NORTHWESTERN UNIVERSITY

Reducing Complexity: A Regularized Non-negative Matrix Approximation (NNMA) Approach to X-ray Spectromicroscopy Analysis

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Physics

By

Rachel Y. C. Mak

EVANSTON, ILLINOIS

December 2014

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

Reducing Complexity: A Regularized Non-negative Matrix Approximation (NNMA)
Approach to X-ray Spectromicroscopy Analysis

Rachel Y. C. Mak

X-ray absorption spectromicroscopy combines microscopy and spectroscopy to provide rich information about the chemical organization of materials down to the nanoscale. But with richness also comes complexity: natural materials such as biological or environmental science specimens can be composed of complex spectroscopic mixtures of different materials. The challenge becomes how we could meaningfully simplify and interpret this information. Approaches such as principal component analysis and cluster analysis have been used in previous studies, but with some limitations that we will describe. This leads us to develop a new approach based on a development of non-negative matrix approximation (NNMA) analysis with both sparseness and spectra similarity regularizations. We apply this new technique to simulated spectromicroscopy datasets as well as a preliminary study of the large-scale biochemical organization of a human sperm cell. NNMA analysis is able to select major features of the sperm cell without the physically erroneous negative weightings or thicknesses in the calculated image which appeared in previous approaches.

Acknowledgements

First and foremost, I would like to thank my advisor Chris Jacobsen for introducing me to this project, and then guiding me through it with incredible patience. Every time I thought I'd understood a concept but hadn't, and had to go back and ask again, was made much easier by his ability to explain a single idea in a dozen different ways, so that I did not feel too bad about not understanding it the first eleven times and instead was able to learn something new each time.

Thank you also to the rest of my thesis committee: Michael Bedzyk, Pulak Dutta, and Jens Koch – your insightful comments were valuable in clearing some murky points in this thesis.

I have benefited much from discussions with Stefan Wild and Sven Leyffer from the Mathematics and Computer Science Division at Argonne National Lab, who helped me realize the importance of rigour in mathematical derivations, and succeeded in at least helping me to not get so lost within all of the matrix indices. Likewise, the Northwestern Machine Learning Meetup Group co-organized by Jeremy Watt and Reza Borhani has been informative and supportive by bringing together people from different academic disciplines who yet share similar data analysis tools.

I have also been fortunate to have had opportunities to take part in pioneering beamtime experiments at the Advanced Photon Source as well as other light sources, where I have learnt much – including how complex microscopes operate and how to interpret images on

the screen – from Si Chen, Stefan Vogt, Sophie-Charlotte Gleber, Keith Brister, and all of the remarkable bionanoprobe team.

I was first introduced to x-ray spectromicroscopy while at Stony Brook University, where I was given much foundational knowledge and advice by former group members: Christian Holzner, Xiaojing Huang, Johanna Nelson, and Jan Steinbrener. The current members in our group at Northwestern: Junjing Deng, Young Pyo Hong, Kenan Li, Yue Sun, and Daikang Yan are now involved in different and fast-evolving projects, but it has been enlightening to (try to) keep up-to-date with the latest developments and provide input into one another's research.

I would like to thank all of my friends – running and cycling buddies, theatre and restaurant co-aficionados, skype and email correspondents – for providing a counter-balance to "academic" life. Finally, to my family: thank you for always being there.

Table of Contents

List of Figures

CHAPTER 1

Introduction

1.1. Spectromicroscopy

Spectromicroscopy is the combination of spectroscopy and microscopy. Microscopy has been used since around Galileo's time as a tool to image objects too small for the naked eye to see: in 1665, Robert Hooke published his detailed drawings of magnified insects in his book *Micrographia*, while Antonie van Leeuwenhoek and Zacharias Janssen were instrumental in developing some of the first microscopes in the 16^{th} and 17^{th} centuries. Spectroscopy separates the components in the images and enables us to carry out quantitative elemental and chemical analysis of these objects. We can combine the advantages of both techniques by dividing an object or sample into units of pixels, acquiring a spectrum over each pixel, and building up an image of the sample using its spectroscopic components – resulting in "spectroscopic imaging" or spectromicroscopy. A familiar example would be a colour image, from which we can obtain information about the wavelength-dependent response of each object in the image, which we could not do in a grayscale image (since it only contains scalar intensity information).

The first microscopes used visible light as the input (excitation) source as well as the output (imaging) medium, as did the first experiments in spectroscopy – Isaac Newton was said to have been one of the first to use a glass prism to split a ray of sunlight into its seven visible components. Today, the excitation source and imaging medium could be a combination of visible light, electrons, x-rays, or even neutrons and protons. Different excitation sources extract different types of information about the sample, and can provide different contrasts in imaging, thus providing different insights. And now, with the technological capability to image large numbers of pixels and acquire spectra across a large range at each pixel, the challenge – and the focus of this thesis – becomes the distillation and interpretation of this rich, complex data.

In the next section, we will briefly describe a few (by no means exhaustive!) spectromicroscopy techniques capable of extracting different types of electronic, molecular, or chemical information from samples. Then we will go further and discuss the physical mechanisms for x-ray absorption in materials and how they could be studied to gain chemical information.

1.2. Different spectromicroscopy techniques

1.2.1. X-ray fluorescence microscopy

X-ray fluorescence (XRF) probes the inner shell electronic transitions of atoms: when an inner-shell electron absorbs an incident x-ray photon and leaves a core hole, another electron from a higher orbital or shell can fill the hole, leaving the atom to emit a photon with characteristic energy unique to the element of the absorbing atom. An energy dispersive spectrometer can be used to measure the energy of the emitted photon. Thus, by scanning an XRF microscope over a sample, a map of its elemental distribution can be obtained. However, XRF is not an efficient radiation process and typically produces low yields for light elements ($Z \lesssim 11$).

1.2.2. Auger electron spectroscopy

For lighter elements down to lithium, the Auger effect provides the dominant energy transition pathway and can be taken advantage of by Auger electron spectroscopy (AES). Similar to XRF, an incoming x-ray photon (or electron) creates a core hole which is then filled by a higher-orbital electron; but instead of radiating a de-excitation photon as in XRF, the energy is absorbed by a second outer-shell (Auger) electron. The Auger electron is ejected from the atom if this de-excitation energy is greater than its orbital binding energy. This energy difference, or the kinetic energy of the Auger electron, is typically between ~ 20 and 500 eV. Due to the short mean free path of electrons in solids, AES is most suited to the analysis of surface compositions [**1**].

1.2.3. X-ray photoelectron spectroscopy

Like XRF, incident x-ray photons are used to eject photoelectrons from samples in x-ray photoelectron spectroscopy (XPS). But unlike XRF, this process is non-radiative. As in AES, XPS also has as a final state an emitted photoelectron, and so is also a surface-sensitive technique. As well as elemental identification, the peaks at different photoelectron energies in the XPS spectrum can be analyzed for shifts (with respect to some reference spectra) to obtain chemical state information, the importance of which will be discussed in more detail in the next section.

1.2.4. Electron energy loss spectroscopy

In electron energy loss spectroscopy (EELS), a near-mono-energetic beam of electrons is focused onto a thin section of the specimen. Some fraction of these electrons will transfer some of their energy via inelastic collisions. In transmission mode, the energy loss of the electrons can be measured as they traverse through the sample, and peaks corresponding to ionization energy loss can give information about the elemental and chemical composition of the sample. The energy of detected electrons range from 0.1 to 1 keV. EELS is sensitive to light elements (down to carbon) and transition metals. It is capable of high energy resolution (< 0.1 eV) and sub-nm spatial resolution [**2**] for thin films.

1.2.5. Infrared and Raman light spectromicroscopy

We include infrared (IR) and Raman spectromicroscopy here since they can be used to obtain complementary sample information to the techniques described above: while x-ray and electron spectroscopies probe electronic state information of the absorbing target, IR can probe the lower-energy vibrational and rotational modes of molecules without much sample perturbation, as well as their polarizability to reveal molecular orientation. In particular, near-IR has deep penetration and has even been used on living subjects [**3**]. IR spectro-microtomography (3D imaging) and reconstruction has recently been carried out to study the spatial distribution of chemical components in biological samples [**4**]. On the other hand, Raman (inelastic) scattering produces much weaker signals, but also offers higher sensitivity.

1.3. X-ray absorption spectroscopy

Section 1.2 summarised a few different variations of x-ray absorption spectroscopy (XAS) techniques, and lead to the point that we would now like to be able to probe samples beyond the surface and obtain not just electronic, but chemical state information from the bulk.

1.3.1. Chemical and oxidation state information

The chemical state of an atom is determined by its type or species as well as its bonding environment, which in turn is related to the nature and number of its neighboring atoms and the types of bonds formed. In theory, a single isolated atom (such as in a monatomic noble gas) would produce the "purest" absorption spectrum of that element. The absorption spectrum would simply consist of a generally decreasing trend as the probability of absorption decreases with increasing energy, punctuated by sharp absorption "edges" which correspond to sharp increases in absorption when the energy of incident photons matches the resonant energy required to excite an allowed electronic orbital transition in an atom. The first (lowest-energy) edge usually corresponds to the absorption of an x-ray photon by a K-shell electron, which is then ejected into the continuum. The optical density (product of absorption coefficient and sample thickness) of the element nickel is shown in Figure 1.1, showing the absorption K-edge at 8.33 keV.

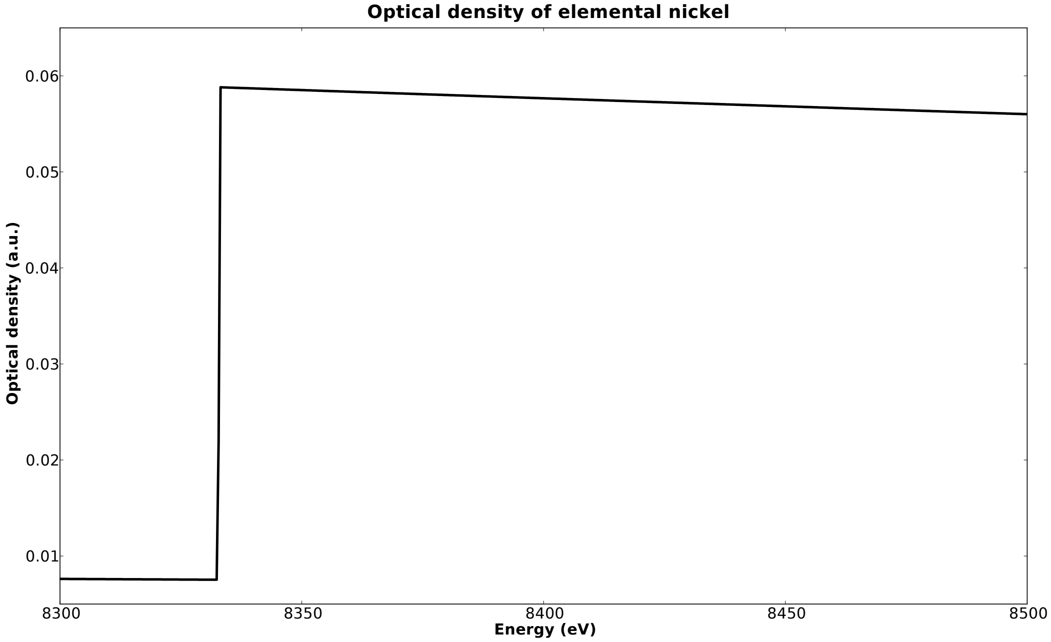


Figure 1.1. The optical density spectrum of an isolated Ni atom[a], showing the absorption K-edge at 8.33 keV.

[a] Data generated from `http://henke.lbl.gov/optical_constants/filter2.html`.

The spectrum is modified when the bonding environment around the absorbing atom changes – the atom could be packed within a crystalline solid and/or bonded with other atomic or molecular species. The forging of new molecular bonds and hybrid orbitals modulates the energy levels of the unbonded configuration. Even pure nickel in an amorphous form (*e.g.*, Ni foil) exhibits rich structures in its absorption spectrum, as shown in Figure 1.2.

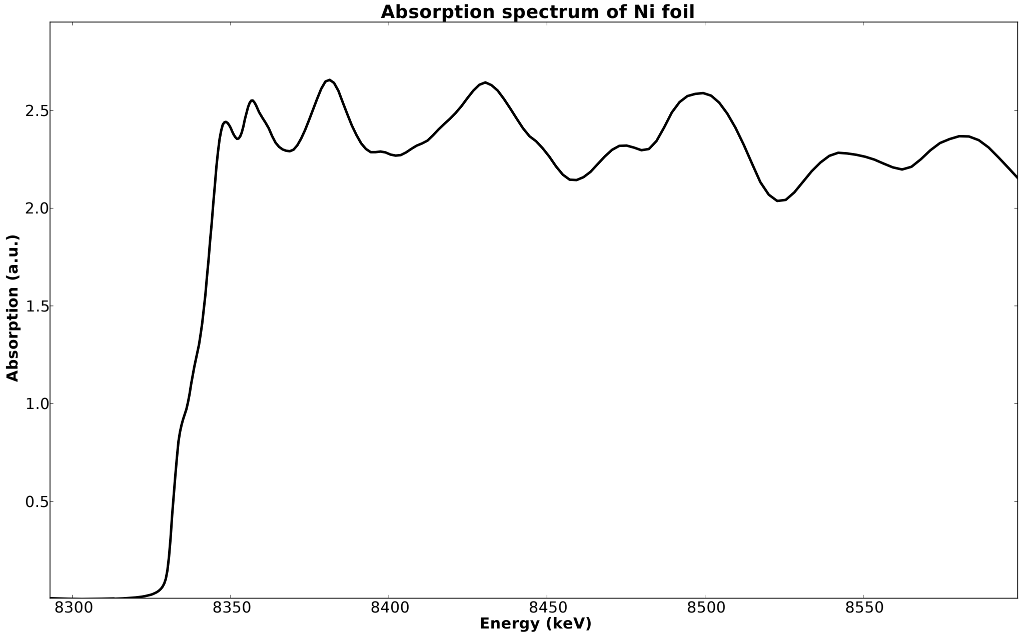


Figure 1.2. The absorption spectrum of a relatively "pure" sample (consisting of just one element) of Ni foil sample[a] can exhibit rich structures due to the absorbing atoms being surrounded (in this case, in an unstructured order) by other atoms. Resonances near the absorption edge and into beginning of the extended region (from $\sim$ 30 eV above the edge) are clearly observed.

[a] Data taken from `http://millenia.cars.aps.anl.gov/~newville/ModelData/ni_rt01.xmu`.

Chemical bonds usually affect the less strongly-bound states; they alter the binding energy levels and oxidation state of the atom, which manifest as shifts in the absorption edges of its spectrum. For example, the absorption spectra of several different types of Ni compounds are shown in Figure 1.3.

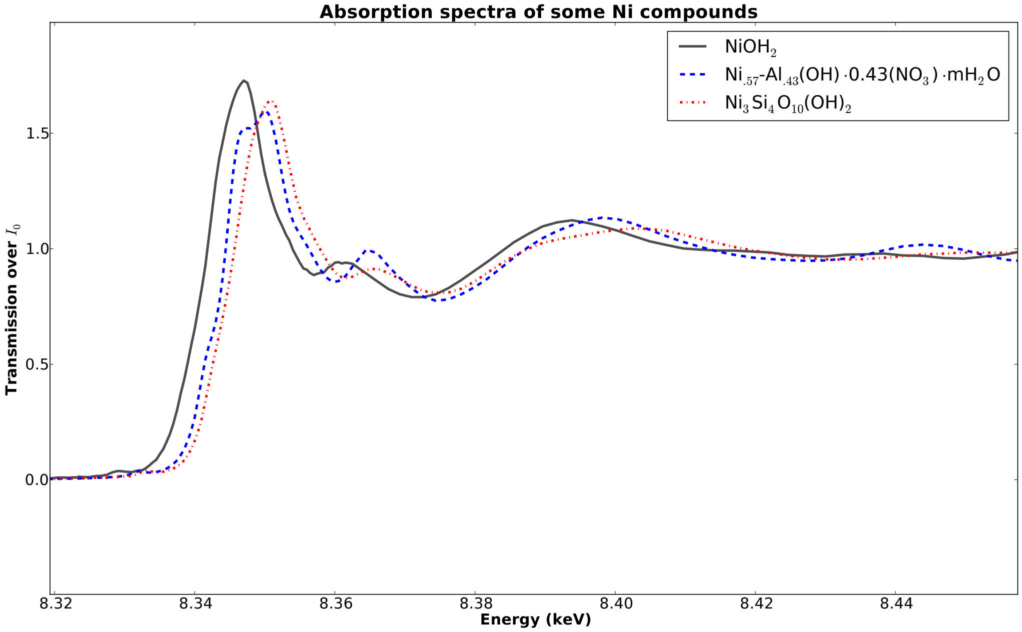


Figure 1.3. Absorption spectra of different compounds of Ni[a], showing relative shifts in their absorption edges, as well as different structures around the edge.

[a] Data taken from `http://www-ssrl.slac.stanford.edu/mes/spectra/compounds/ni/ni_index.html`.

We will next describe the x-ray absorption process which produces the modulations, resonances, and edge shifts in the absorption spectrum.

1.3.2. X-ray absorption cross-section

X-ray absorption involves the interaction of the photon's electromagnetic field with the momentum of a core electron in the absorbing atom and can be described by the electric dipole approximation. If we consider the electron as both a wave and a particle capable of absorbing an incident x-ray photon with energy equal to or greater than its binding energy, and transitioning from an initial quantum state $|i\rangle$ to a final state $\langle f|$, then we can calculate its transition probability per unit time $p_{i\to f}$ semi-classically using Fermi's Golden Rule [**5**]:

$$p_{i\to f} \propto \sum_{\substack{\text{all final states}\\ \text{with energy } E_f}} \left|\langle f|\boldsymbol{\epsilon}\cdot\mathbf{p}|i\rangle\right|^2 \delta(E_f - E_i - \hbar\omega), \tag{1.1}$$

where $\boldsymbol{\epsilon}$ is the polarization vector of the incident photon; $\mathbf{p}$ is the linear momentum operator of the core electron[1]; and the δ function ensures conservation of energy, *i.e.*, energy of the final state (E_f) is equal to the sum of energies of initial state (E_i) and incoming photon ($\hbar\omega$). The probability is summed over all possible final states with energy E_f allowed by the dipole selection rules. In the case of inner shell excitation, $p_{i\to f}$ represents the probability of a transition from the initial state $|i\rangle$ consisting of the incident photon and core electron to the final state $\langle f|$ of the core hole and ejected photoelectron in an available molecular orbital or the continuum – precisely the photoelectric effect. The larger the amount of overlap between the initial and final states, the higher the probability of the transition. Since this probability is measured by the absorption spectrum, it is this amount of overlap between the initial and final states that determines the structure of the absorption spectrum of an atom in a given environment. The shape of the absorption spectrum, or the behaviour of the absorption cross-section, can be explained as a combination of long- and short-range interactions – encoded in the x-ray absorption near-edge structures (XANES) and extended x-ray absorption fine structures (EXAFS) respectively. We will describe the physical mechanisms that contribute to these structures in the following two sections.

