2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

IMPROVING THE ROBUSTNESS OF SURFACE ENHANCED RAMAN SPECTROSCOPY BASED SENSORS BY BAYESIAN NON-NEGATIVE MATRIX FACTORIZATION

Tommy S. Alstrøm[a], Kasper B. Frøhling[b], Jan Larsen[a], Mikkel N. Schmidt[a], Michael Bache[b], Michael S. Schmidt[b], Mogens H. Jakobsen[b], Anja Boisen[b]

[a]Dept. of Informatics and Mathematical Modeling, Technical University of Denmark
Asmussens Allé 321, 2800 Kgs. Lyngby, Denmark

[b]Dept. of Micro- and Nanotechnology, Technical University of Denmark
Ørsteds Plads 345 East, DK-2800, Kgs. Lyngby, Denmark

ABSTRACT

Due to applications in areas such as diagnostics and environmental safety, detection of molecules at very low concentrations has attracted recent attention. A powerful tool for this is Surface Enhanced Raman Spectroscopy (SERS) where substrates form localized areas of electromagnetic "hot spots" where the signal-to-noise (SNR) ratio is greatly amplified. However, at low concentrations hot spots with target molecules bound are rare. Furthermore, traditional detection relies on having uncontaminated sensor readings which is unrealistic in a real world detection setting. In this paper, we propose a Bayesian Non-negative Matrix Factorization (NMF) approach to identify locations of target molecules. The proposed method is able to successfully analyze the spectra and extract the target spectrum. A visualization of the loadings of the basis vector is created and the results show a clear SNR enhancement. Compared to traditional data processing, the NMF approach enables a more reproducible and sensitive sensor.

Index Terms— Biosensing, 17β-Estradiol, Non-negative Matrix Factorization (NMF), Surface Enhanced Raman Spectroscopy (SERS), Unsupervised Learning.

1. INTRODUCTION

Detection of biological and chemical species at very low concentrations have attracted a lot of attention from physicists, chemists and engineers in the past years due to its direct applications in e.g. diagnostics [1], prognostics [2] and environmental safety [3]. Surface Enhanced Raman Spectroscopy (SERS) [4, 5] is capable of single molecule detection [6–8], which makes it a powerful tool for biochemical analysis. SERS enables measurements of weak Raman signals through strong localized enhancement of electromagnetic fields observed in nano gaps (hot spots) between metal surfaces [9]. The SERS setup is illustrated on fig. 1.

One concern regarding SERS is the statistical behavior of the electromagnetic hot spot distribution [10]. Not only is it required to find a hot spot, but there also need to be molecules captured in a hot spot, something that is increasingly unlikely as the concentration of the target molecules decreases. To overcome this problem spectra from a larger area of a substrate is captured, a procedure often refereed to as *Raman Mapping* [11]. This increases the likelihood that spectra from hot spots containing molecules are recorded. Fig. 2 shows a typical Raman map and a typical Raman spectrum. In Raman spectroscopy the presence of peaks coupled with their location is used to determine what types of molecules are present of the surface, if any.

Traditional analysis of Raman maps consists of choosing one or two Raman shifts where the dominant peaks for the molecule of interest are present [11]. The intensities at these Raman shifts due to the presence of molecules are considered to be heavy tail distributed whereas Raman maps for blank substrates are considered to be normal distributed [7, 12].

Fig. 2 shows the traditional approach to data collected over an area of a Raman substrate. The area on fig. 2 is collected over a 30 µm × 30 µm area with a resolution of 1 µm. The Raman map visualizes the recorded intensities at a specific Raman shift. Areas with high intensities are recorded and it is these areas that are denoted as hot spots containing molecules.

