLCA) is effective in many cases because the method can solve the problems of non-negativity and non-orthogonality by using sparsity of the domain of explosives detection. However, multiple compounds tend to be merged into a single basis component, and a single compound tends to be split into multiple basis components in error because the method does not model temporal structure of mass spectra. In this paper, first, we introduce a separation method based on shift-invariant PLCA (SIPLCA) in order to model temporal structure. Next, in order to prevent overfitting, we introduce an attenuation envelope such that it imposes a temporal constraint by focusing on the fact that the intensity of chemical compounds is attenuated with time after passing through the detector. Experimental results indicate that the proposed method outperforms the PLCA-based conventional method and other simple SIPLCA-based methods.

Index Terms— Mass spectrometry, Blind source separation, Probabilistic latent component analysis (PLCA), Time-shift invariance, 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 the intensity is called a mass spectrum. The detector observes the time series of the mass spectra continuously, and it detects a pattern of peaks of a explosive compound from the mass spectra data.

In real environments such as stations, explosive compounds, other chemical compounds, and the chemical background are mixed with each other in the mass spectra. It is necessary to separate the mass spectra into individual compounds. The system does not know either what kind of a spectrum the individual compounds have or when the individual compounds are observed in advance, and so the necessary 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, these methods are not fit to mass spectrometry domain, and separation performance degrades. Recently, there have been several researches that apply non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA) [5] to the area of mass spectrometry [6, 7, 8]. These approaches have the desirable feature that the estimated components are guaranteed to be non-negative, and so distortion is not caused by negative values. Furthermore, the PLCA-based conventional method [8] can solve the uncertainty problem of the number of compounds by using statistical knowledge as sparsity priors. The PLCA-based conventional method is effective in many cases. However, multiple compounds tend to be merged into a single basis component, and a single compound tends to be split into multiple basis components in error because the method does not model temporal structure of mass spectra.

In this paper, we propose a new mass spectra separation method for explosives detection. The proposed method has two key features: First, in order to model the temporal structure, the method makes use of shift-invariant PLCA (SIPLCA) [9]; second, the method imposes a temporal constraint by an attenuation model in order to prevent overfitting.

2. PROBLEM STATEMENT

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

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

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

In this paper, the goal is to estimate the unknown parameters $c(k|t)$ and $s(m|k)$ from the known parameters $x(t, m)$. This task is a blind source separation. In addition, we consider the following three conditions of the explosives detection system. First, all the elements of $c(k|t)$ and $s(m|k)$ are non-negative because mass spectra represent the number of ions for all time and all *m/z*; second, we can not assume the orthogonality between different basis components $s(m|k)$ because multiple compounds 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 PLCA-based conventional mass spectra separation method [8]. The PLCA model [5] considers that $x(t, m)$ is proportional to the following probability distribution:

$$x(t, m) \propto P(t, m) = P(t) \sum_{k=1}^{K} P(k|t)P(m|k) \tag{2}$$

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 $P(k|t)$ a probabilistic temporal activation. $P(m|k)$ corresponds to $s(m|k)$ in (1), and we call $P(m|k)$ a probabilistic spectral basis component. In order to solve the underdetermined problem that K is unknown, the conventional method makes use of sparsity in the temporal activation, sparsity in the spectral basis component, and sparsity between the spectral basis components. These sparsity constraints modeled by entropic priors make it possible to estimate $P(k|t)$ and $P(m|k)$ by EM (expectation-maximization) algorithm. The conventional method can obtain the correct solution in many cases. However, temporal structure of mass spectra is not modeled, so multiple compounds tend to be merged into a single basis component, and a single compound tends to be split into multiple basis components in error.

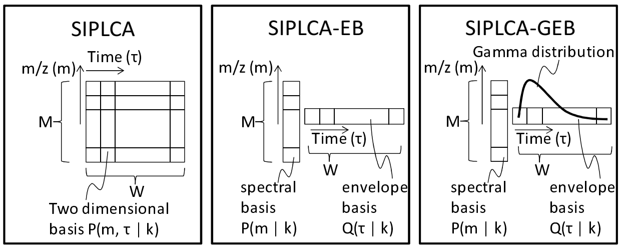


Fig. 1. Basis components of each version of the proposed method (**SIPLCA**, **SIPLCA-EB**, and **SIPLCA-GEB**).