[1] The position operator $\mathbf{r}$ could also be used here, since $\langle f|\boldsymbol{\epsilon}\cdot\mathbf{p}|i\rangle \propto \langle f|\boldsymbol{\epsilon}\cdot\mathbf{r}|i\rangle$. This follows because the commutator of the Hamiltonian H with the position operator $\mathbf{r}$ is proportional to $\mathbf{p}$: $[H,\mathbf{r}] \propto -i\hbar\mathbf{p}$. So $\langle f|\boldsymbol{\epsilon}\cdot\mathbf{p}|i\rangle \propto \langle f|\boldsymbol{\epsilon}\cdot[H,\mathbf{r}]\,|i\rangle = (E_f - E_i)\,\langle f|\boldsymbol{\epsilon}\cdot\mathbf{r}|i\rangle$.

1.3.3. Extended x-ray absorption fine structure (EXAFS)

We discuss the EXAFS first since the interactions which give rise to the structures in this region can be explained in terms of a single-scattering approximation and are therefore simpler to understand. The EXAFS region begins from about 30 to 50 eV above the absorption edge, extending to beyond 1000 eV above the edge. This region contains information about the short-range structure around the target atom, such as the (bond) distances to its nearest neighbour atoms as well as the coordination number and chemical species of its surrounding atoms.

When an incident photon with energy E ejects a photoelectron with binding energy E_0 from an atom, the photoelectron travels as a wave radiating out in all directions and can scatter elastically off electrons from neighbouring atoms. This gives rise to outgoing and incoming waves whereby the electron interferes with itself, with constructive interference occurring when the round-trip distance $2R$ from the absorbing to scattering atom and back is an integer number n of the electron's de Broglie wavelength λ:

$$2R = n\lambda. \tag{1.2}$$

The wavelength of the electron is given by

$$\lambda = \frac{h}{p}, \tag{1.3}$$

where h is Planck's constant and p is the momentum of the electron.

Since the kinetic energy of the photoelectron is $E - E_0 = \frac{p^2}{2m_e}$ (where m_e is the mass of the electron) and its wavenumber $k \equiv \frac{2\pi}{\lambda}$, these two quantities are related by

$$k = \frac{2\pi}{h}\sqrt{2m_e(E - E_0)} \tag{1.4}$$

As we change E by scanning the photon energy in an XAS experiment, the resulting change in the wavenumber k of the photoelectron affects its interference pattern and modulates the absorption probability by a factor

$$\chi(k) \propto \frac{1}{kR^2}\sin(2kR), \tag{1.5}$$

where the $1/R^2$ factor accounts for the decay in scattering probability as the spherical waves spread out. The absorption probability is maximum when the interference is constructive, *i.e.*, when $2kR = \frac{(2n+1)}{2}\pi$.

If $\mu_0(k)$ represents the smoothly decreasing base absorption of an isolated atom above the edge, then the total absorption above the edge for a non-isolated atom can be written as

$$\mu(k) = \mu_0(k)[1 + \chi(k)], \tag{1.6}$$

and so

$$\chi(k) = \frac{\mu(k) - \mu_0(k)}{\mu_0(k)} \tag{1.7}$$

measures the deviation from the base absorption, or the EXAFS due to the interference between the backscattered and outgoing photoelectron waves originating from the absorbing atom.

The complete EXAFS equation is usually written in the form [**6, 7**]

$$\chi(k) = \sum_i N_i \frac{S_i(k)}{kR_i^2} e^{-\frac{2R_i}{\lambda(k)}} e^{-2\sigma_i^2 k^2} \sin[2kR_i + \delta_i(k)], \tag{1.8}$$

which accounts for the backscattering amplitude $S_i(k)$ from each of the N_i neighbouring atoms in the i^{th} shell at distance R_i from the absorbing atom; $e^{-\frac{2R_i}{\lambda(k)}}$ describes the decay of the photoelectron and the core hole as a function of the mean free path $\lambda(k)$ of the electron (the longer the path $2R$ the electron has to travel compared to $\lambda(k)$, the more likely it and the core hole would be to decay and thus lower the probability of scattering); $e^{-2\sigma_i^2 k^2}$ is the Debye-Waller factor associated with the variations in relative distances between the absorbing and backscattering atoms due to structural disorder or thermal fluctuations; and $\delta_i(k)$ is the additional phase shift caused by the potentials of the absorbing and backscattering atoms.

The EXAFS model described above is based on a single-scattering approximation, *i.e.*, the photoelectron scatters elastically off a neighbouring atom a distance R away and returns directly to the absorbing atom. However, it is also possible for the electron to take a more circuitous path and scatter off several different neighbouring atoms before returning to the absorbing atom. In this multiple-scattering case, we can consider the total interference path length $2R$ to be the sum of the scattering lengths [**7**]. At high energies with respect to the absorption edge, these multiple scattering paths correspond to rapidly oscillating terms in k space which tend to average out. If we Fourier transformed $\chi(k)$ into a real-space function describing the radial structure around the absorbing atom, we would not see much contribution from multiple scatterings at large R. However, at energies around E_0, the photoelectron has lower kinetic energy and thus a longer mean free path $\lambda(k)$. Multiple scattering contributions become important here and contribute to near-edge structures in the absorption spectrum (to be discussed in the next section).

By studying the EXAFS equation $\chi(k)$ in Eq. (1.8), we can gain information about the local structure surrounding the absorbing atom, such as the number N of its nearest neighbouring atoms (or coordination number) and their bonding distances R. The information is local because of the decay factor $e^{-\frac{2R}{\lambda(k)}}$, where scattering contributions from large R are suppressed. Typically, EXAFS is sensitive to about 3 Å around the absorbing atom [**8**].

1.3.4. X-ray absorption near-edge structure (XANES)

We now turn our attention to the low-k region of the absorption spectrum with respect to the absorption edge: typically the ~ 20 eV range below and the first ~ 30 eV above the edge. The EXAFS equation breaks down at low-k due to its $1/k$ dependence; also, the low kinetic energy and hence long mean free path of the photoelectron here means that the electron can undergo multiple scatterings and probe further out in R from the absorbing atom – therefore, the single-scattering approximation assumed above in the description of EXAFS is no longer valid. The resonances in this region are known as XANES, or sometimes also near-edge x-ray absorption fine structure (NEXAFS). While an ejected photoelectron in the EXAFS regime has high enough kinetic energy that it would likely encounter weak backscattering from only one neighbouring atom before escaping, an electron from the XANES regime just above the edge would have only just enough energy to reach and probe the empty continuum states just above the Fermi level (*e.g.*, for molecules, just above the highest occupied molecular orbital). The final state of the electron may be unbound or bound (in which case the transition appears as a pre-edge resonance just below the absorption peak), and could be atomic-like or contain orbital mixing due to molecular bond hybridization with neighbouring atoms. On the other hand, a low-k electron just below the absorption edge would not have the energy to escape into the continuum; thus resonances in the pre-edge region are indicative of bound state transitions.

The position of the absorption edge itself contains information about the chemical species of the absorbing atom, as well as being correlated with the atom's oxidation state or valency. In general, the edge energy shift (or chemical shift) is positive when the oxidation number increases, since having a smaller number of electrons per atom means that the effective nuclear potential increases, and each electron becomes more strongly bound to the nucleus.

Thus more energy would be required to remove a core electron, which is reflected in a higher edge energy.

XANES is particularly useful in the study of the K-shell spectra of low-Z molecules [**5**], which typically form short bond lengths (1.1 to 1.5 Å) that are strongly dependent on the types of bond hybridization – precisely the area most likely probed by the low-k electrons. Moreover, covalent bonds between low-Z molecules are strongly directional, and so produce polarization-dependent spectra (when excited by polarized radiation as in the case of a synchrotron) which contain information about the orientation of molecules. These factors combine to create K-shell spectra with strong features that reveal information about the configuration and bonding within and between low-Z molecules.

By probing the unoccupied electronic states using features just above the absorption edge, the bound states using features just below the the edge, and the valency or oxidation state using the position of the edge itself, XANES offers a sensitive tool for studying the chemistry of the environment surrounding the absorbing atom, such as the intramolecular bond type and length, band structure, orbital hybridization, molecular orientation, as well as the identification of chemical species within a sample. Because of its large scattering cross-section, XANES signals are much larger than those from EXAFS – making them easier to measure even at low sample concentrations. However, the nature of the multiple scatterings that give rise to XANES makes its analytical interpretation more complicated than in the case of EXAFS.

1.3.5. X-ray spectral analysis: where to go from here

In an x-ray spectromicroscopy experiment, the transmitted x-ray intensity $I(E)$ through each pixel in a sample is measured as a function of energy E (usually a range of energies that passes through an absorption edge of interest). This quantity is related to the linear absorption coefficient $\mu(E)$ of the material, and the sample thickness t via the Lambert-Beer Law:

$$I(E) = I_0 e^{-\mu(E)t}\,, \tag{1.9}$$

where I_0 is the incident energy (measured for each pixel).

With the capability of synchrotron x-rays to obtain rich, complex intensity spectra over millions of pixels, the challenge becomes how to interpret the collected data. In particular, we would like to find the absorption of the different components within the sample in order to understand its chemical composition. The analysis of image spectra over a continuous energy range is termed “hyperspectral imaging”, and was originally developed for remote sensing (for mineral and oil exploration) [**9, 10**]. The techniques developed have since been used in many applications including the measurement of surface CO_2 emissions in environmental monitoring [**11**], astronomy [**12**], and nano-drug delivery in biomedicine [**13**], and even tea quality classification [**14**]. Some data analysis techniques previously applied to x-ray spectromicroscopy include principal component analysis (PCA) and cluster analysis. We will review these two techniques in Chapter 2, and identify some shortcomings which will provide the motivation for this thesis: a development of non-negative matrix approximation (NNMA) which will be studied in more detail in Chapter 3, with applications to EXAFS and XANES datasets. We will discuss the results of one of the datasets in Chapter 4, before suggesting future work towards refining this technique.

CHAPTER 2

Principal component analysis (PCA) and cluster analysis

After having measured the transmitted intensity spectrum $I(E)$ at each pixel of the sample, we now have a 3D "stack" of 2D transmission images over a range of N energies, as illustrated in Figure 2.1. From this stack, we would like to be able to extract the (unknown) absorption spectrum $\mu(E)$ at each pixel of the sample, as well as its thickness weighting t. In this chapter, we first introduce the optical density, the transformed data quantity that we will work with directly instead of the intensity, and review two analysis methods for extracting components from the data: principal component analysis and cluster analysis.

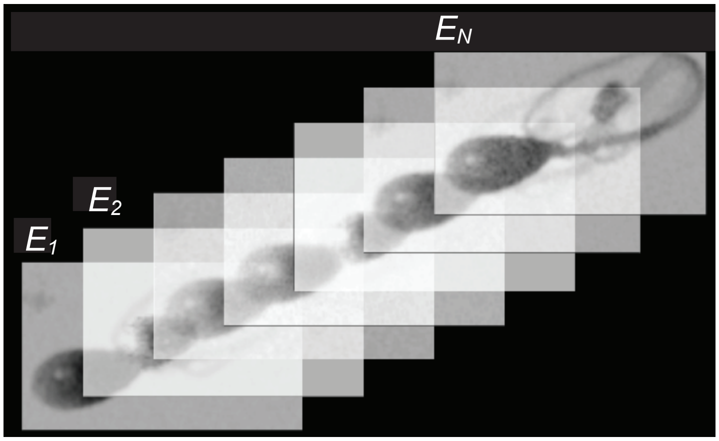


Figure 2.1. A 3D x-ray spectromicroscopy stack consists of a series of 2D transmission images measured at a range of energies from E_1 to E_N.

2.1. Conversion from transmission to optical density

At each pixel of the sample, we can convert $I(E)$ from Eq. (1.9) into an optical density spectrum $D(E)$, given by

$$D(E) = \mu(E)t \tag{2.1}$$

$$= -\log\left(\frac{I(E)}{I_0(E)}\right). \tag{2.2}$$

We choose to work with the optical density instead of the intensity because $D(E)$ is linear with the absorption $\mu(E)$ and sample thickness t – the two quantities that we would like to discover.

If we assign an index $n \in \{1 \ldots N\}$ to each of the N energies at which $I(E)$ has been measured, and an index $p \in \{1 \ldots P\}$ to each of the P pixels in the sample, we can form an optical density matrix $\mathbf{D}_{N\times P}$ where the n^{th} row represents the "flattened" 2D optical density image measured at energy E_n, as illustrated in Figure 2.2. In other words, the entry D_{np} then represents the optical density of pixel p at energy E_n.

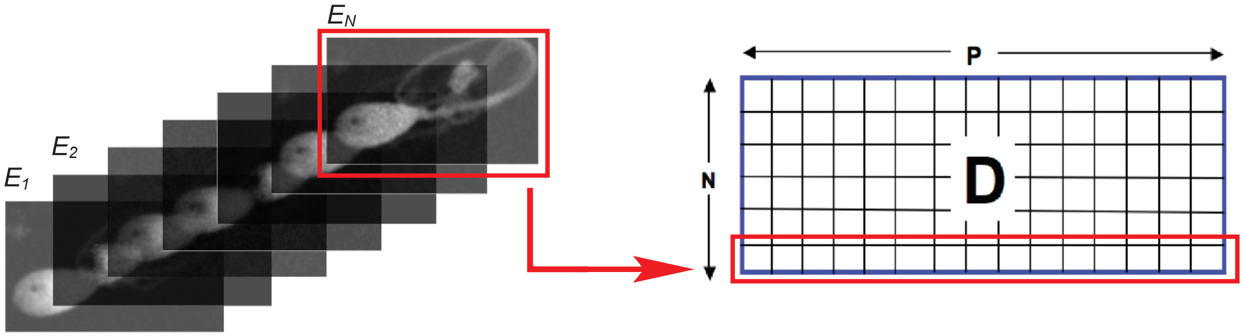


Figure 2.2. Transformation of the 3D stack of optical density spectromicroscopy images into a 2D optical density matrix $\mathbf{D}_{N\times P}$. The optical densities of the pixels in the n^{th} image (measured at energy E_n) are flattened into the n^{th} row of $\mathbf{D}_{N\times P}$.

Further, if the specimen contains S spectroscopically distinct components which we label $s \in \{1 \ldots S\}$, then from Eq. (2.1), we can write D_{np} as a linear combination of the absorption coefficients μ_{ns} of all the components s at energy E_n, weighted by the thicknesses t_{sp} of their respective components:

$$D_{np} = \mu_{n1} t_{1p} + \ldots + \mu_{nS} t_{Sp} \tag{2.3}$$

$$= \sum_{s=1}^{S} \mu_{ns} t_{sp} \, . \tag{2.4}$$

Accordingly, the full optical density matrix $\mathbf{D}_{N \times P}$ can be expressed as a matrix product:

$$\mathbf{D}_{N \times P} = \boldsymbol{\mu}_{N \times S} \mathbf{t}_{S \times P} \, , \tag{2.5}$$

where each column of $\boldsymbol{\mu}_{N \times S}$ represents the absorption spectrum of one component, and each row in $\mathbf{t}_{S \times P}$ represents the thickness weighting of the corresponding component. (We will drop the subscript indices on matrix variables from here on, unless we wish to clarify certain cases.)

The problem we would like to solve is this: given $\mathbf{D}$, can we find factors $\boldsymbol{\mu}$ and $\mathbf{t}$ which satisfy Eq. (2.5)? In addition, $\boldsymbol{\mu}$ and $\mathbf{t}$ are subject to the constraint that all the elements in each matrix must both be non-negative, *i.e.*, $\mu_{ns}, \ t_{sp} > 0 \quad \forall \ \{n, s, p\}$. The requirement for the non-negativity of $\boldsymbol{\mu}$ stems from the fact that negative absorption would imply that the sample is adding energy to the x-ray beam rather than taking energy away from it, leading to a violation of energy conservation. The requirement for the non-negativity of $\mathbf{t}$ stems from the fact that it is meaningless for a physical sample to have a negative thickness.

There are two ways of interpreting the factorization of $\mathbf{D}$ into $\boldsymbol{\mu}\mathbf{t}$. Conventionally, we consider each column of $\mathbf{D}$ to represent the experimentally collected spectrum of a pixel. We can consider this pixel to be a point in an N-dimensional energy space. Each of these points can be generated by a collection of S spectra (columns) in $\boldsymbol{\mu}$, *i.e.*, each column in $\mathbf{D}$ can be expressed as a linear combination of the S spectra in $\boldsymbol{\mu}$, with $\mathbf{t}$ specifying the weighting coefficients in the combination, as depicted in Figure 2.3.

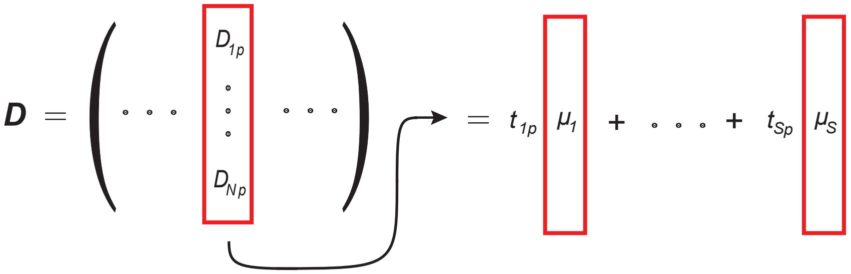


Figure 2.3. The p^{th} column of $\mathbf{D}$ can be thought of as a linear combination of the S spectra (columns of $\boldsymbol{\mu}$), weighted by the corresponding thickness coefficients in the p^{th} column of $\mathbf{t}$.

Alternatively, we can think of the problem in a transverse way: each row of $\mathbf{D}$ represents a P-dimensional point at a given energy (not pixel). Each of these points can be expressed as a linear combination of the S components (rows) in $\mathbf{t}$, with $\boldsymbol{\mu}$ specifying the weighting coefficients in the combination, as depicted in Figure 2.4.