Non-negative Matrix Factorization (NMF) [13, 14] is a popular unsupervised learning method used for discovery of meaningful patterns in data. It has been applied in a large range of areas, from image processing [14], decomposing signals from gas sensors [15] to reducing the noise in wind signals [16]. NMF has also proven as a successful tool for decomposing Raman spectra gathered using ordinary Raman

We acknowledge the support from the Danish Council for Strategic Research, under the grant “MUSE - Multisensor dvd-platform”

978-1-4799-3694-6/14/$31.00 ©2014 IEEE

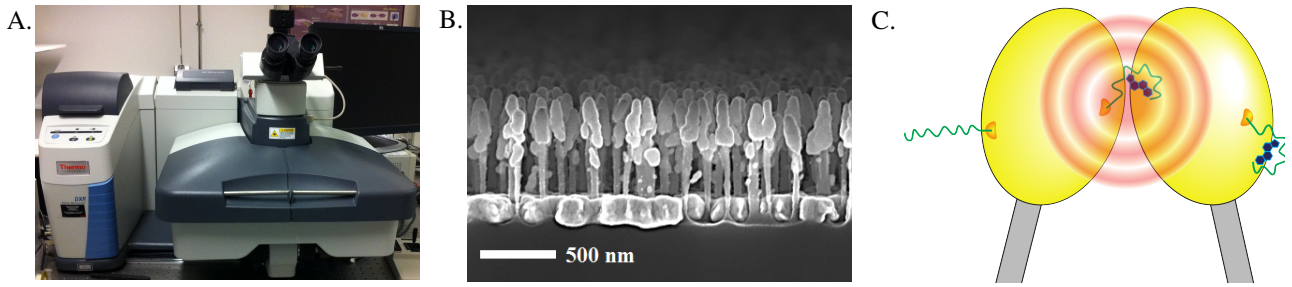

Fig. 1. A. The DXR™ Raman Microscope used to collect data. B. A side-view up of a Raman substrate depicting the nanopillars, courtesy of Kaiyu Wu, DTU. C. Illustration of the principle behind the SERS substrates. The two left pillars have molecules on them but in order to get the improved SNR the molecule needs to be captured in the hot spot as shown on the right. This is achieved by leaning the pillars through solvent evaporation.

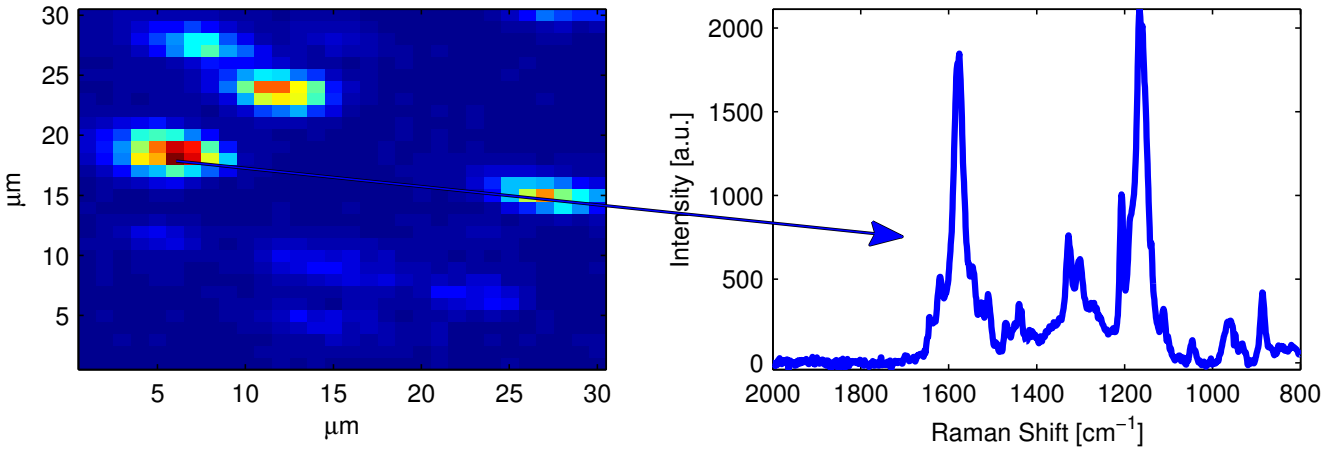

Fig. 2. Illustration of an uncontaminated SERS measurement using a Raman map at 1166 cm^{-1} (left) and an example spectrum for Estradiol Glow (right). Hot spots containing molecules are readily identified on the Raman map as the red areas. On the spectrum the peaks at 1166 cm^{-1} and 1580 cm^{-1} are considered the major discriminative.

spectroscopy [17, 18].