4. PROPOSED METHOD

4.1. Mass spectra separation using shift-invariant PLCA

In this section, we introduce shift-invariant PLCA (SIPLCA) [9] into the PLCA-based conventional method in order to make it possible to model temporal structure of mass spectra. While we think of the spectral basis component of each compound as a one-dimensional probability distribution of *m/z* in PLCA, in SIPLCA, we think of a basis component of each compound as a two-dimensional probability distribution $P(m, \tau|k)$ as Fig. 1 shows. Now, we define $\tau = 1, \cdots, W$ as the time index in the basis component, where W is the frame size of the basis component. In SIPLCA, we assume that the input signal $x(t, m)$ is generated by convolving $P(m, \tau|k)$ over time as follows:

$$\begin{aligned} & x(t, m) \propto P(t, m) \\ = & \sum_{k} P(k) \sum_{\tau} P(m, \tau|k) P(t-\tau|k) \qquad (3) \\ = & \sum_{\tau} P(t-\tau) \sum_{k} P(m, \tau|k) P(k|t-\tau) \qquad (4) \end{aligned}$$

Smaragdis in [9] uses the formulation of (3), but (3) and (4) are equivalent. Equation (4) can be also used without loss of generality. In this paper, in order to keep consistency between the conventional method and the proposed method, we use the formulation of (4).

From (4), we can obtain **SIPLCA** algorithm (Algorithm 1) to estimate $P(k|t)$ and $P(m, \tau|k)$. Similarly to the conventional method [8], in order to concentrate a stationary chemical background on the first basis component, i.e., $k = 1$, we set $P(k|t)$ of $k = 1$ to a higher value than $P(k|t)$ for all $k \neq 1$ in (7). In the case of $W = 1$, **SIPLCA** is equivalent to the conventional method.

In the conventional method [8], the number of the unknown parameters, i.e., $P(k|t)$ and $P(m|k)$, is $(KT + MK)$. In contrast, in **SIPLCA**, the number of the unknown parameters, i.e., $P(k|t)$ and $P(m, \tau|k)$, is $(KT + MWK)$. Because **SIPLCA** has more unknown parameters than the conventional method, **SIPLCA** is likely to suffer from overfitting. In order to avoid overfitting, we introduce an attenuation model into **SIPLCA** in the next subsection.

Algorithm 1 SIPLCA

1. Initialization process

Set all the unknown parameters to random values.

2. Iteration process

Iterate the following E step and M step.

E step:

$$P(k,\tau|t,m) = \frac{P(t-\tau)P(k|t-\tau)P(m,\tau|k)}{\sum_{k',\tau'} P(t-\tau')P(k'|t-\tau')P(m,\tau'|k')}, \tag{5}$$

M step:

$$\hat{c}(k|t) = \sum_{m,\tau} x(t+\tau,m)P(k,\tau|t+\tau,m), \tag{6}$$

$$P(k|t) = \begin{cases} \frac{1}{1+\sum_{k'\neq 1} g(\beta_{\mathsf{a}},\{\hat{c}(k'|t)\}_k)} & \text{if } k=1, \\ \frac{g(\beta_{\mathsf{a}},\hat{c}(k|t))}{1+\sum_{k'\neq 1} g(\beta_{\mathsf{a}},\{\hat{c}(k'|t)\}_k)} & \text{otherwise,} \end{cases} \tag{7}$$

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

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

$$P(t) = \frac{\sum_{k,\tau,m} x(t+\tau,m)P(k,\tau|t+\tau,m)}{\sum_{t,k,\tau,m} x(t+\tau,m)P(k,\tau|t+\tau,m)}, \tag{10}$$

where $g(\beta,\{\gamma_i\}_i)$ is the entropic prior of Grindlay and Ellis [10]: $g(\beta,\{\gamma_i\}_i) = \frac{{\gamma_i}^\beta}{\sum_i {\gamma_i}^\beta}$.