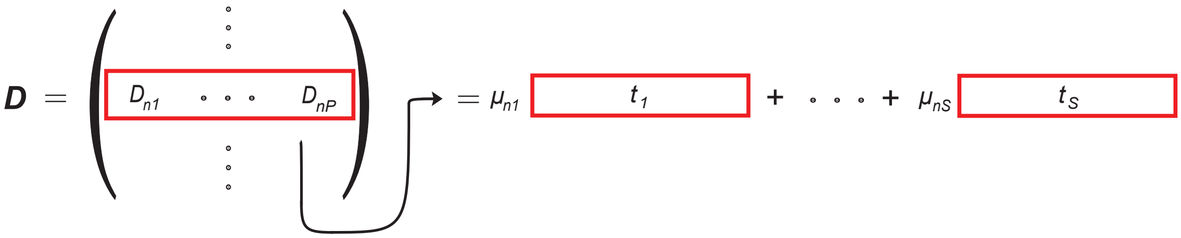


Figure 2.4. The n^{th} row of $\mathbf{D}$ can be thought of as a linear combination of the S components (rows of $\mathbf{t}$), weighted by the corresponding absorption coefficients in the n^{th} row of $\boldsymbol{\mu}$.

These two interpretations are equivalent, and both will be useful for our study. The first (column combination) view is useful because we are interested in recovering the absorption spectra of each spectroscopic component (which the column decomposition provides); while the second (row combination) view is useful because we are interested in reconstructing the spatial distribution of each component (which the row decomposition provides).

2.2. Principal component analysis

The problem of analyzing the measured data $\mathbf{D}$ in terms of a set of spectra $\boldsymbol{\mu}$ has been the subject of a variety of multivariate statistical analysis approaches in energy loss electron microscopy [**15, 16**] and in infrared spectromicroscopy [**17, 18**]. In x-ray spectromicroscopy, approaches using spectral standards or hand-defined regions assumed to be of uniform, pure composition can be used to obtain a set of S spectra $\boldsymbol{\mu}$ from which thickness maps $\mathbf{t}$ can be calculated using singular value decomposition (SVD) for matrix inversion [**19, 20**]:

$$\mathbf{t} = \boldsymbol{\mu}^{\dagger}\mathbf{D} \tag{2.6}$$

where $\boldsymbol{\mu}^{\dagger}$ is the pseudo-inverse of $\boldsymbol{\mu}$. However, the thickness maps obtained in this way are not guaranteed to be non-negative, which is one of the two non-negativity contraints required when finding factor solutions to Eq. (2.5).

Principal component analysis (PCA) is a similar but more general approach that has been used to understand complex samples in x-ray microscopy [**21, 22**] by reducing the dimensionality of the dataset in order to identify the significant components or a basis set of $\bar{S}$ (reduced S) orthonormal spectral signatures.

PCA is a very well-established statistical analysis method, used as early as 1901 to fit planes by means of orthogonal least squares [**23**]. From a basic viewpoint, PCA is like a generalized version of regression, which seeks to find the best-fit line or hypersurface to the data while at the same time trying to eliminate correlations (or redundancies) among variables. As mentioned, one use of PCA is dimensionality reduction in order to describe as many variations in a multivariate system with as few variables as possible. In other words, we seek to maximize the variance of a linear combination of variables. The first principal component is the linear combination with maximal variance; essentially, we search for a dimension along which the observations are maximally separated. The second principal component is the linear combination with maximal variance but in a direction orthogonal to the first principal component, and so on, up to the n^{th} dimension or principal component.

PCA is therefore useful in sorting data according to importance: large variances (represented by the most significant components) carry important information regarding the dynamics in the data, while small variances (least significant components) represent mostly noise. By cutting out the latter, PCA produces better signal-to-noise ratio and leads to a more structured description of the data. There may be some loss of information, but this should be minimized.

To find the significant components in PCA, we first calculate the covariance matrix $\mathbf{Z}_{N\times N}$ given by

$$\mathbf{Z}_{N\times N} = \mathbf{D}_{N\times P}\,\mathbf{D}^T_{P\times N}. \tag{2.7}$$

The eigenvectors $\boldsymbol{\mu}_{N\times S}$ and eigenvalues $\boldsymbol{\Lambda}_{N\times N}$ associated with $\mathbf{Z}_{N\times N}$ are related by

$$\mathbf{Z}_{N\times N}\,\boldsymbol{\mu}_{N\times S} = \boldsymbol{\mu}_{N\times S}\,\boldsymbol{\Lambda}_{S\times S}\,, \tag{2.8}$$

where $\boldsymbol{\Lambda}_{S\times S}$ is a diagonal matrix with the s^{th} entry on the diagonal corresponding to the s^{th} column (eigenvector or eigenspectrum) in $\boldsymbol{\mu}_{N\times S}$. The magnitudes of the eigenvalues can be displayed on a scree plot in order to choose the number of significant spectroscopic components S. For example, the scree plot of a simulated dataset that will be studied later in Chapter 3 is shown in Figure 2.5, clearly indicating four significant components. Their associated eigenspectra are shown in Figure 2.6.

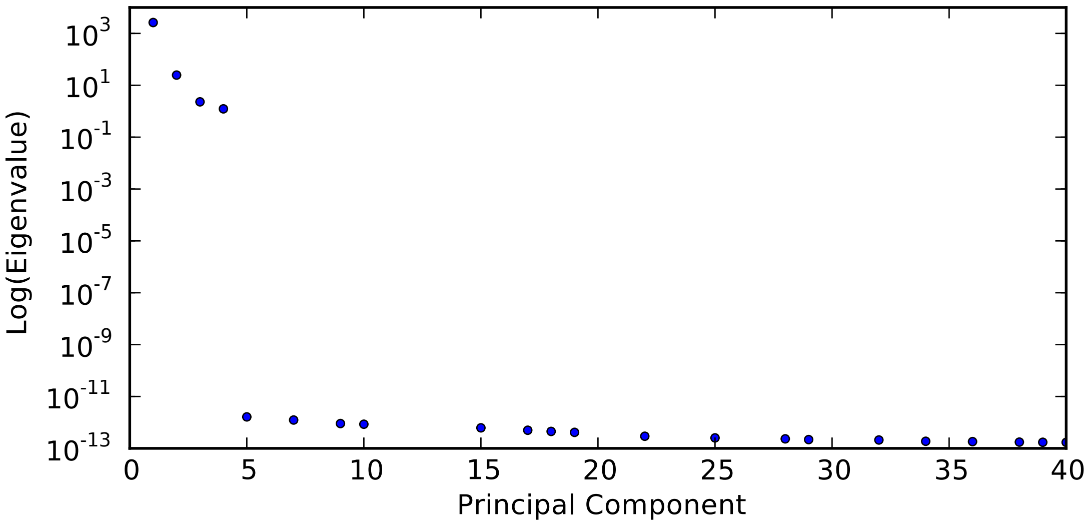


Figure 2.5. Scree plot showing the ranking of eigenvalues from PCA on a simulated four-component dataset that will be studied later in Chapter 3. In this case, four significant components are clearly indicated. However, in many real datasets, the number of truly significant components may not be so clear-cut.

We have seen that the SVD inversion of $\boldsymbol{\mu}^{\dagger}$ does not guarantee a non-negative thickness map $\mathbf{t}$, while PCA produces a basis set $\boldsymbol{\mu}$ which may include negative spectral values. Therefore, neither of these approaches satisfies the non-negative condition of our desired solution described in Eq. (2.5).

2.3. Cluster analysis

While PCA does not provide a set of spectra which are individually all interpretable as non-negative absorption spectra of materials present in the sample, it does provide an orthogonal and reduced-dimensionality search space for cluster analysis [**24, 25**] by specifying a minimum number of clusters in which to group pixels with similar spectra.

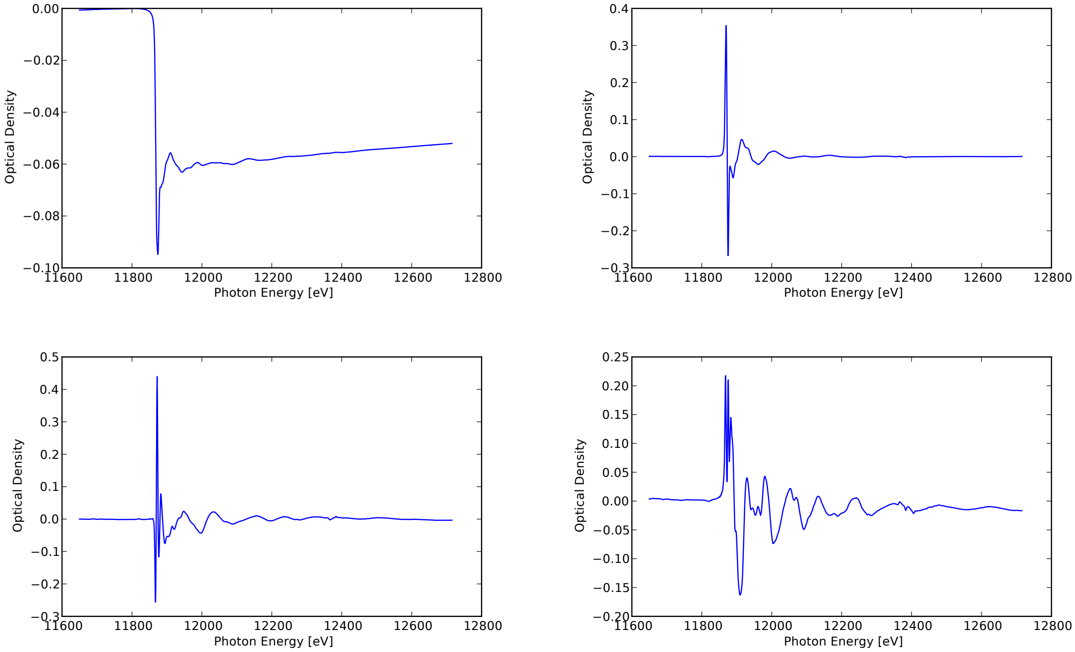


Figure 2.6. The four PCA eigenspectra (columns of $\boldsymbol{\mu}_{N\times S}$) corresponding to the four most significant eigenvalues in Figure 2.5. As can be seen here, PCA can produce negative values in the eigenspectra, which does not satisfy our constraint of a non-negative $\boldsymbol{\mu}$.

The goal of cluster analysis is to find an optimal grouping for which the observations or objects within each cluster have a high degree of similarity, but the clusters themselves are dissimilar to one other. The measure of similarity could be defined by a Euclidean distance or some other metric. Cluster analysis is one of the most common unsupervised learning algorithms and has also been referred to as unsupervised pattern recognition.

Cluster analysis is different from classification analysis in that the latter allocates the observations to a known number of groups, while in cluster analysis, neither the number of groups nor the types of groups are known in advance – hence unsupervised learning. There are different types of clustering algorithms: hierarchical clustering [**26**], Gaussian mixture models [**27**], neural networking [**28**], k-means clustering [**24, 25**], spectral clustering [**29**], and even a combination such as hierarchical clustering of a mixture model [**30**], amongst others.

In k-means clustering, we would like to group pixels into clusters according to the similarity in their spectral content. The number of clusters k (equivalent to our number of components S) is estimated using the number of significant eigenvalues from PCA as a minimum. We begin with a preliminary, random choice for the locations of cluster centres – each cluster is characterized by some random linear combination of the weightings of each component. Then, we choose a pixel p and find the cluster centre closest to p according to our chosen distance metric, and move this cluster center towards the pixel by a scaled distance as described in [**24**]. This is repeated for each of the P pixels, and then the whole process iterated several times until the positions of the cluster centers have stabilized. This learning vector quantization algrorithm is described in more detail in [**24**]. Once the cluster centres are found, the spectra calculated from each cluster provide a set $\boldsymbol{\mu}$ for the calculation of thickness weightings $\mathbf{t}$ according to Eq. (2.6).

Although cluster analysis has proven to be useful for a variety of applications [**31, 32**], it has also been observed to yield some regions with slightly negative values in the thickness maps $\mathbf{t}$, which, as mentioned above, are non-physical and thus represent limitations in the analysis.

In order to understand the way in which non-negative errors can arise in cluster analysis, we consider a simple example of a specimen with uniform thickness and a continuum of composition starting with 100% of material A which is strongly absorptive at energy E_1, and ending with 100% of material B which is strongly absorptive at energy E_2 (see Figure 2.7).

A scatterplot of the location of individual pixels based on their responses at the energies $\{E_1, E_2\}$ is shown schematically in Figure 2.8A. If we were to apply cluster analysis to these

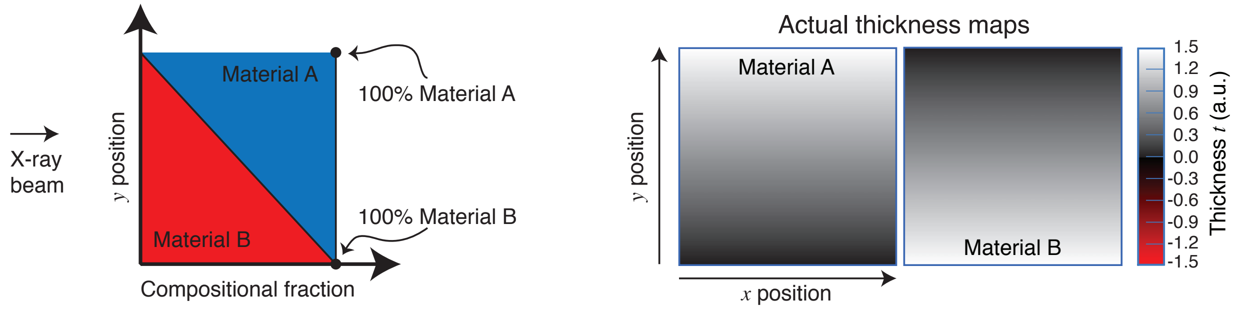


Figure 2.7. A sample of uniform overall thickness, composed of two different materials with continuously varying ratios as a function of y. The x-ray beam is travelling along the z direction on the plane of the page, and as it scans along the y direction, experiences different amounts of absorption depending on the relative amounts of the two materials.

pixels and tried to sort them into two clusters, this would result in the groupings shown in Figure 2.8B, where the vectors shown point to the respective cluster centers. These cluster basis vectors are the spectra or columns of $\boldsymbol{\mu}$ (with $S = 2$ in this example), from which we could calculate the thickness maps according to Eq. (2.6). The spectral composition of any given pixel in the sample can be described by some linear combination of these basis vectors. However, consider the case of a pixel which is far from the median in composition, such as the one at the upper left in Figure 2.8C. The composition of that pixel involves a negative weighting of the red cluster basis vector – *i.e.*, one that produces negative values in the thickness map $\mathbf{t}$, which we have already mentioned is not physically interpretable and therefore undesirable. Of course, if the variation among the spectral responses of the pixels assigned to a cluster is small, these errors can be negligibly small; however, as Figure 2.8 shows, there is no guarantee that cluster analysis will produce a thickness map with few negative pixels. Indeed, this negative thickness error is exactly what is observed in an actual cluster analysis run applied to data of the form of Figure 2.7, and will be discussed in more detail in Chapter 3.

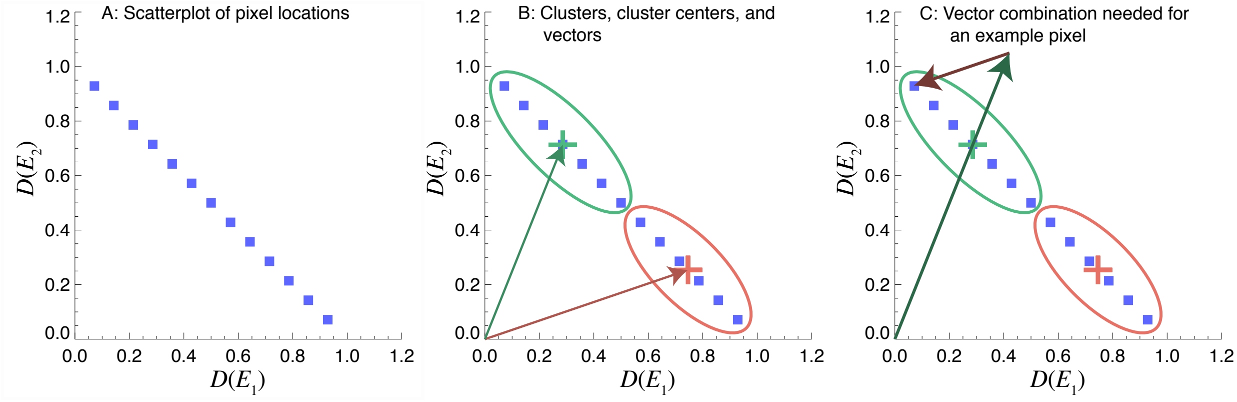


Figure 2.8. Schematic illustration to visualize how the linear combination of cluster analysis components could result in negative thickness weightings.

2.4. The next step: non-negative matrix approximation

The negative weightings produced in PCA and cluster analysis can be interpreted as negative thicknesses, which does not make physical sense. Alternatively, the weightings can be interpreted as a probability distribution of components over each pixel – again, it does not make sense to have negative probabilities. It is then natural to try to find an approximate factorization of $\mathbf{D}$ while imposing a non-negative constraint on the values of $\boldsymbol{\mu}$ and $\mathbf{t}$, with the goal of looking for a physically meaningful interpretation of spectromicroscopy data. This is the motivation for this thesis, the rest of which will be devoted to such an analysis using non-negative matrix approximation, its applications to EXAFS and XANES datasets, as well as a discussion of its shortcomings and possible improvements.

CHAPTER 3

Non-negative matrix approximation (NNMA)

3.1. Introduction

Given a set of input data (such as a set of images), the goal of principal component analysis, cluster analysis, and other factor analysis methods is to find a set of basis vectors (or basis images), which by linear combination can reconstruct any of the original input data. However, there are many instances where the concept of simply linearly combining (adding and subtracting) factors to obtain the final result does not make intuitive sense [**33**]. Consider the case of face recognition in image analysis: if a face were to be reconstructed by adding *and subtracting* faces from a set of basis images, these basis images with negative facial features would not be physically intuitive. What does it mean to subtract a part of the face? What is the meaning of the part of the face which has been subtracted (a negative face component, in a sense)? It would be more meaningful if we were to reconstruct a face by *additively* building up from a set of facial features. In text mining, words are grouped into a basis set, then added up to reconstruct a document; then it does not make sense to "negatively" include a set of words – they should either exist or not exist in a particular document.