In this paper, we propose a variant of NMF that is particularly suitable for handing SERS spectra. For Raman maps a great deal of the data are empty spectra or spectra contaminated with other compounds. We show that NMF can demix data from Raman maps and identify spectra that can be considered true in a physical sense. These results can then be used to identify the presence of target molecules by setting a threshold on the loadings matrix in the NMF model.

2. MATERIALS AND METHODS

2.1. Surface Enhanced Raman Spectroscopy

SERS has been widely studied for its use in biosensors [11, 19–22]. By creating localized electromagnetic hot spots enhancement factors up to 10^{12} [9] have been demonstrated compared to traditional Raman spectroscopy. SERS has ultra high sensitivity combined with specific information of molecular vibrations, which yields a very powerful tool for biosensing. We use silver coated silicon nanopillars as a SERS substrate [23] that have previously shown the capability for biosensing [11].

2.1.1. Detection of 17β-Estradiol

Increasing health risks posed by endocrine-disrupting chemicals (EDCs) have been a growing concern for the public in later years. EDCs are compounds or molecules found in the environment, food and consumer products that affects hormone synthesis and control in humans and animals. Evidence have been presented that points out the severe impact EDCs have on reproduction capability (both male and female), metabolism, obesity and various types of cancer [3]. The female hormone 17β-Estradiol (E2) is an EDC and its presence in the environment is being watched closely by the European Commission [24].

In this work we have focused our attention towards a la-

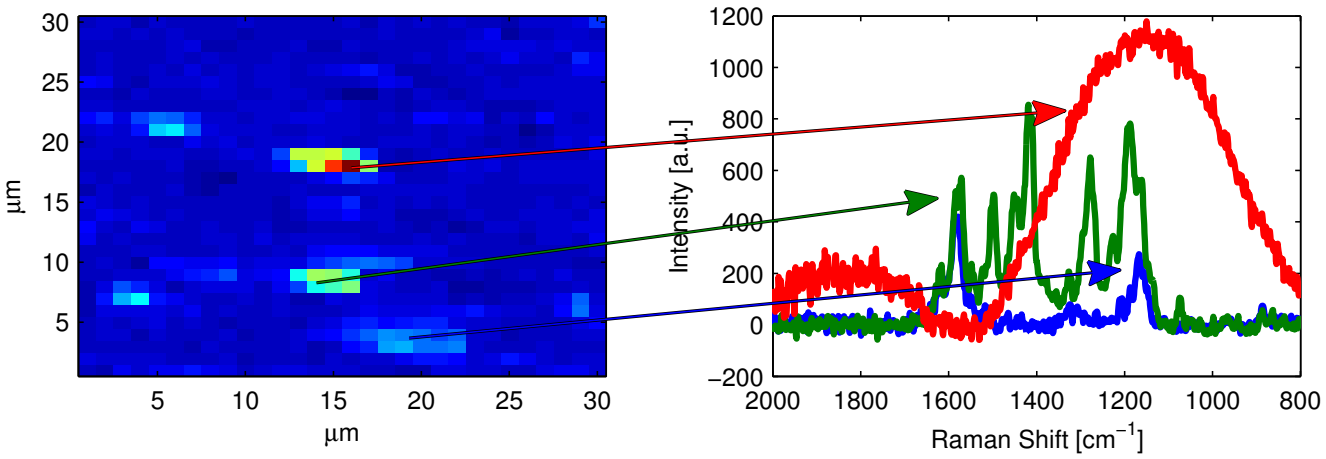

Fig. 3. An example of a Raman map that has been locally contaminated (Red and Green spectra). The map shows the relative intensities of the 1166 cm^{-1} Raman shift. The distribution of intensities are heavy-tailed but not only due to presence of EG but also due to other contaminants.

beled version of E2 due to the low Raman activity of pure E2. Jena Bioscience GmbH have developed a fluorescent labeled version called Estradiol Glow (EG) which has shown to yield strong and reproducible SERS spectra.

2.1.2. Data acquisition

Each SERS mapping was acquired with Thermo Scientific DXR Raman Microscope (fig. 1) using a 785 nm excitation laser. A laser powered at 0.5 mW power was used in conjunction with a 50x optical objective which yields an approximately 1 µm diameter laser spot size. The acquisition time was 1 s and each spot was sampled twice before a 6$^{\text{th}}$ order polynomial baseline correction was performed. Each map consisted of spectra gathered in larger squares (up to 10^4 points) with a 1x1 µm grid. High concentration solutions consists of 1 µM EG and low concentrations consists of 10 nM EG.