4.2. Attenuation model

In this section, we introduce an attenuation model in order to avoid overfitting. Now, we assume that a person with a explosive compound passes through the detector. Within two or three seconds, ions of the compound are measured, and peaks rise rapidly in the mass spectrum. After the person goes away from the detector, the intensity of the peaks decreases continuously and slowly. The intensity of $P(m,\tau|k)$ changes largely depending on τ, but the spectral shape of $P(m,\tau|k)$ does not change largely depending on τ. Thus, in the application of explosives detection, we can assume that $P(m,\tau|k)$ can be decomposed into a spectral basis component $P(m|k)$ and a envelope basis component $Q(\tau|k)$, which are mutually independent. We represent the envelope basis component as $Q(\tau|k)$, not $P(\tau|k)$ in order to distinguish the envelope basis component not depending on t with the temporal activation $P(k|t)$ or $P(t|k)$ depending on t. By this assumption, we obtain the following equation:

$$P(m,\tau|k) = P(m|k)Q(\tau|k) \tag{11}$$

Therefore, we can also decompose the estimation process of $P(m,\tau|k)$ in **SIPLCA** into the estimation process of $P(m|k)$ and that of $Q(\tau|k)$. In the estimation process of $P(m|k)$, the sparsity in the spectral basis component can be used similarly to the conventional method [8]. However, it is not obvious that an sparsity constraint is effective in the estimation process of $Q(\tau|k)$. By considering these facts, we can obtain SIPLCA with an Envelope Basis (**SIPLCA-EB**) algorithm (Algorithm 2). As Fig. 1 shows, we can think that **SIPLCA-EB** has two kinds of the basis components, i.e., the spectral basis component $P(m|k)$ and the envelope basis component $Q(\tau|k)$.

Algorithm 2 SIPLCA-EB

In **PLCA**, replace (8) (9) with the following equations:

$$\hat{s}(m|k) = \sum_t x(t,m)P(k,\tau|t,m) - \beta_{\mathsf{c}} \sum_{k'\neq k} P(m|k'), \tag{12}$$

$$P(m|k) = g(\beta_{\mathsf{b}},\hat{s}(m|k)), \tag{13}$$

$$Q(\tau|k) = \frac{\sum_{t,m} x(t,m)P(k,\tau|t,m)}{\sum_{\tau',t,m} x(t,m)P(k,\tau'|t,m)}. \tag{14}$$

Compute (11).

Algorithm 3 SIPLCA-GEB

In **SIPLCA-EB**, replace (14) with the following equation:

$$Q(\tau|k) = \begin{cases} \frac{1}{W} & \text{if } k=1, \\ \mathcal{G}(\tau;\theta,\phi) & \text{otherwise.} \end{cases} \tag{15}$$

In addition, in order to enhance robustness, we impose a constraint the envelope basis component $Q(\tau|k)$. We focus on the fact that the intensity of chemical compounds is attenuated with time after passing through the detector. As explained above, $Q(\tau|k)$ rises rapidly first, and then it decreases continuously and slowly. In order to model such temporal structure of attenuation, we approximate $Q(\tau|k)$ by the gamma distribution $\mathcal{G}(\tau;\theta,\phi) = \frac{\phi^\theta}{\Gamma(\theta)}\tau^{\theta-1}e^{-\phi\tau}$, where $\Gamma(\theta)$ is the gamma function, and both θ and ϕ are the parameters of the gamma function. The attenuation curves of different compounds are similar to each other empirically, and so we assume that both θ and ϕ are constant, and they do not depend on compounds. Therefore, we can achieve SIPLCA with an gamma Envelope Basis (**SIPLCA-GEB**) algorithm (Algorithm 3). In (15), only for k = 1, $Q(\tau|k)$ is set to the uniform distribution in order to represent the fact that the intensity of the chemical background does not change in a short time. As Fig. 1 shows, we can think that **SIPLCA-GEB** has both the spectral basis component $P(m|k)$ and the envelope basis component $Q(\tau|k)$ similarly to **SIPLCA-EB**, where $Q(\tau|k)$ is defined by the gamma distribution. After the algorithm converges, finally, we can compute an estimate $\hat{c}(k|t)$ of $c(k|t)$ from (6), and we can also compute an estimate $\hat{s}(m|k)$ of $s(m|k)$ from (12).