If we constrain the weighting coefficients in the linear combinations to be non-negative, the basis images can then only be added (and not subtracted) to reconstruct the desired images. This creates basis images that represent parts. This additive, parts-based approach to factorization is called non-negative matrix approximation (NNMA) and has been used since at least 1994, when it was first known as positive matrix factorization [**34**]. The concept of NNMA is simple: factorize a given non-negative matrix $\mathbf{D}_{N\times P}$ (where all of its entries are non-negative) into two non-negative matrix factors $\boldsymbol{\mu}_{N\times S}$ and $\mathbf{t}_{S\times P}$, where $S < \min\{N, P\}$, such that $\mathbf{D}$ can be approximated as

$$\mathbf{D} \approx \boldsymbol{\mu}\mathbf{t}. \tag{3.1}$$

There is much interest in the study of finding non-negative factorizations in the form of Eq. (3.1) due its many applications in generating useful, insightful (and perhaps commercially profitable!) information. We have already mentioned the application of NNMA to feature extraction in face recognition analysis and text mining in document classification. A tiny sampling of other applications includes digital image processing (intensity in each image pixel is non-negative) [**35**], gene expression analysis (probability of expression is non-negative) [**36, 37**], PageRank for the ordering of web pages (rank is non-negative) [**38, 39**], recommendation systems used by online retail and service providers such as Amazon and Netflix (consumer ratings are non-negative) [**40**], and hyperspectral remote sensing (reflectances of objects are non-negative) [**41**].

3.2. Minimization of the cost function F

The approximation in Eq. (3.1) can be achieved by minimizing the cost function F (also known as the objective function), where

$$F(\boldsymbol{\mu}, \mathbf{t}) = \frac{1}{2}||\mathbf{D} - \boldsymbol{\mu}\mathbf{t}||_F^2. \tag{3.2}$$

Here, $||\mathbf{X}||_F^2 = \sum_{ij} x_{ij}^2$ is the squared Frobenius norm of the matrix $\mathbf{X}$, as described in more detail in Appendix A.1. We constrain each element in $\boldsymbol{\mu}$ and $\mathbf{t}$ to be non-negative.

There are many algorithms for minimizing Equation (3.2), including alternating least squares [**42, 43**], gradient-based methods [**44, 45**], and multiplicative updates derived from gradient descent [**46**]. We will base our approach on the last of these, because the rules are simple to derive and easy to implement, and the results could be interpreted in a direct way. However, we note that though multiplicative update rules have been widely used in different applications thanks to their low complexity, they suffer from a lack of good convergence properties [**44**]. But, as we shall see in our results, even if we cannot always guarantee a global minimization, locally minimized solutions can still contain useful information about the data.

The minimization of the cost function F in Eq. (3.2) with respect to both $\boldsymbol{\mu}$ and $\mathbf{t}$ simultaneously is classified as a non-convex optimization problem. While convex problems have the desirable property of a globally optimal solution (a global minimum or maximum) that could be found within a reasonable amount of time (*i.e.*, within polynomial time $\mathcal{O}(n^k)$ where k is some constant independent of input size n), this is not necessarily the case for non-convex problems, which are more intractable than convex ones [**47**]. Non-convex optimization problems are considered to be NP-hard[1] problems which cannot be solved in polynomial time. To achieve useful results in polynomial time, we resort to local optimization techniques to search for locally optimal solutions (local minima or maxima) instead. While local optimization methods can be fast, they require an initial guess for the solution or a starting point for the search, and the value of the final local solution can depend sensitively on this starting point; also, locally optimal solutions are not guaranteed to be global (the cost function is not guaranteed to be the smallest possible over the whole domain), and may not

[1] NP is the set of all decision problems that can be verified in polynomial time. Decision problems are problems with simple answers that are binary: yes/no, true/false, etc. Many optimization problems can be formulated as decision problems. NP-hard problems are outside the set of NP, and hence cannot be solved in polynomial time. NP stands for "non-deterministic polynomial time".

contain information about how far away the true, globally optimal solution lies. Nonetheless, there could still be useful information to be gained from a locally optimal solution.

In our case, though the problem of minimizing F with respect to $\boldsymbol{\mu}$ and $\mathbf{t}$ simultaneously is non-convex, it becomes a convex optimization problem if we minimize with respect to only one of the variables $\boldsymbol{\mu}$ or $\mathbf{t}$ while keeping the other variable fixed. Therefore, a local optimization method would be to alternate the search between the $\boldsymbol{\mu}$ and $\mathbf{t}$ domains.

3.3. The NNMA algorithm

We use the multiplicative rules derived by Lee and Seung [**46, 48**] to iteratively update $\boldsymbol{\mu}$ and $\mathbf{t}$ alternately. These update rules originate from the more familiar gradient descent approach – at each iteration, we take a small step in the direction of the steepest negative gradient:

$$\begin{aligned}\mathbf{t} &\leftarrow \mathbf{t} - \epsilon_{\mathbf{t}} \frac{\partial F}{\partial \mathbf{t}} \\ &= \mathbf{t} + \epsilon_{\mathbf{t}}(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D} - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t}) \qquad (3.3)\\ \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} - \epsilon_{\boldsymbol{\mu}} \frac{\partial F}{\partial \boldsymbol{\mu}} \\ &= \boldsymbol{\mu} + \epsilon_{\boldsymbol{\mu}}(\mathbf{D}\mathbf{t}^{\mathrm{T}} - \boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}}) \qquad (3.4)\end{aligned}$$

where $\epsilon_{\boldsymbol{\mu}}$, $\epsilon_{\mathbf{t}} > 0$ are small step sizes. The expansions of the partial derivatives of F in Eqs. (3.3) and (3.4) are shown in Appendix B.

Lee and Seung set the step sizes to be:

$$\epsilon_{\mathbf{t}} = \frac{\mathbf{t}}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t}} \qquad (3.5)$$

$$\epsilon_{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}}{\boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}}} \qquad (3.6)$$

(division is element-wise, so that $\epsilon_{\mathbf{t}}$ has the same dimensions as $\mathbf{t}$, and $\epsilon_{\boldsymbol{\mu}}$ has the same dimensions as $\boldsymbol{\mu}$.) in order to obtain the multiplicative update rules:

$$\mathbf{t} \leftarrow \mathbf{t}\,\frac{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D}}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t}} \tag{3.7}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}\,\frac{\mathbf{D}\mathbf{t}^{\mathrm{T}}}{\boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}}}\ . \tag{3.8}$$

After making an initial guess for the number of components S and initializing $\boldsymbol{\mu}$ and $\mathbf{t}$ with non-negative random entries, we update each variable independently using the multiplicative update rules, and use these to calculate the cost function F. We then use the updated $\boldsymbol{\mu}$ and $\mathbf{t}$ to find the next iteration of updates, and calculate F again. We repeat this procedure until we find the minimum of F.

In practice, when the number of iterations is large enough, F may decrease at ever smaller increments, so we stop the iterations if one of two conditions are met:

(1) the change in the cost function, ΔF, satisfies $-10^{-6} < \Delta F < 0$ (this interval can be adjusted depending on how much mileage one would like to get out of the algorithm – *i.e.*, how close one would like to get to the exact local minimum, but we have not found much difference in the quality of the reconstructed $\boldsymbol{\mu}$ and $\mathbf{t}$ as long as the magnitude of ΔF is sufficiently close to zero);

(2) the number of iterations exceeds some set maximum number (*e.g.*, `maxIters` $=$ 10^4, but this can be varied depending on the behaviour of the convergence with a particular set of parameters).

The basic NNMA procedure is outlined in Figure 3.1.

```
Basic NNMA procedure
(1) Estimate the number of components $S$.
(2) Initialize $\boldsymbol{\mu}_{N\times S}$, $\mathbf{t}_{S\times P}$ with random, non-negative entries.
(3) While (iters < maxIters) and ($\Delta F(\boldsymbol{\mu},\mathbf{t}) < -10^{-6}$) do:
    (i) Update $\mathbf{t} \to \mathbf{t}' = \mathbf{t} \cdot \frac{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D}}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t} + \epsilon}$;
    (ii) Update $\boldsymbol{\mu} \to \boldsymbol{\mu}' = \boldsymbol{\mu} \cdot \frac{\mathbf{D}\mathbf{t}'^{\mathrm{T}}}{\boldsymbol{\mu}\mathbf{t}'\mathbf{t}'^{\mathrm{T}} + \epsilon}$;
    (iii) Calculate $F(\boldsymbol{\mu}',\mathbf{t}')$.
```

Figure 3.1. The basic NNMA algorithm, using the multiplicative update rules of Lee and Seung. In the updates, the symbol $\cdot$ denotes element-wise matrix multiplication; division is also element-wise. A small $\epsilon \sim 10^{-9}$ is added to the denominator in the update rules to prevent division by zero.

3.4. Application of NNMA to spectromicroscopy

As discussed in Chapter 2, we measure the transmission intensity of the sample at each pixel and for each energy, and convert this into a matrix of optical densities, $\mathbf{D}$, for all pixels and energies.

Each column in $\mathbf{D}$ represents the optical density spectrum at each pixel (or if viewed row-wise, each row in $\mathbf{D}$ represents the optical density image at one particular energy). Each column in $\boldsymbol{\mu}$ contains the absorption spectrum of each component as a function of energy; and each row in $\mathbf{t}$ represents the (flattened) image of one component (or, if viewed column-wise, each column in $\mathbf{t}$ represents the components detected at each pixel). From an NNMA perspective, the S components in the sample are the "hidden features" that we would like to recover.

In sections 3.4.2 and 3.4.3, we simulate two different datasets using available reference absorption spectra; we will describe the results of the basic NNMA procedure on these datasets, and compare the reconstructed spectra with their respective references. In section 3.4.4, we will do the same for XANES data from a human sperm cell. But first, we describe the reference spectra used to simulate our datasets.

3.4.1. Reference spectra for the simulation of datasets

In our simulated datasets, we use a set of reference EXAFS spectra from different arsenic (As) compounds[1] in various mixtures to test whether NNMA could be used to decompose the mixtures into the individual input components. For reference, the As K-edge is at 11.867 keV [**49**]. The As compounds we have chosen are As_2O_3 (arsenic trioxide), As_2O_5 (arsenic pentoxide), As_2S_3 (arsenic trisulfide), and GaAs (gallium arsenide). Their EXAFS spectra are shown in Figure 3.2, with measurement points at 320 different energies from 11.650 to 12.715 keV. The spectra are also shown superimposed on one another in Figure 3.3(a), while a close-up highlighting the relative shifts in their absorption edges are shown in Figure 3.3(b). Taking the point of inflection of the first sharp rise in each spectrum as the location of the absorption edge, we see that there is an energy difference of only about 6 eV separating the edges furthest apart from one another (between As_2O_3 and As_2S_3). It is thus important to test whether NNMA would be able to identify the closely-spaced but distinguishable absorption edges of the different spectra.

[1] The As compound spectra were provided by Matt Newville from GSECARS (GeoSoilEnviro Consortium for Advanced Radiation Sources), and are also available from `cars.uchicago.edu/~newville/ModelLib/search.html`.

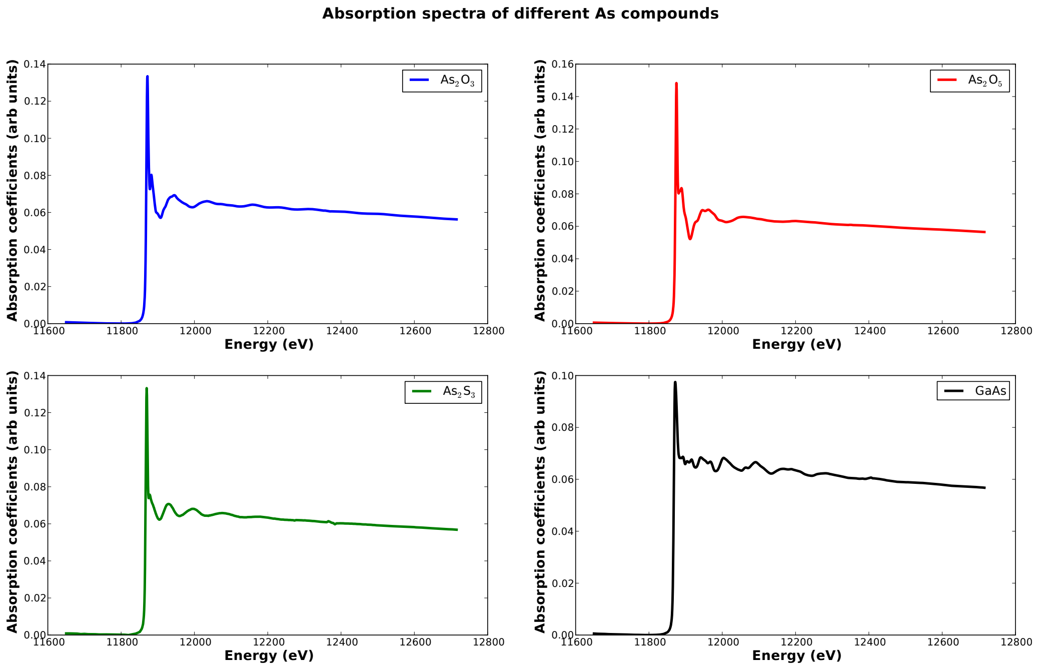
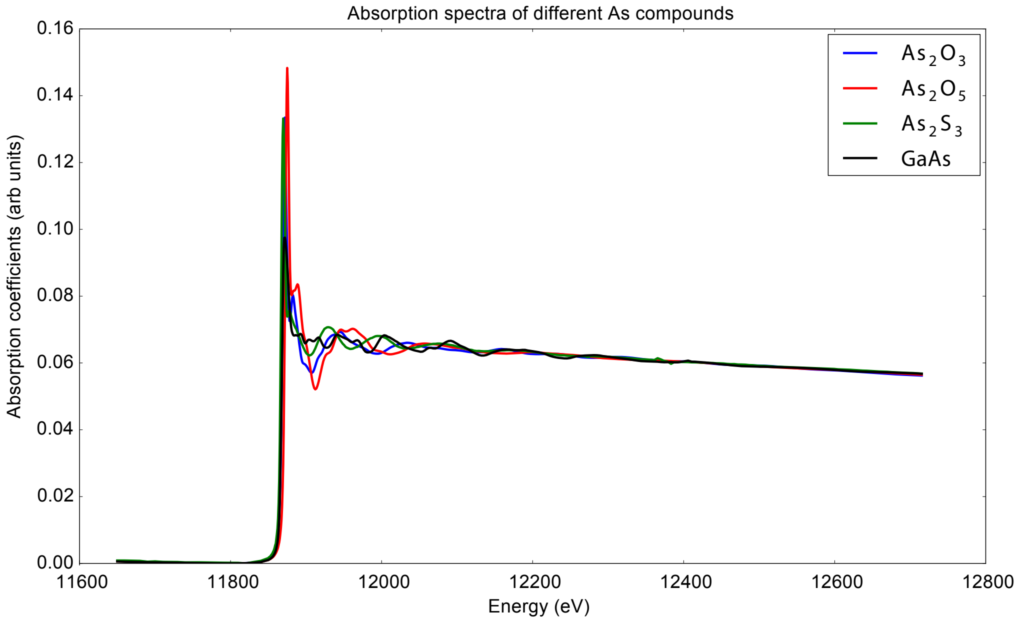
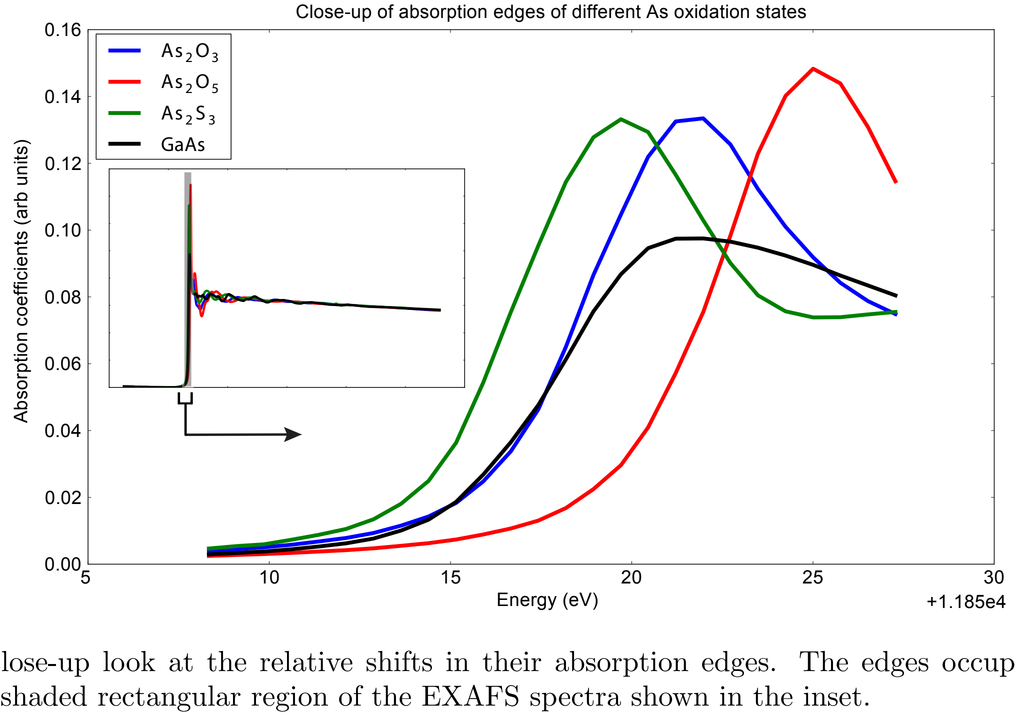


Figure 3.2. Reference arsenic (As) compound spectra used in simulating datasets in Sections 3.4.2 and 3.4.3.



(a) The four reference As compound spectra superimposed on one another.



(b) A close-up look at the relative shifts in their absorption edges. The edges occupy the very narrow shaded rectangular region of the EXAFS spectra shown in the inset.

Figure 3.3. The reference As compound spectra used in simulating datasets described in Sections 3.4.2 and 3.4.3.