2.2. Bayesian Non-negative Matrix Factorization

The non-negative matrix factorization model can be stated as $\mathbf{X} = \mathbf{SL} + \mathbf{E}$, where $\mathbf{X} \in \mathbb{R}^{S\times N}$ is the data matrix that contain the entire Raman map. $\mathbf{X}$ is factorized into two matrices, the spectra matrix $\mathbf{S} \in \mathbb{R}_+^{D\times S}$ and the loadings matrix $\mathbf{L} \in \mathbb{R}_+^{S\times N}$, that contain only non-negative real elements. The residual matrix is denoted as $\mathbf{E} \in \mathbb{R}^{S\times N}$ and models the measurement noise.

A Raman spectrum consists of one or more spectral components (unless there are no molecules present, in which case there is only noise). We would like the matrix $\mathbf{S}$ to contain spectra that are easily interpretable, and each vector in $\mathbf{S}$ should contain as "distinct" information as possible. Further, a recorded spectrum is typically a combination of only a few basis spectra hence the rows in loading matrix $\mathbf{L}$ should favor sparse solutions. This calls for a model where both $\mathbf{S}$ and $\mathbf{L}$ have sparse priors. Following [25], we consider exponential priors on the elements in $\mathbf{S}$ and $\mathbf{L}$

$$p(\mathbf{S}) = \prod_{d=1}^{D}\prod_{s=1}^{S} \mathcal{E}\left(S_{d,s};\alpha\right) \tag{1}$$

and similarly for $\mathbf{L}$ we have

$$p(\mathbf{L}) = \prod_{s=1}^{S}\prod_{n=1}^{N} \mathcal{E}\left(L_{s,n};\beta\right) \tag{2}$$

where $\mathcal{E}(x;\lambda) = \lambda\exp(-\lambda x)u(x)$ is the exponential probability density function, and $u(x)$ is the unit step function. We assume that the product of $\mathbf{SL}$ is able to model the data such that $\mathbf{E}$ is only measurement noise. We model the residuals in $\mathbf{E}$ as i.i.d. normal distributed with zero mean and variance σ^2. Thus, the likelihood function can be written as

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{d=1}^{D}\prod_{n=1}^{N} \mathcal{N}\left(X_{d,n};\left(\mathbf{SL}\right)_{d,n},\sigma^2\right) \tag{3}$$

where $\boldsymbol{\theta} = \{\mathbf{S},\mathbf{L},\sigma^2\}$ denotes all parameters in the model and $\mathcal{N}(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp(-(x-\mu)^2/(2\sigma^2))$ is the normal probability density function. In order to apply the Bayesian framework, a prior distribution on the noise σ^2 is also required. Here, an inverse gamma density with shape k and scale θ is chosen

$$p(\sigma^2) = \mathcal{G}^{-1}\left(\sigma^2;k,\theta\right) \tag{4}$$

as it makes it more convenient to derive the posterior density for the parameters.