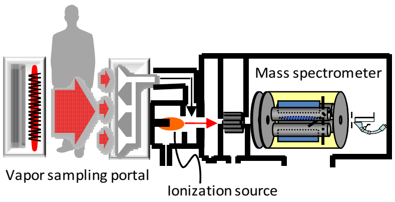
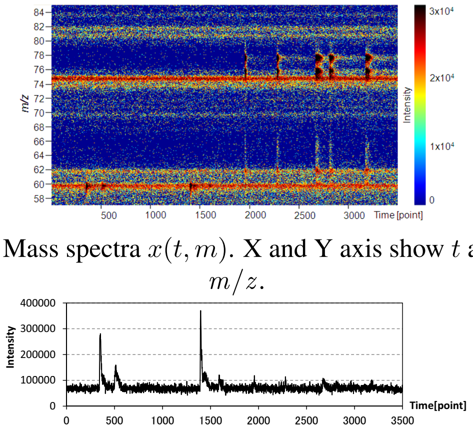


Fig. 2. Explosives detector.



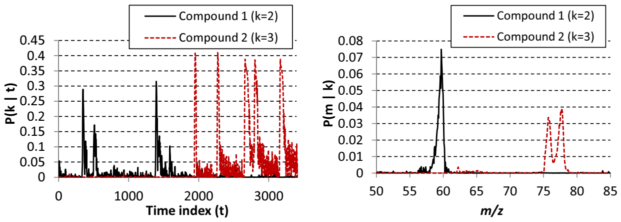
(a) Mass spectra $x(t, m)$. X and Y axis show t and m/z.

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

Fig. 3. Input signal.

5. EXPERIMENTAL RESULTS

We evaluated the separation performance of the proposed method. We used the device of the walk-through portal explosives detector [1] to record the input mass spectra. Four of the authors 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 2 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 indices of *m/z*, M was 512. Figure 3 (a) shows the input mass spectra, and Fig. 3 (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, i.e., $k = 1$ (Fig. 3 (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. 3 (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., $t = 1950$).



(a) Temporal activations $P(k|t)$.

(b) Spectral basis components $P(m|k)$.

Fig. 4. Estimates for Compound 1 ($k = 2$, black) and Compound 2 ($k = 3$, red).

We applied **SIPLCA**, **SIPLCA-EB**, and **SIPLCA-GEB** described in Section 4. In the case of $W = 1$, **SIPLCA** is equivalent to the PLCA-based conventional method [8]. In the estimation process, all the unknown parameters were initialized by random values. On each condition, the estimation process was run twenty times. We set the number of basis components K in the estimation process at eight. β_{a} was 1.01, β_{b} was 1.005, β_{c} was 0.5, θ was 2.5, and ϕ was 0.3. The measurements were $\mathrm{SNR}_{k,i,j}$ as follows:

$$\mathrm{SNR}_{k,i,j} = 10\log_{10}\frac{\max_{t\in\mathcal{A}_{k,i}}|\hat{c}(k|t)_j|}{\sqrt{\frac{1}{|\mathcal{N}_k|}\sum_{t\in\mathcal{N}_k}|\hat{c}(k|t)_j|^2}}\ [\mathrm{dB}], \quad (16)$$

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:

$$\mathrm{SNR} = \left\{\sum_{k=2,3}\sum_{i,j}\frac{1}{\mathrm{SNR}_{k,i,j}}\right\}^{-1} \quad (17)$$