3.4.2. Simulated dataset: a two-wedge prism

We construct a hypothetical rectangular prism put together from two right-angled triangular wedges, as shown in Figure 3.4. The blue wedge is composed of As_2O_3 (component 1), while the red wedge is composed of As_2O_5 (component 2). The xy surface of the prism is 256×64 pixels. (The absolute thickness in the z direction does not matter in this example as we are only interested in the relative ratios of the two components at each pixel.) The prism is constructed in such a way that when the x-ray beam is oriented along the z direction and scanned across the x and y directions, it will "see" the ratio of the mixture of the two components vary smoothly in the y direction, *i.e.*, the weighting or thickness map of each component will vary in a smooth gradient as a function of y. Each row of the optical density matrix, $\mathbf{D}_{16384 \times 320}$, then consists of one sweep of the x-rays at one energy across the

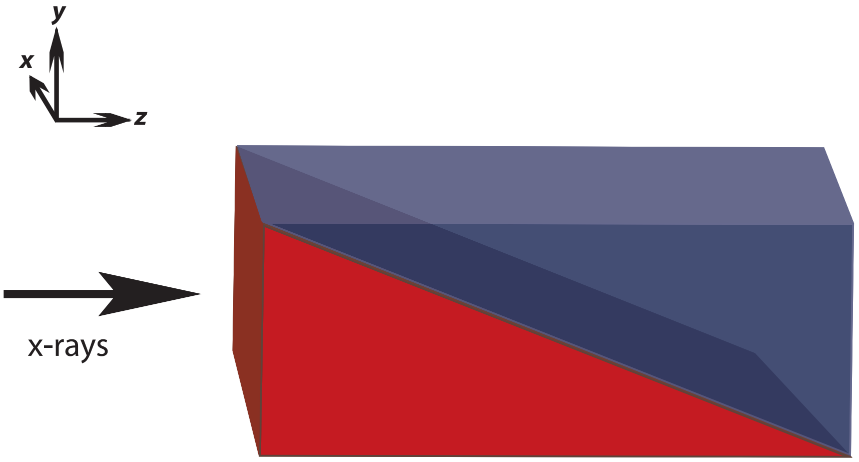


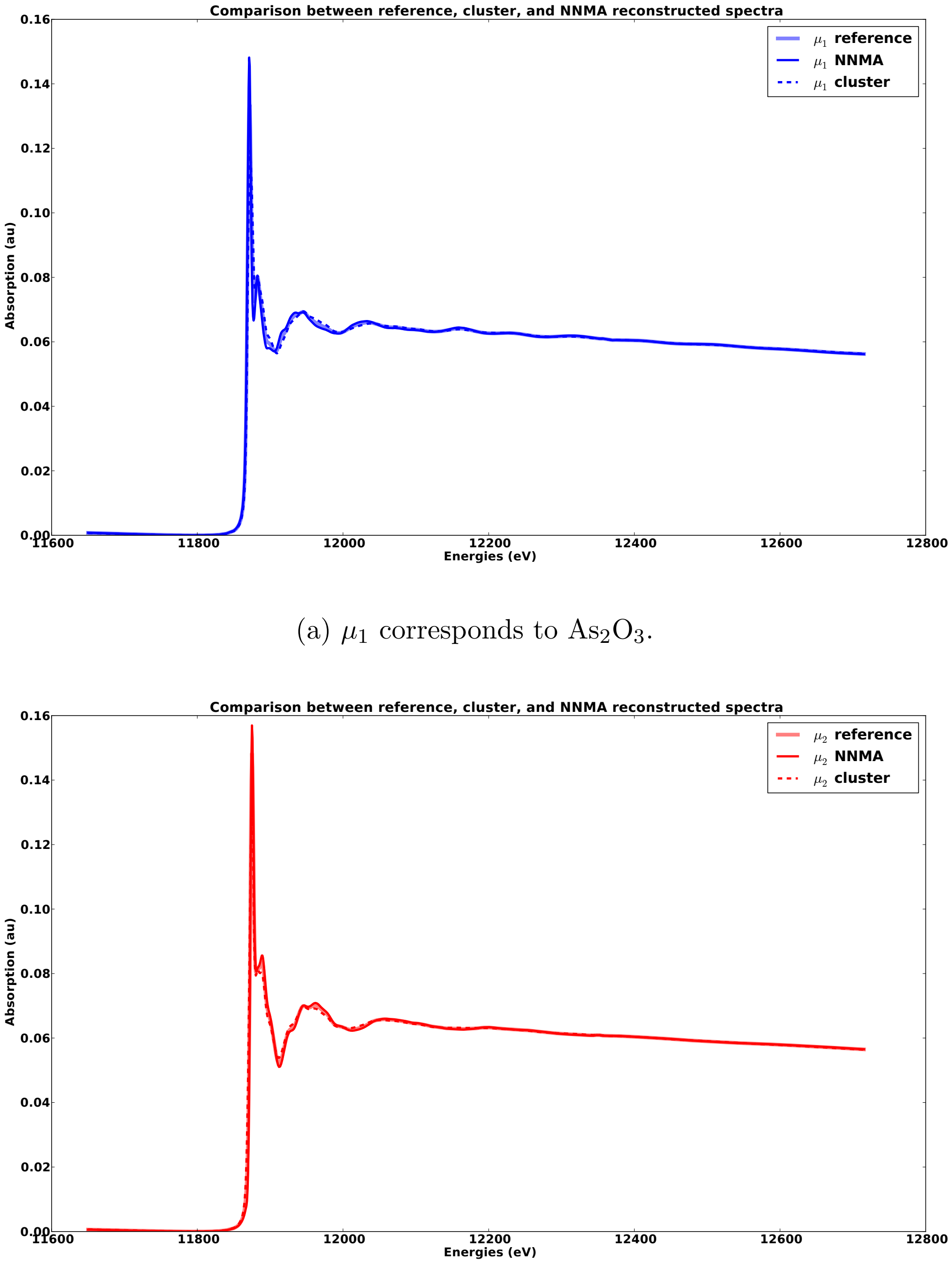
Figure 3.4. A model of the two-wedge prism, composed of two different chemical components in varying ratios (varying in the y direction as seen by the x-rays), used as a simulated dataset to test NNMA. In this example, the material in the blue wedge is composed of As_2O_3, while the red wedge is composed of As_2O_5.

xy surface of the prism. And at each pixel of the surface, the absorption spectrum will be different along the y direction depending on the ratio of component 1 to component 2.

We take the As_2O_3 and As_2O_5 absorption spectra in Figure 3.2 and set them as the two columns in our $\boldsymbol{\mu}_{320\times 2}$ matrix, while each row of the $\mathbf{t}_{2\times 16384}$ matrix specifies the relative thickness of each component across all the pixels on the surface of the prism (*e.g.*, the entries corresponding to the row of pixels on the prism halfway up in the y direction would be 0.5 for both components 1 and 2). Of course, we would not have the luxury of knowing $\boldsymbol{\mu}$ and $\mathbf{t}$ beforehand in a real experiment; but for the purpose of this simulation, we multiply these two matrices to obtain our "experimental" optical density, $\mathbf{D}$. The idea is to find out whether, given only $\mathbf{D}$, how well we can recover $\boldsymbol{\mu}$ and $\mathbf{t}$ using the NNMA procedure described in Section 3.3.

The reconstructed absorption spectra are shown in Figure 3.5, along with a close-up of the absorption edge reconstruction in Figure 3.6. The NNMA parameters used for this reconstruction are $S = 2$ and 10,000 iterations. Cluster analysis reconstructions are also shown for comparison. The corresponding thickness maps are shown in Figure 3.7 and 3.8.

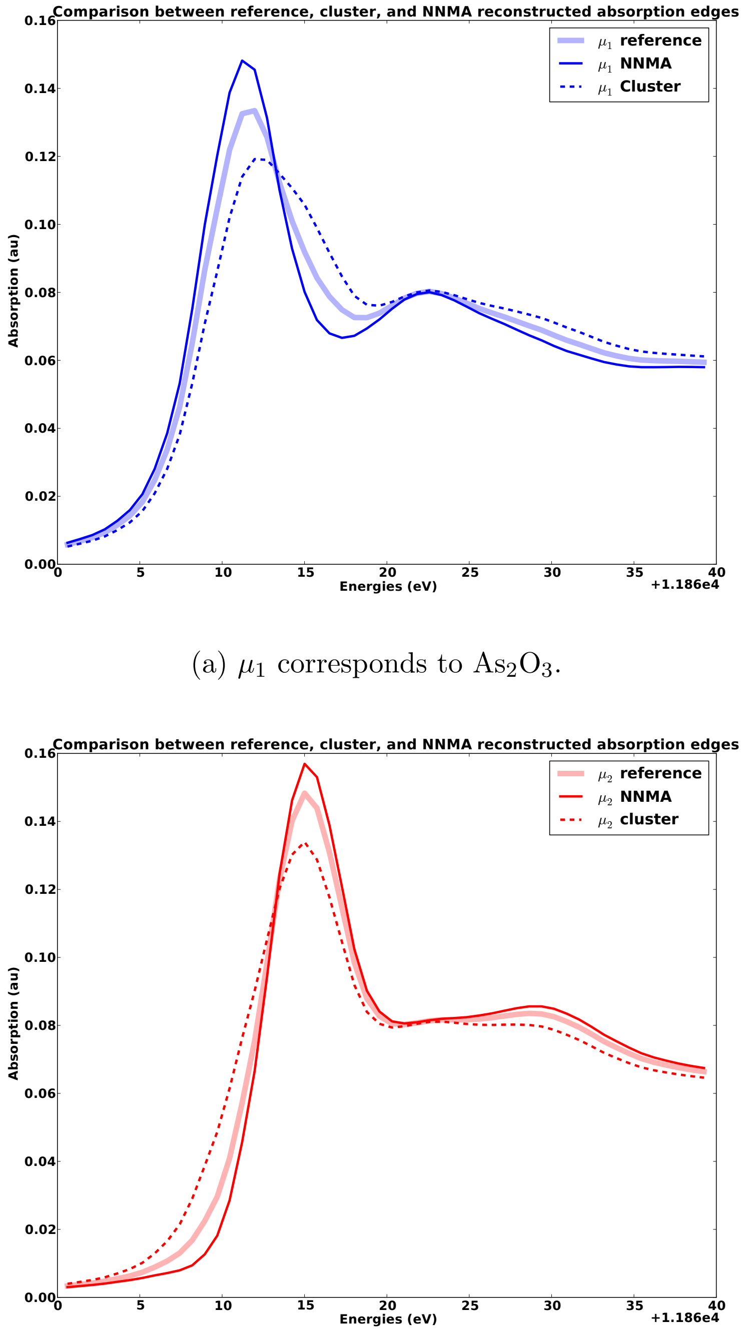
From the point of view of cluster analysis, each pixel on the surface of the prism is assigned exclusively to one of two clusters (or components). However, as can be seen from Figure 3.4, this is only true for the very top row of pixels where y is maximum (all pixels there are seen by the x-rays as composed purely of component 1 in blue), and the very bottom row of pixels where $y = 0$ (composed purely of component 2 in red). All of the other pixels in between are composed of some mixture of components 1 and 2. Therefore, when using cluster spectra to fit the observed pixel spectra in $\mathbf{D}$, the error is largest in the mid-y region where the prism is composed of exactly the same amount of component 1 as component 2, as shown in the cluster error map in Figure 3.8(b).



(a) μ_1 corresponds to As_2O_3.

(b) μ_2 corresponds to As_2O_5.

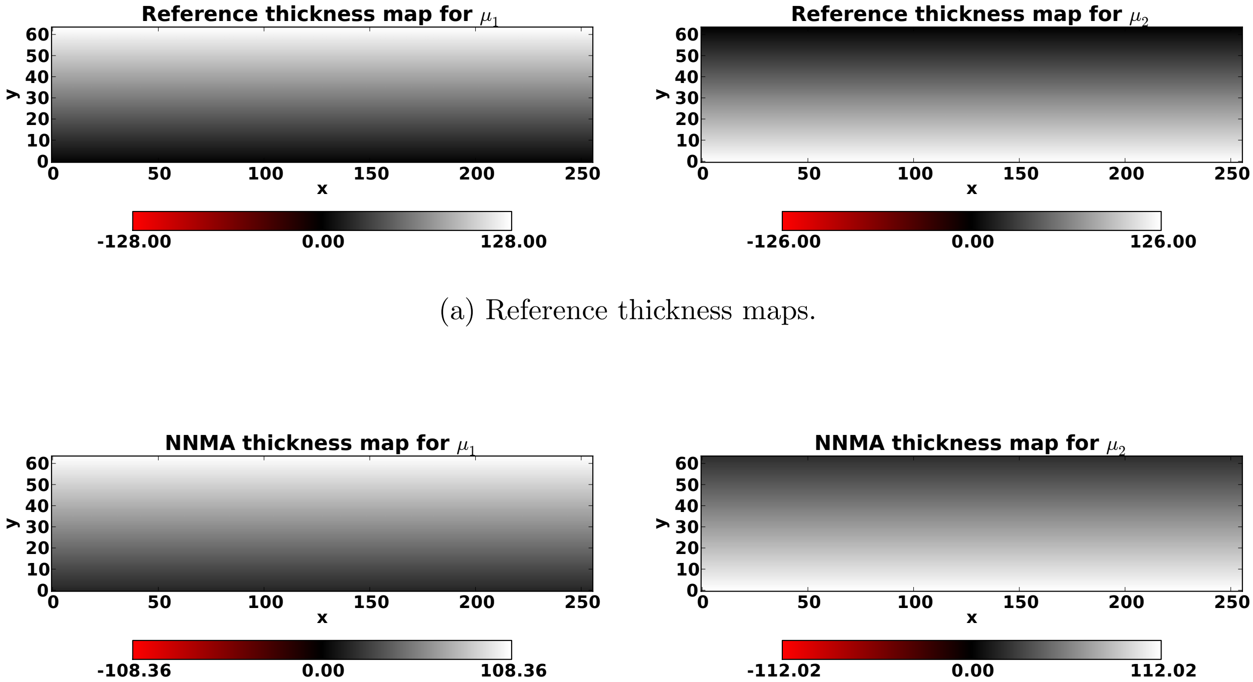
Figure 3.5. Comparison between reference, cluster, and NNMA reconstructed spectra for the simulated two-wedge prism sample. Close-ups of the reconstruction around the absorption edge regions are shown in Figure 3.6.



(a) μ_1 corresponds to As_2O_3.

(b) μ_2 corresponds to As_2O_5.

Figure 3.6. Comparison of the region around the absorption edges between reference, cluster, and NNMA reconstructed spectra for the simulated two-wedge prism sample.



(a) Reference thickness maps.

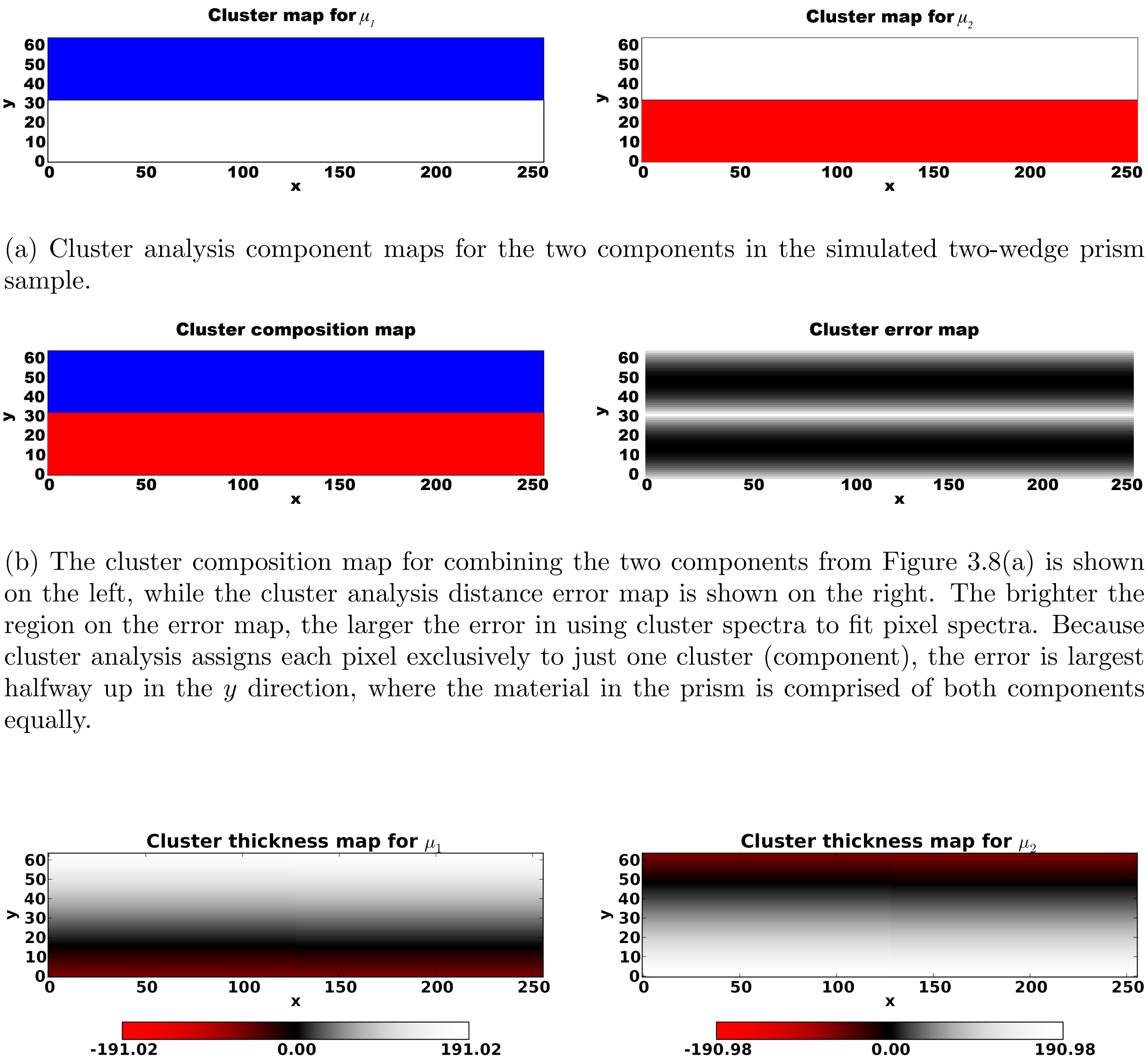
(b) NNMA reconstructed thickness maps.

Figure 3.7. NNMA reconstruction of the thickness maps for the two components in the simulated two-wedge prism sample, compared with the reference thickness maps. In this case, μ_1 corresponds to As_2O_3, while μ_2 corresponds to As_2O_5. We can also compare these to the thickness maps obtained from cluster analysis shown in Figure 3.8.

For the simulated two-component prism sample, NNMA is able to more accurately reconstruct the edge regions of the absorption spectra compared with cluster analysis, along with a better approximation of the thickness maps without any negative regions.

3.4.3. Simulated dataset: alphabetical letters

Next, we simulate a dataset consisting of four alphabetical letters ('A', 'B', 'C', and 'D') using all four of the As compound EXAFS spectra from Section 3.4.1. The composition of



(a) Cluster analysis component maps for the two components in the simulated two-wedge prism sample.

(b) The cluster composition map for combining the two components from Figure 3.8(a) is shown on the left, while the cluster analysis distance error map is shown on the right. The brighter the region on the error map, the larger the error in using cluster spectra to fit pixel spectra. Because cluster analysis assigns each pixel exclusively to just one cluster (component), the error is largest halfway up in the y direction, where the material in the prism is comprised of both components equally.