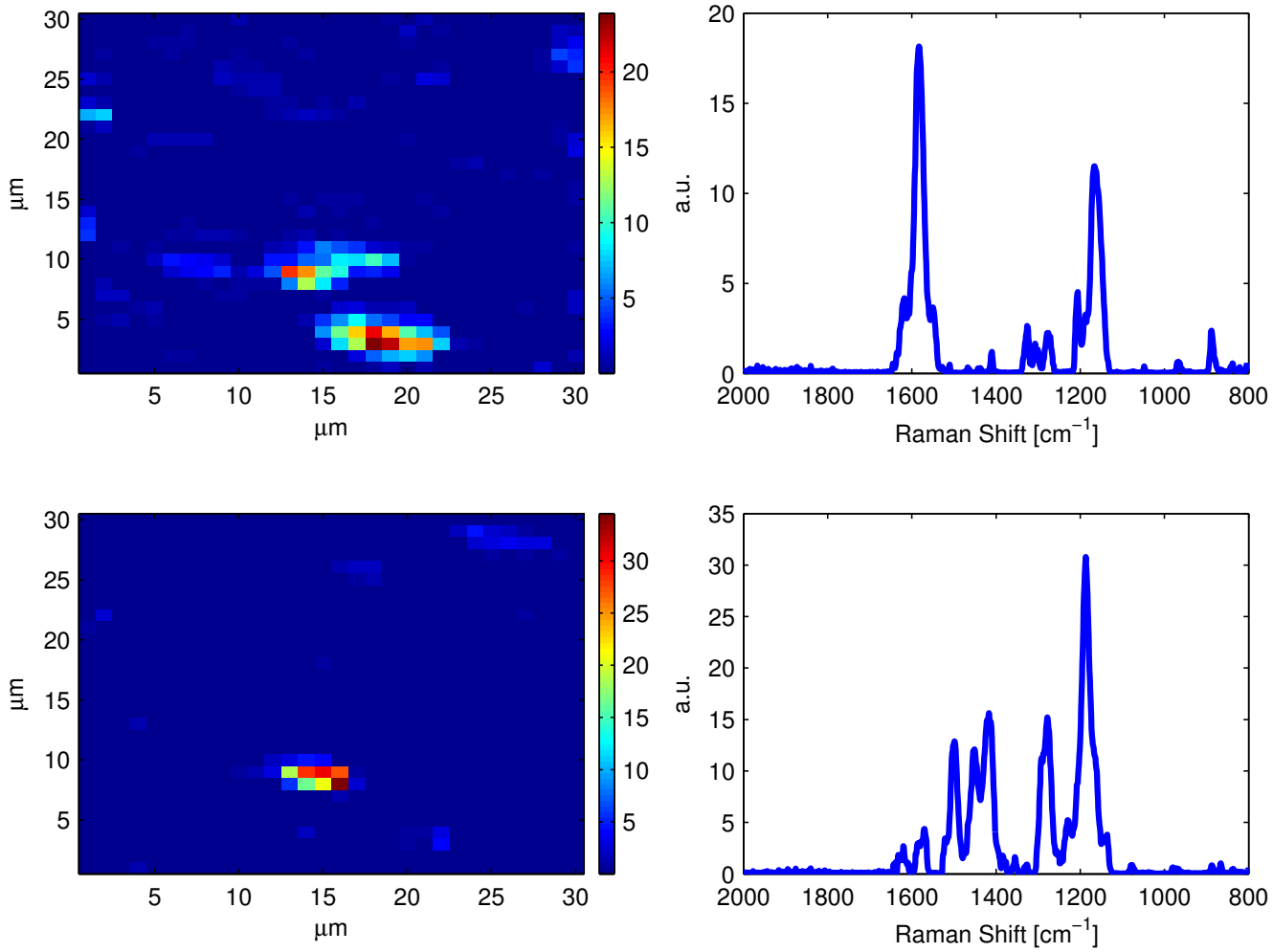

Fig. 4. Noise filtering using NMF. Left hand side is the loadings in $\mathbf{L}$ drawn as Raman maps and right hand side is the corresponding basis vector in $\mathbf{S}$

Using Bayes rule, the posterior for the parameters $\boldsymbol{\theta}$ can now be written as the product of eq. (1–4). This model has been proposed in earlier work [25], where an effective Gibbs sampler was derived which we use for inference.

In our experiments, we used a burn in period of 20 Gibbs sweeps and then used 50 sweeps to generate 50 samples from each component. The mean value from these samples were used as the parameter estimates. Further, α and β were chosen to 1 as we found this to yield a suitable sparsity. For the noise σ^2 we chose a flat improper prior, $k = 0$ and $\theta = 0$, to let the data dictate the inference regarding the noise.

2.2.1. Probing mode

For very low concentrations or even single molecule detection the Raman spectrum of interest will only make up a minor portion in the data. In order to cope with this scenario we have the spectra of interest as fixed vectors in $\mathbf{S}$ as initial values and then simply do not sample from these. This spectrum of interest can for example be learned from high concentration measurements.