Figure 4 shows estimates of the temporal activations $P(k|t)$ and estimates of the spectral basis components $P(m|k)$ respectively. As these results, the proposed method estimated both the temporal activations and the spectral basis components correctly. As Fig. 5 shows, the longer the frame size W is, mostly the higher the separation performance is. The separation performances of all the versions of the proposed method (**SIPLCA**, **SIPLCA-EB** and **SIPLCA-GEB**) of $W = 20$ were higher than that of the conventional method [8] (**SIPLCA** of $W = 1$). These results indicate that it is effective to model the temporal structure by SIPLCA in the proposed method. In the cases that the range of W is 1 to 5, the separation performances of each version are not significantly different. However, when W was set to be 20, SNR of **SIPLCA-GEB** is higher than those of the other versions at

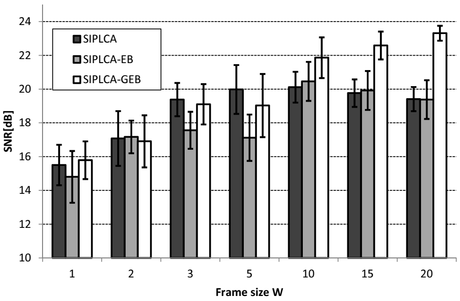
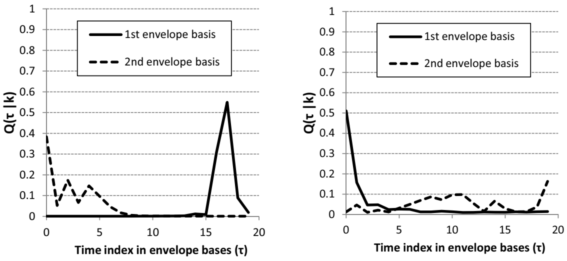


Fig. 5. SNR for each method. X and Y show the frame size of the basis component W and SNR [dB]. Error bars represent 95% confidence intervals.



(a) Compound 1.

(b) Compound 2.

Fig. 6. Splitted envelope basis components in the case of **SIPLCA-EB** and $W = 20$. X and Y show τ and $Q(\tau|k)$.

about 4dB. These results indicate that **SIPLCA** and **SIPLCA-EB** tend to suffer from overfitting, and that **SIPLCA-GEB** can prevent overfitting successfully by using the attenuation model. Actually, in the case of **SIPLCA-EB**, as Fig. 6 shows, both of the two compounds were split into two basis components respectively. The envelope basis components did not follow the attenuation model described in Section 4.2, i.e., some envelope basis components did not rise rapidly, and other envelope basis components were not continuous. It is clear that the envelope basis components were not estimated correctly. These results indicate that **SIPLCA-EB** suffered from overfitting.

6. CONCLUSION

We proposed a new method to separate mass spectra into individual chemical compounds for explosives detection. In order to model temporal structure, the proposed method makes use of SIPLCA. Moreover, in order to prevent overfitting, by focusing on temporal attenuation of chemical compounds, we introduced an attenuation envelope such that it imposes a temporal constraint into SIPLCA by focusing on the fact that the intensity of chemical compounds is attenuated with time after passing through the detector. In the experiment using the data in a real environment, it was shown that the proposed method (**SIPLCA-GEB**) outperforms the PLCA-based conventional method and other simple SIPLCA-based methods (**SIPLCA** and **SIPLCA-EB**).

7. REFERENCES

[1] Y. Takada, Y. Suzuki, H. Nagano, M. Sugiyama, E. Nakajima, M. Sugaya, Y. Hashimoto, and M. Sakairi, “High-throughput walkthrough detection portal as a measure for counter terrorism: Design of a vapor sampler for detecting triacetone triperoxide vapor by atmospheric-pressure chemical-ionization ion-trap mass spectrometry,” *IEEE Sensors Journal*, vol. 12, no. 6, pp. 1673–1680, jun 2012.

[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. Smaragdis, B. Raj, and M. Shashanka, “A probabilistic latent variable model for acoustic modeling,” in *NIPS*, 2006.

[6] 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.

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

[8] 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.

[9] P. Smaragdis, B. Raj, and M. Shashanka, “Shift-invariant probabilistic latent component analysis,” *MERL Technical Report TR2007-009*, 2007.

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