(c) While the cluster analysis component map is discrete (each pixel belongs to one and only one cluster), we can use SVD to invert the cluster spectra in order to obtain a continuous thickness maps via Eq. (2.6). As can be seen here, these thickness maps contain some negative (red) regions towards the lower y values on the map corresponding to μ_1, and higher y values on the map corresponding to μ_2.

Figure 3.8. For comparison with Figure 3.7, here are the cluster analysis component maps (top); the combined component map and error map (middle); and the thickness maps obtained from the singular value decomposition of cluster spectra (bottom).

the letters are 'A' → As_2O_3, 'B' → GaAs, 'C' → As_2S_3, and 'D' → As_2O_5. The letters have been written into a 128 × 128-pixel background with no absorption. Each column of $\boldsymbol{\mu}_{320\times4}$ contains one of the four As reference spectra, while each row (component) of $\mathbf{t}_{4\times16384}$ contains a 1 where a pixel belongs to the letter composed of that component, and 0 otherwise (*e.g.*, the entries in the first row of $\mathbf{t}$ which correspond to the pixels of 'A' are set to 1, while the rest of the entries are 0). We again multiply $\boldsymbol{\mu}$ and $\mathbf{t}$ to obtain a simulated optical density matrix $\mathbf{D}_{320\times16384}$, on which we perform the NNMA procedure as before.

The NNMA parameters we used were $S = 4$ and 10,000 iterations. Figure 3.9 shows the reconstructed spectra alongside their corresponding thickness maps.

We find that NNMA does not perform as well on the letters dataset as on the prism sample: the structure of the reconstructed spectra differ from those of the reference; and each of the thickness maps, instead of containing just a single letter (representing one of the four original components), contains some combination of letters. This indicates that the extracted components are made up of some linear combination of the reference components. The simulated letters dataset differs from the two-wedge prism sample used in Section 3.4.2 mainly in the sparseness of $\mathbf{t}$. For the prism, most of the pixels are composed of some mixture of two components, and so most of the entries in the $\mathbf{t}$ matrix are non-zeros. On the other hand, each of the letters is composed purely of just one component, on top of a non-absorbing background, and so most ($> 70\%$) of the entries in $\mathbf{t}$ should be zeros. If we have prior information about the sparseness (or other properties) of the dataset, as in the case with the letters example and also in many real experimental samples that we might be interested in studying, then we could implement a strategy to introduce this characteristic into the NNMA procedure, in order to obtain reconstructions that better model the data. This scheme, called "regularizations", will be discussed in Section 3.6.

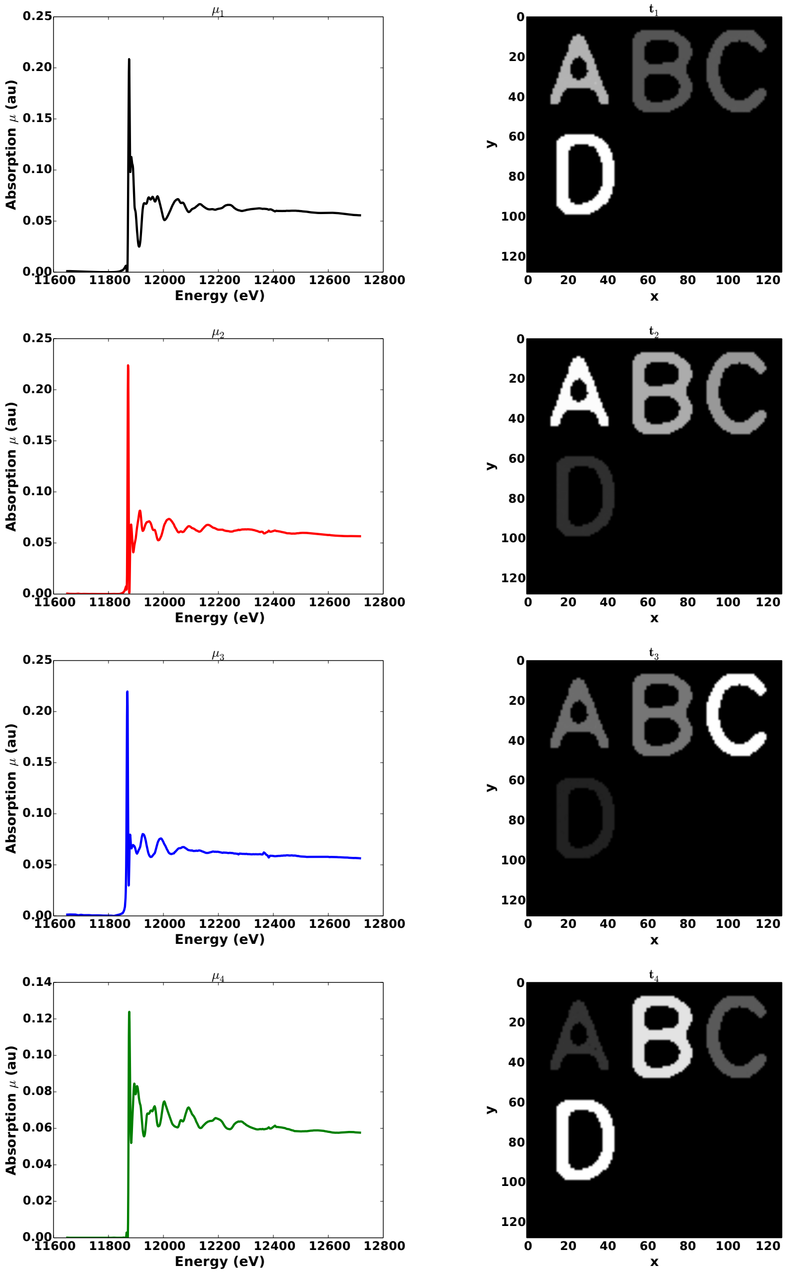


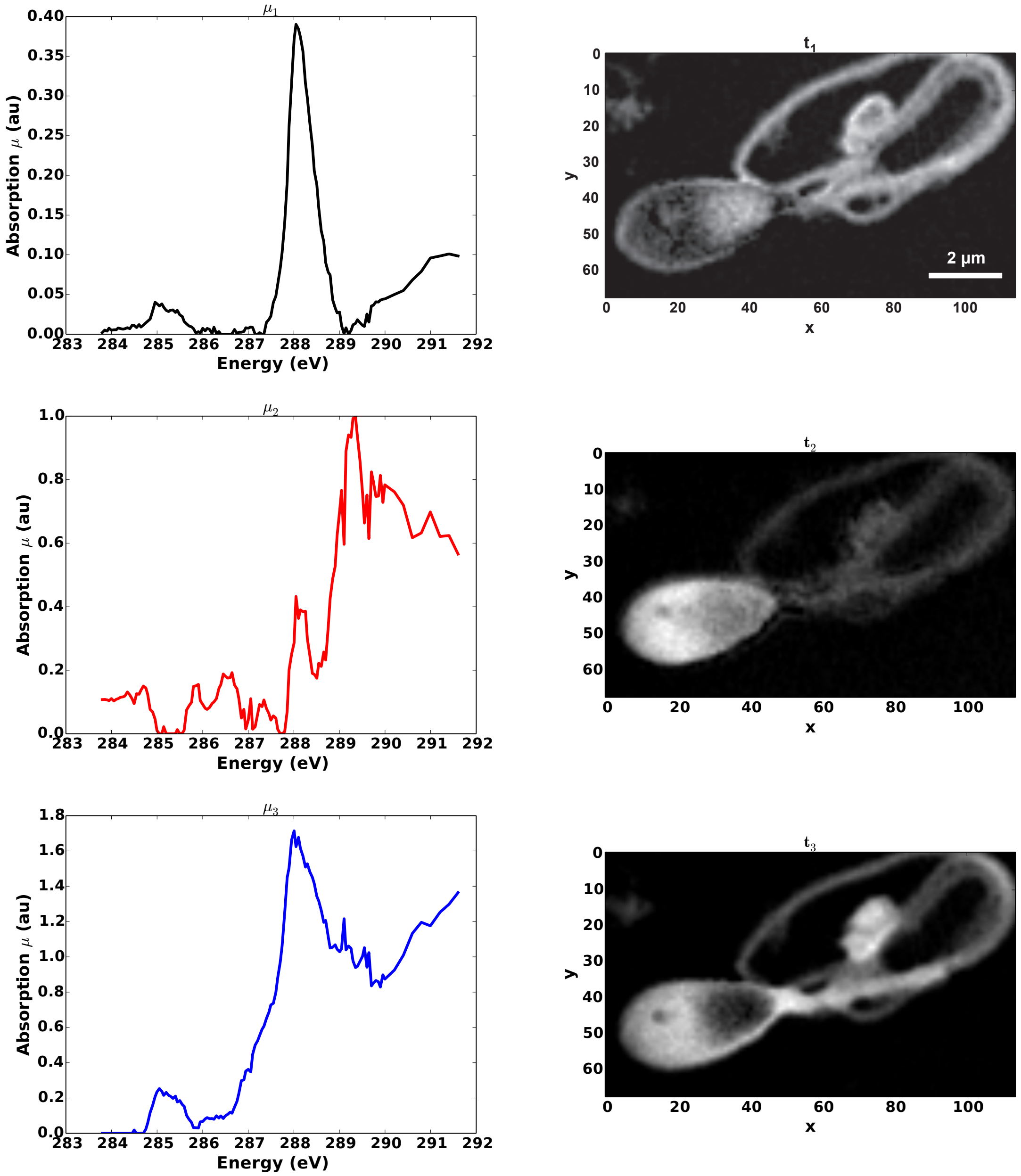
Figure 3.9. NNMA reconstructed spectra and thickness maps for the letters dataset, with $S = 4$. There is no *a priori* determination of which letter will correspond to which thickness map. Since each letter can be found in varying amounts in all of the thickness maps, NNMA has not been successful here in teasing out each feature (letter) and assigning it to a component. Rather, each letter is represented as a mixture of all four components, which could be possible in some cases, but which we know is not the case here since we simulated each letter with just one component.

3.4.4. Experimental dataset: human sperm cell

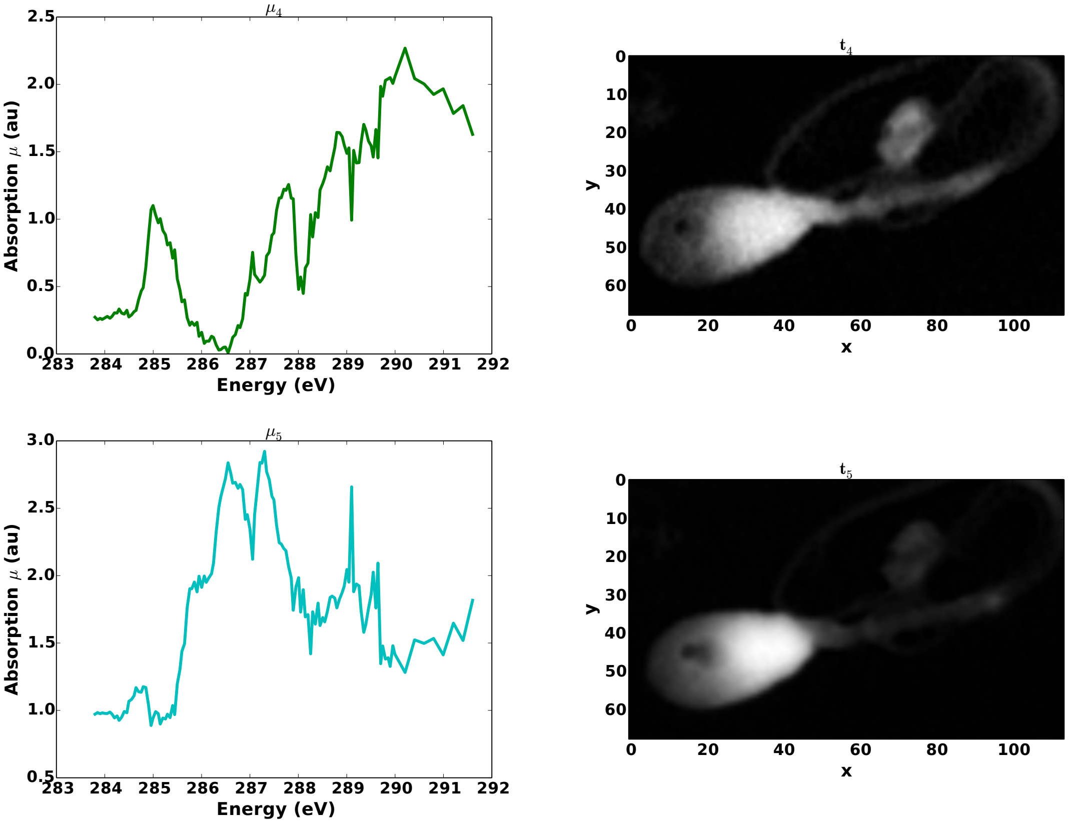
In both the two-wedge prism and letters examples discussed above, the datasets were simulated from a reference set of absorption spectra, and thus we have a set of true spectra and thickness maps with which to compare our NNMA reconstructions. We now turn our attention to an experimental spectromicroscopy dataset of a human sperm cell, taken with a scanning transmission x-ray microscope (STXM) by H. Fleckenstein at beamline X1A of the National Synchrotron Light Source [**50**]. The importance of studying the human sperm cell will be discussed in Section 4.1; here, we focus on the application of NNMA to the dataset. Since we do *not* have a set of reference spectra for an experimental dataset, the challenge becomes whether we could gauge how "good" or "correct" our reconstructions are – if indeed there is such a thing as a correct reconstruction. The goal is to be able to extract meaningful features or components from the cell.

The optical densities of the sperm cell were experimentally measured and span 133 energy points between 283.8 and 291.6 eV – the XANES region surrounding the carbon K-edge at 284.2 eV [**49**]. Each 2D image slice of the spectromicrospy stack is $114 \times 68 = 7752$ pixels, so that $\mathbf{D}$ has dimensions 133×7752. PCA and cluster analysis suggest that the number of components $S \sim 5$. Therefore, we applied the basic NNMA procedure outlined in Figure 3.1, with $S = 5$ and 10,000 iterations. The reconstructed spectra and their corresponding thickness maps are shown in Figure 3.10. For comparison, the cluster analysis spectra and thickness maps are shown in Figure 3.11.

Figure 3.10. NNMA reconstructed absorption spectra and corresponding thickness maps for the sperm dataset, with $S = 5$.



(a) Reconstructions for components $s = 1, 2, 3$ are shown here, while $s = 4, 5$ are shown in the next figure.



(b) *(Continued from Figure 3.10(a).)* Reconstructions for components $s = 4, 5$.

There are several observations to be made from these results. First, the reconstructed spectrum of the first component μ_1 has a smaller signal than the others – its maximum is less than half the magnitude compared to that of the next smallest component μ_2. This could indicate either a background component; or, as its corresponding thickness map suggests, an overall distribution of carbon mass in the sperm cell. Second, the reconstructed spectra are not very smooth; there are some sharp spikes and dips as well as other small but rapid fluctuations that remain even when the algorithm was run over a larger number of iterations, and that are not characteristic of absorption spectra. Experimentally measured spectra do contain fluctuations from detector and other instrument noise, but also involves a monochromator response function that serves to blur the signal and contribute to smoothing

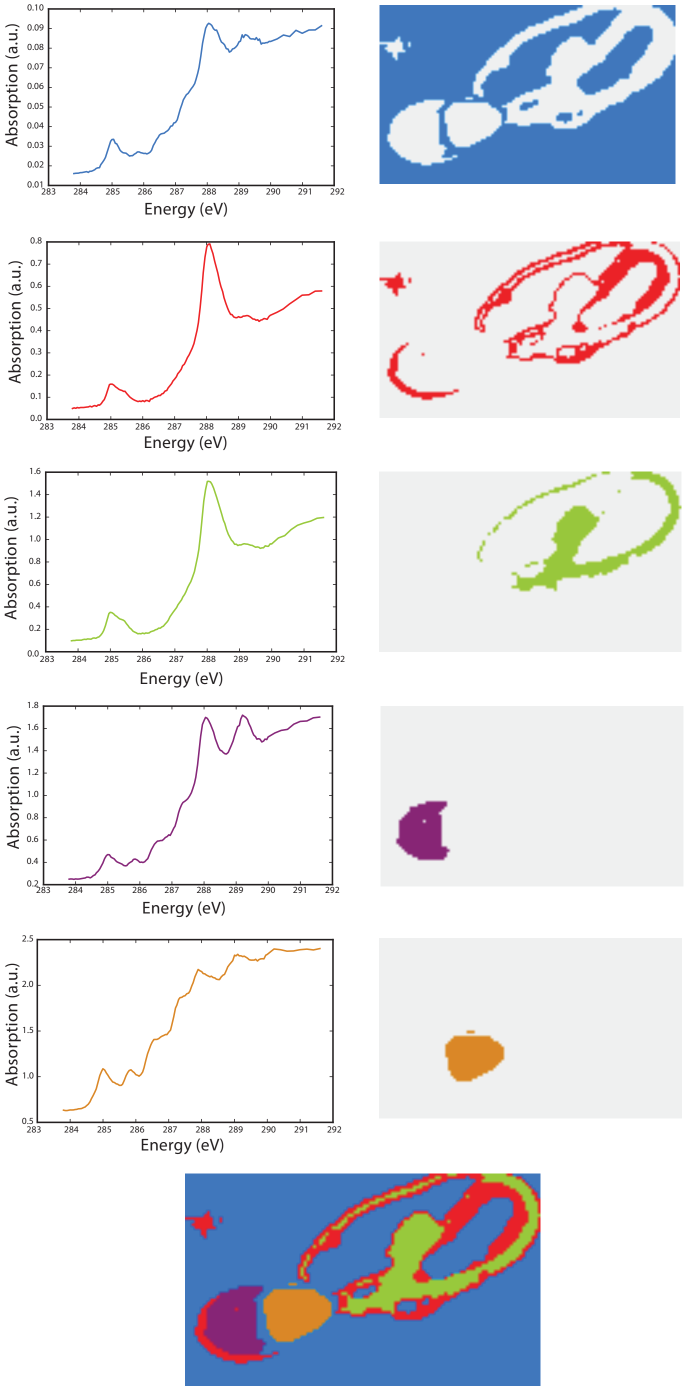


Figure 3.11. As a comparison to the NNMA reconstructed spectra and thickness maps for the sperm dataset shown in Figure 3.10, here are the spectra and component maps generated from cluster analysis with $S = 5$. Each pixel in the image is assigned exclusively to one component spectrum. A composite image showing all of the components is shown at the bottom.