3. RESULTS AND DISCUSSION

A typical "dirty" measurement is visualized on fig. 3. Utilizing the traditional data processing approach on this substrate would lead to misleading results as some of the very high intensities do not originate from EG. The large elevated area on the red spectra is caused by fluorescence (another vibrational phenomenon) that clearly interferes with Raman spectroscopy. It is probably seen because of insufficient washing or cleaning of the substrate after treatment. Different salts used in buffer solutions are auto-fluorescent, meaning that they have a naturally high degree of fluorescence. Another type of interference is seen in the green spectra, where clear

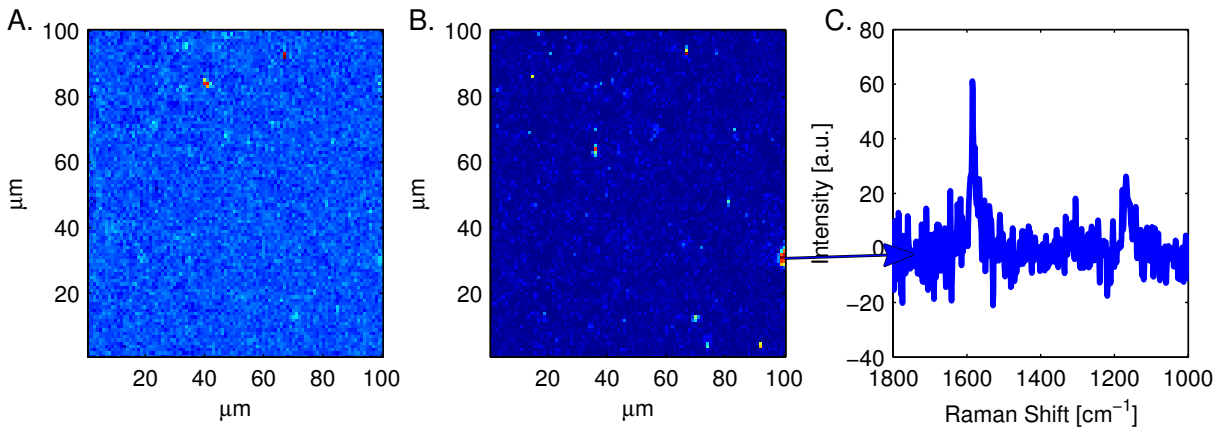

Fig. 6. Demonstration of NMF at low concentrations. A. Raman map of 1166 cm^{-1}. B. Loadings of the EG basis vector plotted as Raman Map. C. The spectrum at a location that is identified as having EG present.

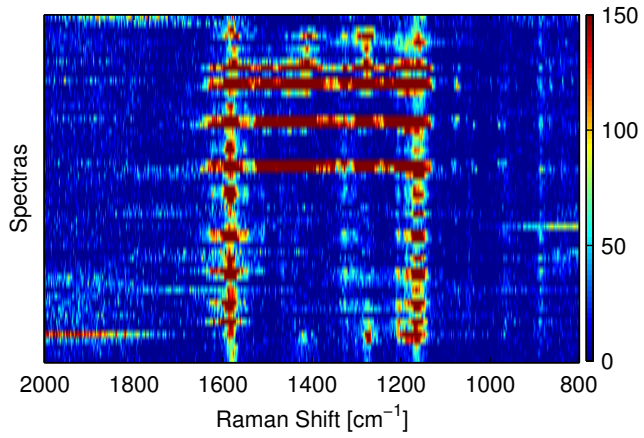

Fig. 5. Illustration of the raw spectra that is identified to have a loading of EG spectrum greater than two. The two main EG Raman components are present if all spectra, and EG is successfully identified.

peaks are easily distinguished. The spectra varies in shape compared to the normally observed spectra (blue). This type of interference is clearly caused by the Raman effect as seen by the very distinct peaks. This can come from one of the salts used in the buffer solution. As these are clearly located in a hot spot their signature is greatly amplified as is the EG.

Using the proposed method on the data, two Raman spectra are identified in $\mathbf{S}$ shown on figure fig. 4. Using the spectrum in $\mathbf{S}$ that is identified as an EG spectrum, the most dominant area for 1166 cm^{-1} has been correctly removed. In addition the mixed area (green spectrum on fig. 3) has been demixed into two distinct components, the EG spectrum and the contaminants. The loadings for the basis vector in $\mathbf{S}$ can now be used to identify where EG is likely to be present. Plotting the spectra that have a loading great than 3 (fig. 5) shows that the method correctly identifies spots where EG is present. At high concentrations the proposed method successfully identifies a Raman spectrum as one of the basis vectors in $\mathbf{S}$. The method is applied on the data shown in fig. 2, and one basis vector readily corresponds to a spectrum for EG.

At low concentrations, the NMF is not able to automatically identify the EG spectrum as one of the basis vectors. This is likely because the molecule is giving a weak signal as well as being a rare occurrence. Using the basis vector identified as EG (shown on fig. 4) as initial value in $\mathbf{S}$ and not sampling from the column that contains that basis vector, we are able to successfully identify areas with EG. An example is shown on fig. 6. Clearly the SNR is magnitudes lower compared to the high concentration situation that had a SNR of about 30.

4. CONCLUSIONS

A Bayesian NMF approach was used to analyze SERS data. The method was able to effectively decouple signals and at high concentrations it was able to identify the EG base spectrum as one of the basis vectors. This allows for more accurate and robust data analysis when the SERS substrate is contaminated by unknown molecules.

A further advantage of using NMF for SERS data is that the method is interpretable in the sense that the basis vectors identified can be related to the expected physical effects.

Possible future work is to use the loading matrix in a classifier in order to make an identification of the molecules that have bound with the substrate. Further, using a Bayesian framework allows for further definitions of the prior and a parametric prior that is physically explainable can by incorporated.

5. REFERENCES

[1] Anjum Qureshi, Yasar Gurbuz, and Javed H. Niazi, "Biosensors for cardiac biomarkers detection: A review," *Sensors and Actuators B: Chemical*, vol. 171-172, pp. 62–76, Aug. 2012.

[2] Maria Thunø, Betina Macho, and Jesper Eugen-Olsen, "suPAR: the molecular crystal ball.," *Disease markers*, vol. 27, no. 3, pp. 157–72, Jan. 2009.

[3] Evanthia Diamanti-Kandarakis, Jean-Pierre Bourguignon, Linda C. Giudice, Russ Hauser, Gail S. Prins, Ana M. Soto, R. Thomas Zoeller, and Andrea C. Gore, "Endocrine-disrupting chemicals: an Endocrine Society scientific statement.," *Endocrine reviews*, vol. 30, no. 4, pp. 293–342, June 2009.

[4] M Fleischmann, P. J. Hendra, and A. J. McQuillan, "RAMAN SPECTRA OF PYRIDINE ADSORBED AT A SILVER ELECTRODE," *Chemical Physics Letters*, vol. 26, no. 2, pp. 163–166, 1974.

[5] David L. Jeanmaire and Richard P. Van Duyne, "Surface raman spectroelectrochemistry:: Part I. Heterocyclic, aromatic, and aliphatic amines adsorbed on the anodized silver electrode," *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, vol. 84, no. 1, pp. 1–20, 1977.

[6] Kneipp Kneipp, Yang Wang, Harald Kneipp, Lev T. Perelman, Irving Itzkan, Ramachandra R. Dasari, and Michael S. Feld, "Single molecule detection using surface-enhanced Raman scattering (SERS)," *Physical Review Letters*, pp. 1667–1670, 1997.

[7] Eric C. Le Ru, Matthias Meyer, and Pablo G. Etchegoin, "Proof of single-molecule sensitivity in surface enhanced Raman scattering (SERS) by means of a two-analyte technique," *The journal of physical chemistry. B*, vol. 110, no. 4, pp. 1944–8, Feb. 2006.

[8] Zee H. Kim, "Single-molecule surface-enhanced Raman scattering: Current status and future perspective," *Frontiers of Physics*, vol. 9, no. 1, pp. 25–30, May 2013.

[9] Eric C. Le Ru, Evan J. Blackie, Matthias Meyer, and Pablo G. Etchegoin, "Surface Enhanced Raman Scattering Enhancement Factors: A Comprehensive Study," *The Journal of Physical Chemistry C*, vol. 111, no. 37, pp. 13794–13803, Sept. 2007.