– *i.e.*, if the monochromator has an energy resolution of 0.1 eV or better (as is the case with the STXM used in this experiment), then by sampling at energy intervals finer than this resolution, we would expect a smooth spectral response. Third, the thickness maps do not appear to be very well-differentiated – *e.g.*, there are areas in t_2 and t_3 that overlap, as also in t_4 and t_5.

The first issue involves the problem of choosing an appropriate estimate for S, the number of spectral components, which we will investigate in the next section. The second and third issues will be addressed with the addition of regularization terms, to be discussed in Section 3.6.

3.5. Choosing the number of spectral components

As mentioned in the previous section, performing the NNMA procedure with $S = 5$ on the sperm dataset produces one component that appears weaker than the others. If we had chosen $S = 4$ instead, then we would obtain the spectra and thickness maps shown in Figure 3.12. It appears that the $s = 1$ and 3 components from Figure 3.10(a) have merged to yield the new $s = 3$ component.

To get a more quantitative sense of how the quality of the NNMA reconstructions depends on S, we could look at how the cost function F varies as a function of S, since F is a measure of how closely the reconstructed optical density matrix $\mathbf{D}$ matches that from experiment. For the sperm dataset, we set S between 2 and 9, and applied the NNMA procedure as before. The plot of F versus S is shown in Figure 3.13. There is a large (55%) decrease in the cost function from $S = 2$ to 3, and a smaller but still significant (31%) decrease from $S = 3$ to 4. From $S = 4$ to 5, the decrease is only 6%, and continues to get smaller with increasing S. Of course, we expect the cost function to decrease as S increases. In fact, in the limit $S \to P$,

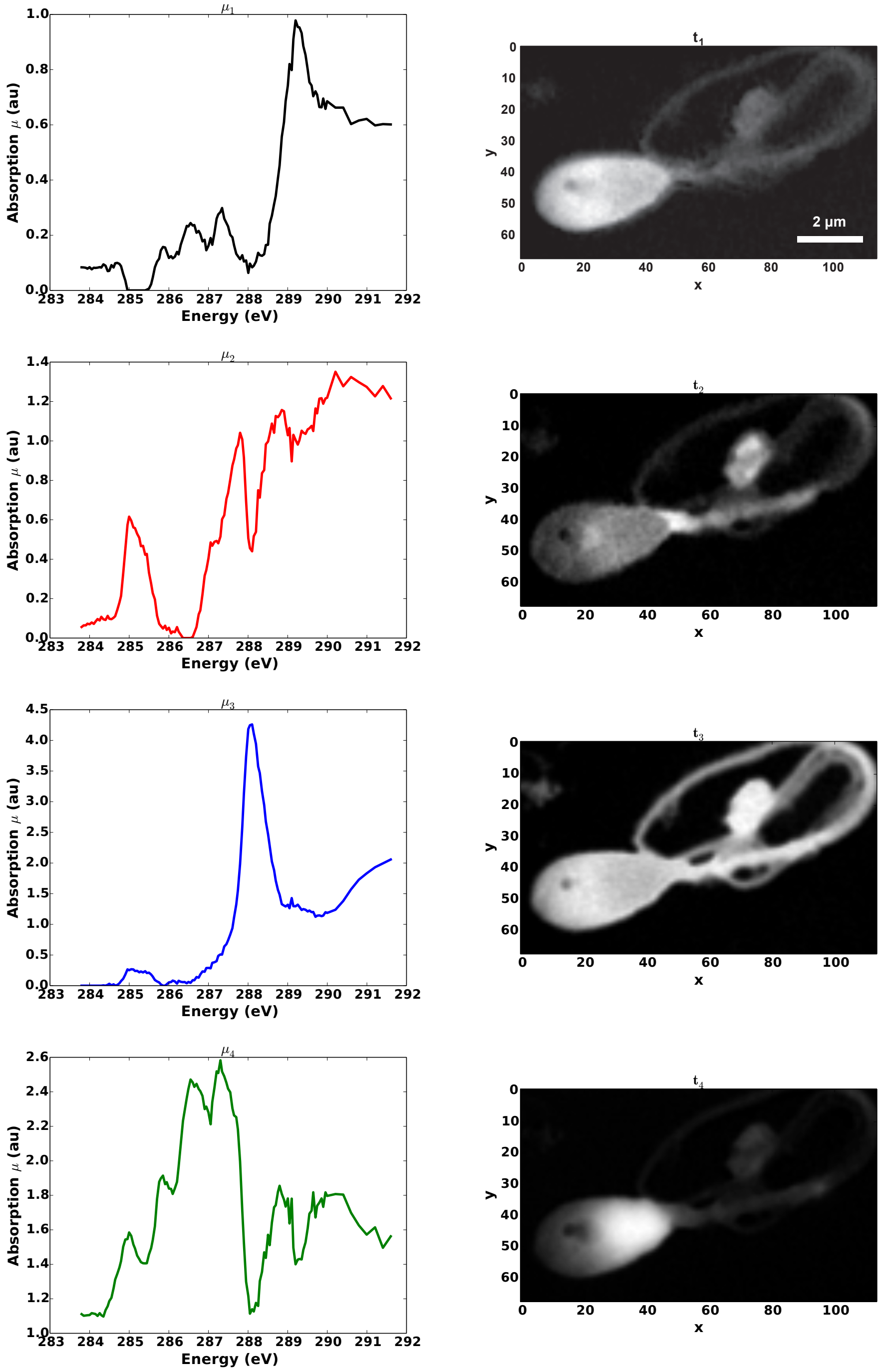


Figure 3.12. NMNA reconstructed absorption spectra and corresponding thickness maps for the sperm dataset, with $S = 4$. Comparing this to Figure 3.10(a) where $S = 5$ was used, it appears that the $s = 1$ and 3 components from that figure have merged to yield the $s = 3$ component in this figure.

we would expect a near-perfect match between reconstruction and data ($F \to 0$), since we could simply assign each pixel its own component. However, this would overfit the data and does not tell us anything useful. The purpose of using NNMA (as well as PCA and cluster analysis) is to perform dimensionality reduction in order to identify meaningful features in the data. From this cost function analysis, we estimate the number of components to be 4 or 5, consistent with the number suggested by PCA. There is no definitive method to decide the exact number for S, and we do not have the true set of spectra in an experiment. Another interesting method for estimating S, which we have not used in this work, involves a measure based on the cophenetic correlation coefficient ρ_S [**36**]. It evaluates the stability of assigning pixels to S clusters over several runs. As the clustering becomes more stable (pixels do not change clusters between different runs), ρ_S becomes smaller. S is chosen where ρ_S begins to decrease significantly.

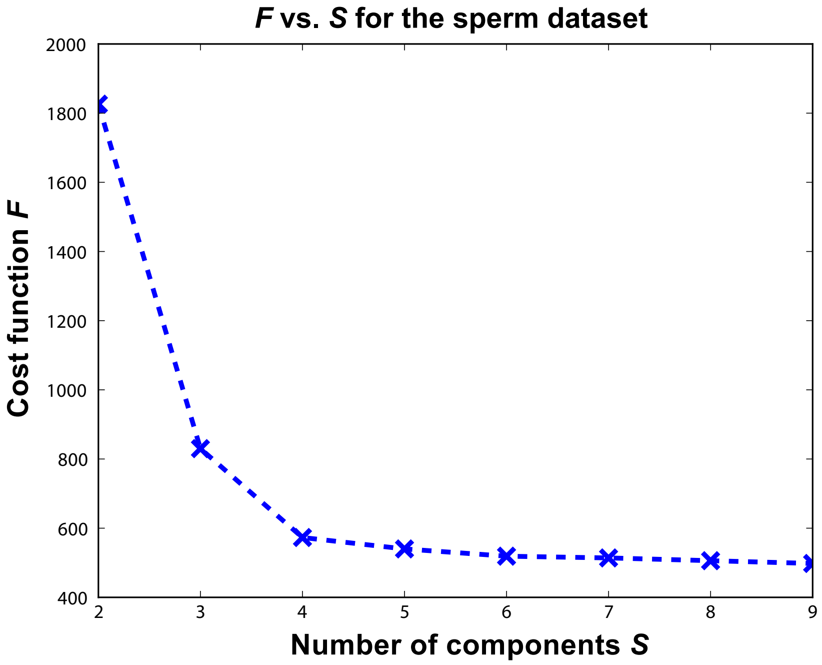


Figure 3.13. Cost function F versus number of components S for the sperm dataset. There is a large decrease in the cost function from $S = 2$ to 3, and a smaller but still significant decrease from $S = 3$ to 4. However, from $S = 4$ onwards, the decreases become less significant, and we estimate S to be 4 or 5. We would then perform NNMA at these two different values and see how the results change.

3.6. Addition of regularizations

The basic NNMA procedure works relatively quickly and is able to find a consistent minimization (at least locally) for the cost function F. While the results for the simulated artificial dataset are good, the reconstructed absorption spectra for the sperm dataset are not what we would expect from realistic physical spectra. This is because the NNMA algorithm has no underlying expectation of what the reconstructed results should look like: from the NNMA perspective, it does not matter if the spectra are not smooth, or if the thickness maps contain many components that overlap one another spatially – as long as F is minimized.

However, our goal is to be able to separate the components within the sample as cleanly as possible, while generating realistic, smooth absorption spectra. In addition, NNMA does not guarantee a unique solution to the factorization problem: given a solution pair ($\boldsymbol{\mu}$, $\mathbf{t}$) which minimizes Eq. (3.2), it is possible to find an invertible matrix $\mathbf{X}_{S\times S}$ such that we could achieve the same minimization with the solution pair ($\boldsymbol{\mu}\mathbf{X}$, $\mathbf{X}^{-1}\mathbf{t}$).

To solve these problems, we can add regularizations to the cost function in Eq. (3.2) to incorporate any prior information that we might have about the data, or certain characteristics that we might expect from the results. The addition of regularization terms modifies the cost function to be minimized, so that it becomes:

$$F(\boldsymbol{\mu}, \mathbf{t}) = \frac{1}{2}||\mathbf{D} - \boldsymbol{\mu}\mathbf{t}||_2^2 + \lambda_1 J_1(\boldsymbol{\mu}, \mathbf{t}) \ + \ ... \ + \ \lambda_n J_n(\boldsymbol{\mu}, \mathbf{t}), \tag{3.9}$$

where n is the number of regularization schemes we wish to incorporate; J_i is the regularizer that is a function of $\boldsymbol{\mu}$ and/or $\mathbf{t}$; and $\lambda_i > 0$ is its corresponding continuous regularization parameter.

While hard constraints strictly exclude infeasible solutions, regularizations are softer and allow for adjustments – by tuning λ – on how much penalty we want to assign when the reconstruction does not fit the model well. Essentially, λ represents a trade-off between finding reconstruction solutions that minimize the errors in data matching (that is, the basic cost function in Eq. (3.2)), and solutions that fit our expected model. If each λ_i is small compared to one, then most of the weight falls onto the first term of Eq. (3.9), and it pays to minimize the error in data matching at the cost of not following the data model. As λ_i increases, then more weight falls onto the i^{th} regularization term, and so we place more emphasis on extracting desired characteristics of the data model at the cost of larger errors in data matching. Regularizations also serve to narrow the search space for solutions to the minimization problem, and thus reduce the non-uniqueness of the solutions.

3.6.1. Sparseness regularization

One of the characteristics we look for in the thickness maps is features consisting of well-separated components. To aid in this objective, we make two observations. First, many biological specimens contain distinct, localized regions or structures (such as the nucleus, mitochondria, etc.) with unique absorption signatures. Since these structures are localized, they appear in only a small (relative to the total) number of pixels. Second, each type of structure is composed of components that are distinct from those of other types (since the combination is what gives each type its unique "chemical fingerprint" – it is what distinguishes a nucleus from a cell wall, for example), and so their absorption signatures should typically be a combination of only a few components.

The upshot of these observations is that we would expect the **t** matrix to be sparse: if each component is concentrated only within a small region or number of pixels (structure localization), then we can expect each row of $\mathbf{t}$ (representing each component) to contain mostly zeros – that is, $\mathbf{t}$ is row-sparse. In addition, if each pixel contains only a very small number of pure components (components in identified structures should be as unmixed as possible), then we can expect each column of $\mathbf{t}$ to contain more zeros than not – that is, $\mathbf{t}$ is column-sparse.

The sparseness of $\mathbf{t}$ is thus related to the degree of mixing between different components in the sample: the more sparse $\mathbf{t}$ is – that is, the more zero or near-zero entries $\mathbf{t}$ contains – the greater the degree of spectral separation in $\boldsymbol{\mu}$, which is what we seek. Consider the extreme case of a solution for $\mathbf{t}$ in which, for each of the P columns, at most one of the S entries is non-zero. This would mean that the sample has a single distinct spectrum at each pixel in the measured data (or just background signal in the case of an all-zeros column), with no spectral mixing, giving a very sparse solution for $\mathbf{t}$. Such a highly sparse solution exactly like the one just described would be highly unlikely in x-ray spectromicroscopy data of anything other than a completely phase-segregated sample, yet it is still desirable to seek solutions with some degree of sparseness in $\mathbf{t}$, since they offer simpler and often more insightful interpretations of the data.

Sparseness regularizations and constraints have proven useful in a variety of applications, including text mining [**51, 52**], compressed sensing in signal processing [**53, 54**], the study of neural networks [**55**], and gene expression in bioinformatics [**56**].

Since the sparsest solution for $\mathbf{t}$ means that it should contain as many zero entries as possible, this would be achieved by minimizing its "ℓ_0 norm" [1], $||\mathbf{t}||_0$, which counts the

[1] The ℓ_0 norm is not technically a norm, since it does not satisfy the scaling property of a true norm – *i.e.*, $||\alpha\mathbf{t}||_0 \neq \alpha||\mathbf{t}||_0$.

number of non-zero entries in $\mathbf{t}$:

$$||\mathbf{t}||_0 = \#\{(i,j)\,|\,t_{ij} \neq 0\} \tag{3.10}$$

However, the minimization of the ℓ_0 norm is computationally expensive to solve (see Appendix A.2). As an approximation to the ℓ_0 norm, one popular way to achieve sparse solutions is to regularize $||\mathbf{t}||_1$, the ℓ_1 norm of $\mathbf{t}$, also known as the lasso technique [**57, 58**] (also see Appendix A.3). The addition to the basic cost function in Eq. (3.2) is then:

$$\begin{aligned} \lambda_{\mathbf{t}} J_{\mathbf{t}}(\mathbf{t}) &= \lambda_{\mathbf{t}} ||\mathbf{t}||_1 \\ &= \lambda_{\mathbf{t}} \sum_{p=1}^{P} \sum_{k=1}^{S} t_{kp} \end{aligned} \tag{3.11}$$

where each element $t_{kp} \geq 0$ by constraint. The sparseness regularization parameter $\lambda_{\mathbf{t}}$ can be tuned according to prior information we might have about the chemical structure of the sample. The addition to the update rule for $\mathbf{t}$ is:

$$\lambda_{\mathbf{t}} \frac{\partial J_{\mathbf{t}}(\mathbf{t})}{\partial \mathbf{t}} = \lambda_{\mathbf{t}}\, \mathbf{ones}_{S\times P} \tag{3.12}$$

(where $\mathbf{ones}_{S\times P}$ is an $S \times P$ matrix with all entries equal to one), so that the $\mathbf{t}$ update rule in Eq. (3.7) becomes modified to

$$\mathbf{t} \leftarrow \mathbf{t}\, \frac{(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D})}{(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t} + \lambda_{\mathbf{t}})}. \tag{3.13}$$

There are some limitations associated with the lasso. For example, when $S > P$, at most P components or features can be selected before saturation [**59**]. Although this is not usually the case in spectromicroscopy datasets (where the number of pixels is much greater than the number of components, or $S \ll P$), this problem commonly arises in applications such as microarray data analysis, where there can be many thousands of genes expressed in fewer than 100 samples. In these cases, other sparseness regularization schemes have been proposed [**60**].

Sparseness regularization on the letters dataset. We apply NNMA with this new update to the letters dataset, setting $\lambda_{\mathbf{t}} = 0.8$. From Figure 3.14, we can see from the thickness maps that sparseness regularization does yield a sparser $\mathbf{t}$ (compared to Figure 3.9) – most of the pixels in the image are identified with just one of the four components. The only exceptions are those pixels associated with the letter 'B', which contain a mixture of mainly μ_2 with some contribution from μ_3. The reconstructed spectra are similar to the reference spectra in Figure 3.2, again with the exception of μ_2. For comparison, Figure 3.15 show NNMA reconstructions with the same parameters, except this time with $\lambda_{\mathbf{t}} = 0.1$. Since a lower penalty is being placed on the sparseness regularization term, more of the features (letters) begin to contain mixtures of two or more components. On the other end of the scale, placing too high of a penalty on the sparseness regularization term can also drastically affect the quality of the reconstructions, as shown in Figure 3.16. A large $\lambda_{\mathbf{t}}$ can dominate the denominator in the $\mathbf{t}$ update rule in Eq. (3.13), and if it is also large compared to the numerator $\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D}$, the $\mathbf{t}$ would not change much between iterations. Indeed, for $\lambda_{\mathbf{t}} = 2.5$, we find that $\Delta F < -10^{-7}$ after fewer than 50 iterations and the algorithm is stopped.

A more systematic investigation on how we might use the behavior of the cost function F as a function of $\lambda_{\mathbf{t}}$ to choose appropriate regularization parameter values will be presented in Section 3.6.3.