[10] Pablo G. Etchegoin, Matthias Meyer, and Eric C. Le Ru, "Statistics of single molecule SERS signals: is there a Poisson distribution of intensities?," *Physical chemistry chemical physics : PCCP*, vol. 9, no. 23, pp. 3006–10, 2007.

[11] Jaeyoung Yang, Mirko Palla, Filippo G. Bosco, Tomas Rindzevicius, Tommy S. Alstrø m, Michael S. Schmidt, Anja Boisen, Jingyue Ju, and Qiao Lin, "Surface-enhanced Raman spectroscopy based quantitative bioassay on aptamer-functionalized nanopillars using large-area Raman mapping.," *ACS nano*, vol. 7, no. 6, pp. 5350–9, June 2013.

[12] Pablo G. Etchegoin, Matthias Meyer, Evan J. Blackie, and Eric C. Le Ru, "Statistics of single-molecule surface enhanced Raman scattering signals: fluctuation analysis with multiple analyte techniques.," *Analytical chemistry*, vol. 79, no. 21, pp. 8411–5, Nov. 2007.

[13] Pentti Paatero and Unto Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[14] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[15] Tommy S. Alstrø m, Jan Larsen, Claus H. Nielsen, and Niels B. Larsen, "Data-driven modeling of nano-nose gas sensor arrays," in *Proceedings of SPIE*, Ivan Kadar, Ed. SPIE, 2010, vol. 7697, pp. 76970U–76970U–12.

[16] Mikkel N. Schmidt, Jan Larsen, and Fu-Tien Hsiao, "Wind Noise Reduction using Non-Negative Sparse Coding," in *IEEE Workshop on Machine Learning for Signal Processing*. 2007, pp. 431–436, IEEE.

[17] Stephan Niebling, Hannes Y Kuchelmeister, Carsten Schmuck, and Sebastian Schlücker, "Quantitative, label-free and site-specific monitoring of molecular recognition: a multivariate resonance Raman approach.," *Chemical communications (Cambridge, England)*, vol. 47, no. 1, pp. 568–70, Jan. 2011.

[18] Hualiang Li, Tülay Adal, Wei Wang, Darren Emge, and Andrzej Cichocki, "Non-negative Matrix Factorization with Orthogonality Constraints and its Application to Raman Spectroscopy," *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 48, no. 1-2, pp. 83–97, Feb. 2007.

[19] S. Nie, S. Emory, "Probing Single Molecules and Single Nanoparticles by Surface-Enhanced Raman Scattering," *Science*, vol. 275, no. 5303, pp. 1102–1106, Feb. 1997.

[20] Michael a Ochsenkühn and Colin J Campbell, "Probing biomolecular interactions using surface enhanced Raman spectroscopy: label-free protein detection using a G-quadruplex DNA aptamer.," *Chemical communications (Cambridge, England)*, vol. 46, no. 16, pp. 2799–801, Apr. 2010.

[21] Cynthia V Pagba, Stephen M Lane, Hansang Cho, and Sebastian Wachsmann-Hogiu, "Direct detection of aptamer-thrombin binding via surface-enhanced Raman spectroscopy.," *Journal of biomedical optics*, vol. 15, no. 4, pp. 047006, 2013.

[22] Teodora Ignat, Roberto Munoz, Kleps Irina, Isabel Obieta, Miu Mihaela, Monica Simion, and Mircea Iovu, "Nanostructured Au/Si substrate for organic molecule SERS detection," *Superlattices and Microstructures*, vol. 46, no. 3, pp. 451–460, Sept. 2009.

[23] Michael Stenbaek Schmidt, Jörg Hübner, and Anja Boisen, "Large area fabrication of leaning silicon nanopillars for surface enhanced Raman spectroscopy.," *Advanced materials (Deerfield Beach, Fla.)*, vol. 24, no. 10, pp. OP11–8, Mar. 2012.

[24] EU, "Environment and Water: proposal to reduce water pollution risks," Tech. Rep. January, European Commission, 2012.

[25] Mikkel N. Schmidt, Ole Winther, and Lars K. Hansen, "Bayesian Non-negative Matrix Factorization," in *Independent Component Analysis and Signal Separation*. 2009, vol. 5441 of *Lecture Notes in Computer Science*, pp. 540–547, Springer Berlin Heidelberg.