We are not sure why the component associated with 'B' seems to be more difficult to reconstruct correctly; it might have something to do with the fact that the original μ_2 used

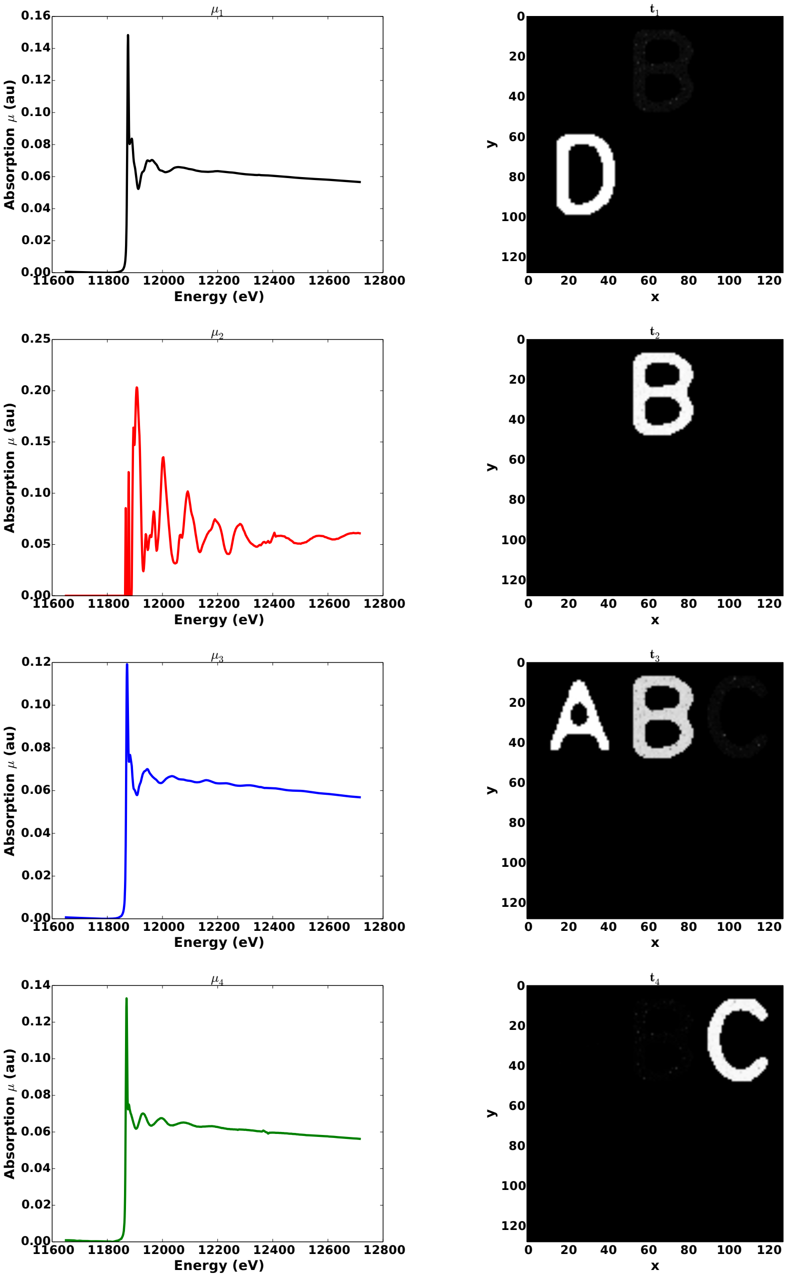


Figure 3.14. NNMA reconstructed spectra and thickness maps for the letters dataset, with sparseness regularization. Parameters are $\lambda_{\mathbf{t}} = 0.8$, $S = 4$, and number of iterations = 1000. The thickness maps show that $\mathbf{t}$ has been made very sparse, with most of the pixels assigned to just one component. The only exceptions are those associated with the letter 'B', which appear to contain mostly μ_2 with some mixture of μ_3. All of the spectra are comparable to the reference spectra, again with the exception of μ_2. We will address this problem in Section 3.6.2.

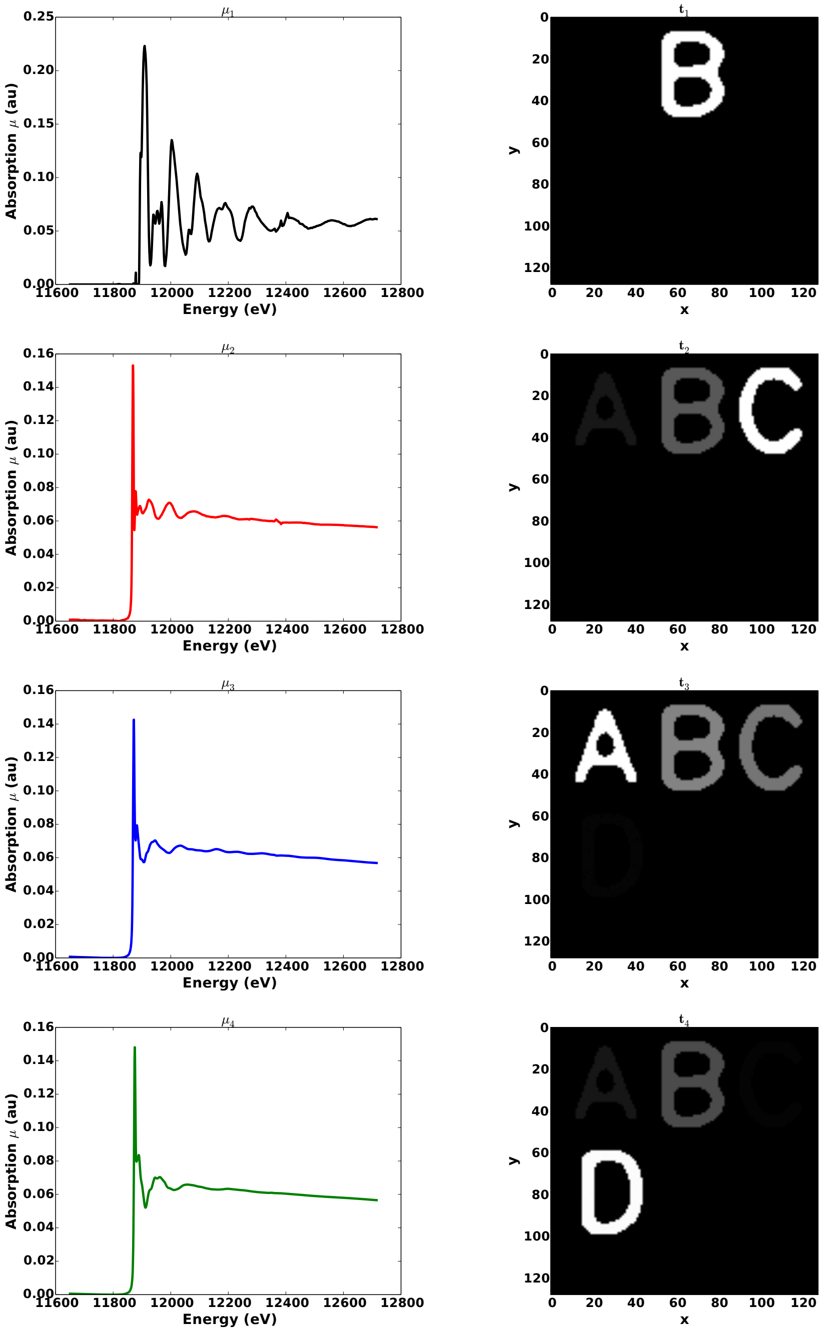


Figure 3.15. NNMA reconstructed spectra and thickness maps with the same set of parameters as for Figure 3.14, except $\lambda_{\mathbf{t}} = 0.1$. Since a lower penalty is being placed on the sparseness regularization term, the thickness maps accordingly show more mixtures than those with a higher penalty imposed. We observe again that the reconstructed spectrum for ‘B’ deviates the most from the reference.

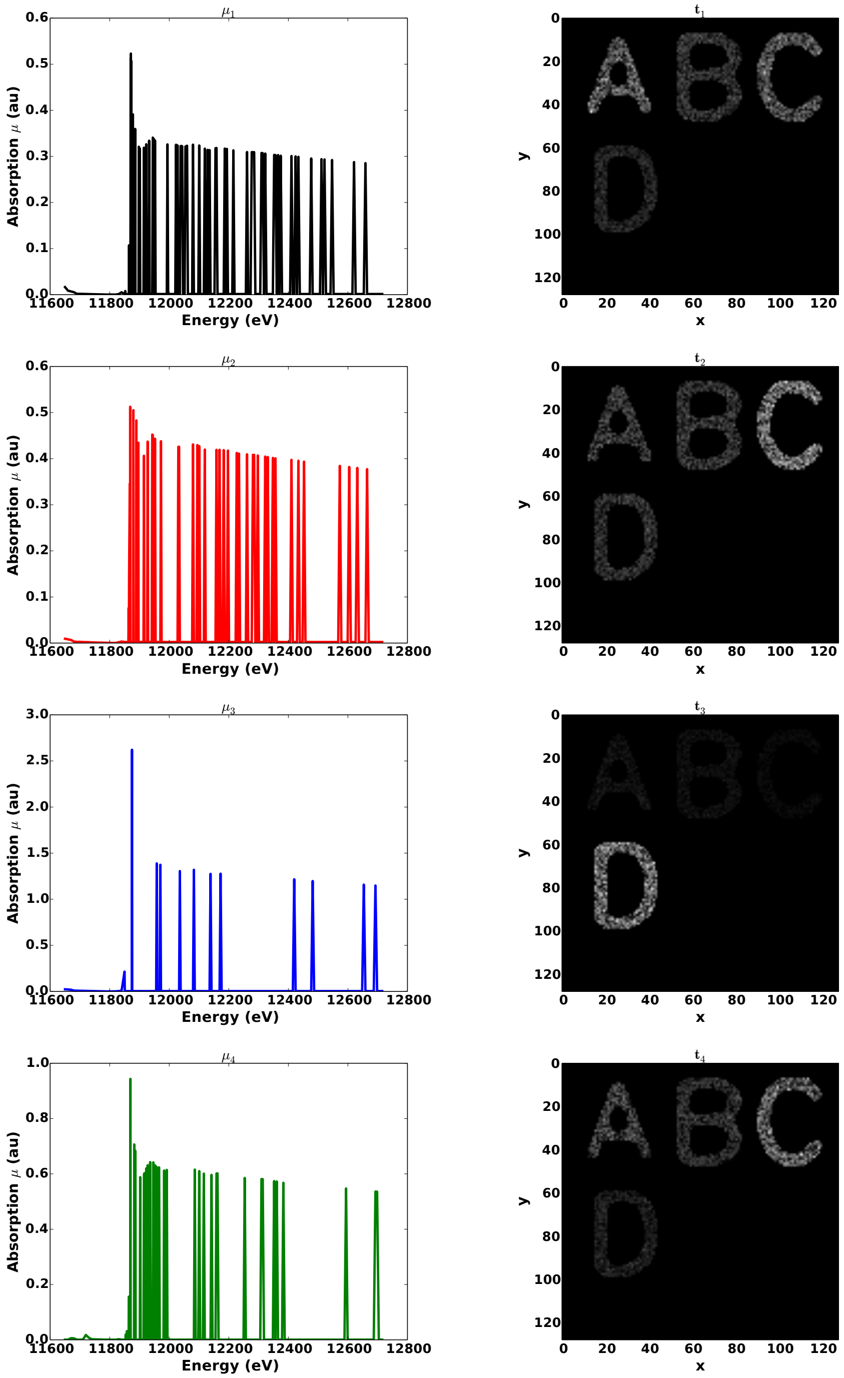
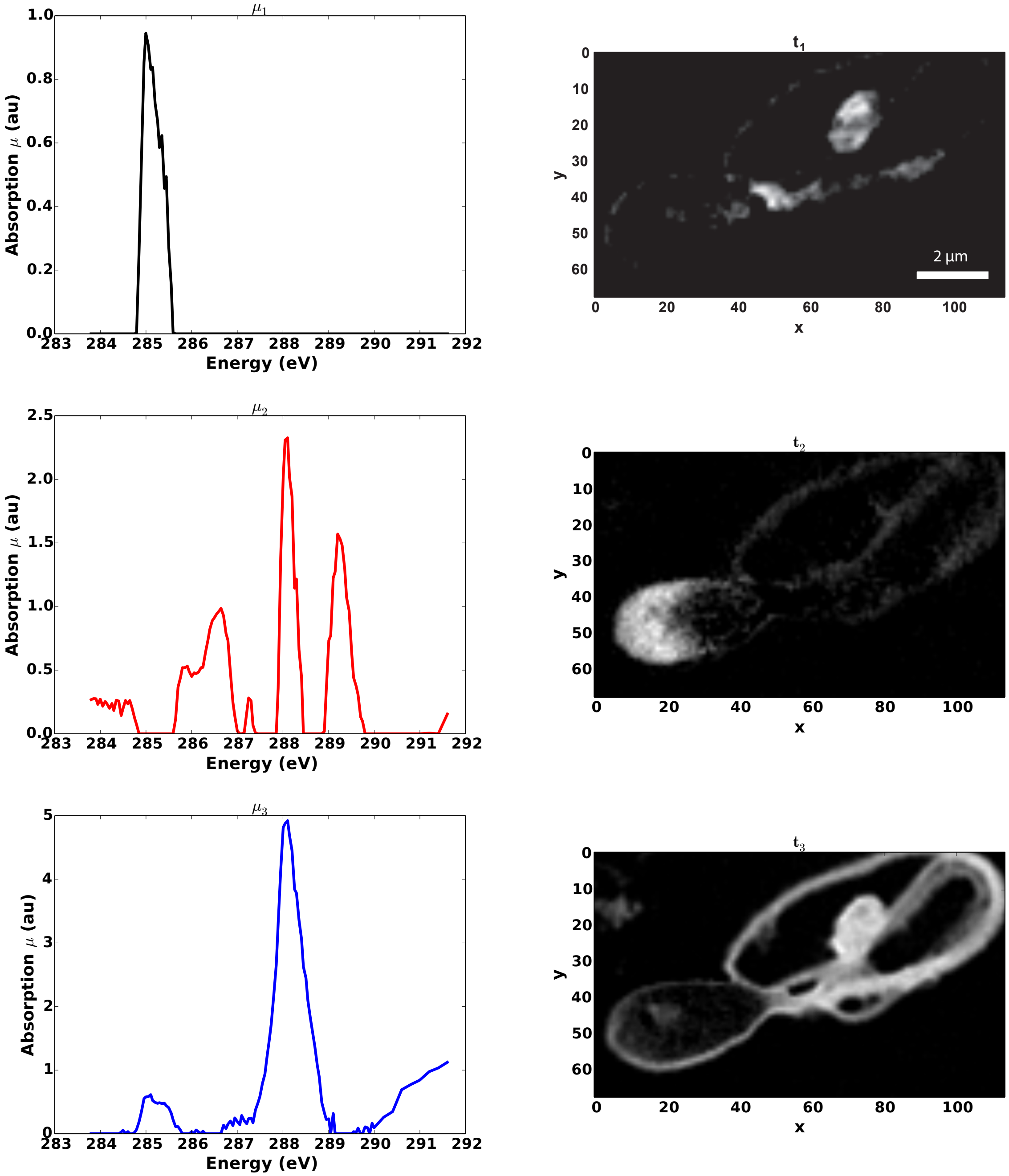


Figure 3.16. Setting the sparseness regularization parameter $\lambda_{\mathbf{t}}$ too high can drastically reduce the quality of the reconstructions. Here, we have set $\lambda_{\mathbf{t}} = 2.5$ while keeping the other parameters the same as for Figures 3.14 and 3.15.

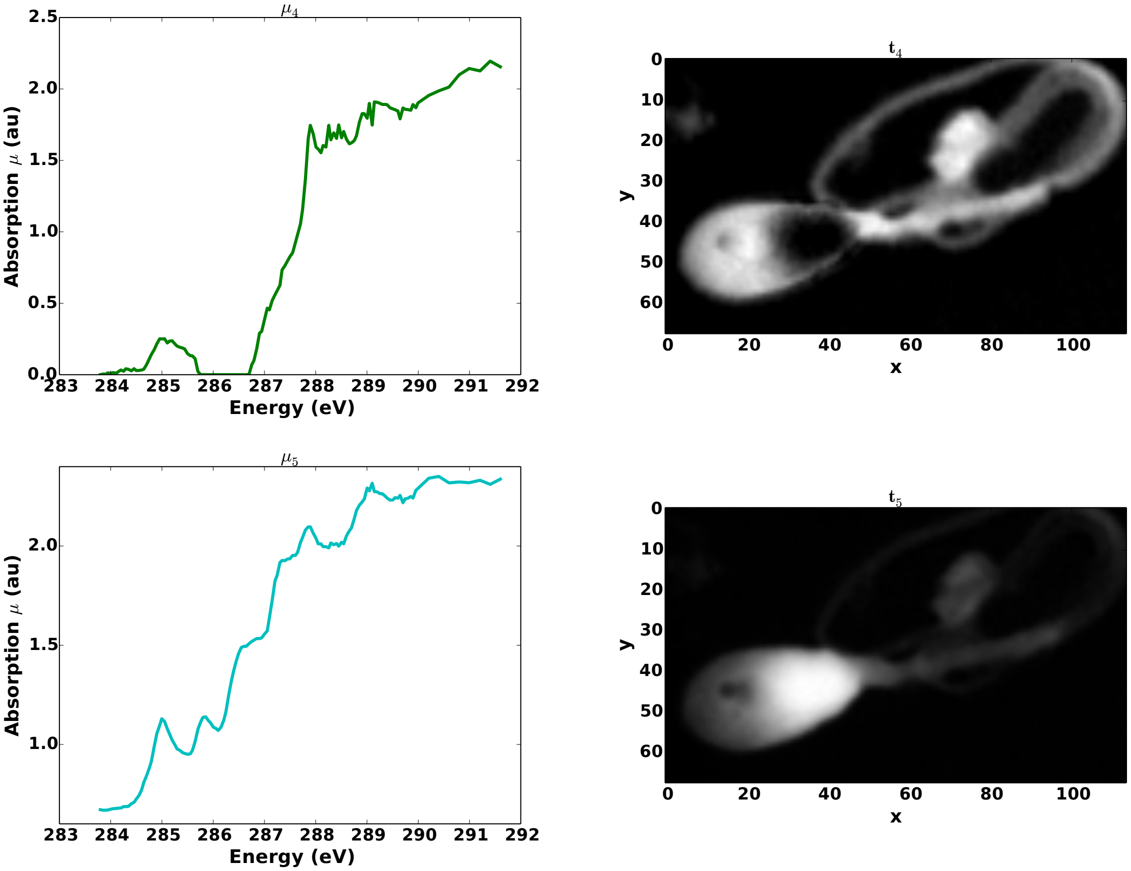
to simulate ‘B’ (the GaAs spectrum in Figure 3.2) has a smaller absorption resonance signal than the other three spectra, thus making its optical density weaker in the simulated data. (Strangely, its reconstructed spectrum happens to have the largest resonance magnitude.) The problem of obtaining better spectra reconstructions from NNMA will be discussed in the next section.

Sparseness regularization on the sperm dataset. We also apply NNMA with sparseness regularization to the sperm dataset. Figure 3.17 (with $\lambda_{\mathbf{t}} = 0.2$) show improved feature selection in the thickness maps compared to Figure 3.10 (no regularization). Again, spectra improvement will be discussed in the next section. As was seen in the letters dataset, setting too high a penalty for sparseness can lead to worse feature selection (Figure 3.18) if the sparseness of the data does not match.

Figure 3.17. NNMA spectra and thickness maps reconstructions for the sperm dataset, with sparseness regularization. Parameters are $\lambda_{\mathbf{t}} = 0.2$, $S = 5$, number of iterations = 10,000.

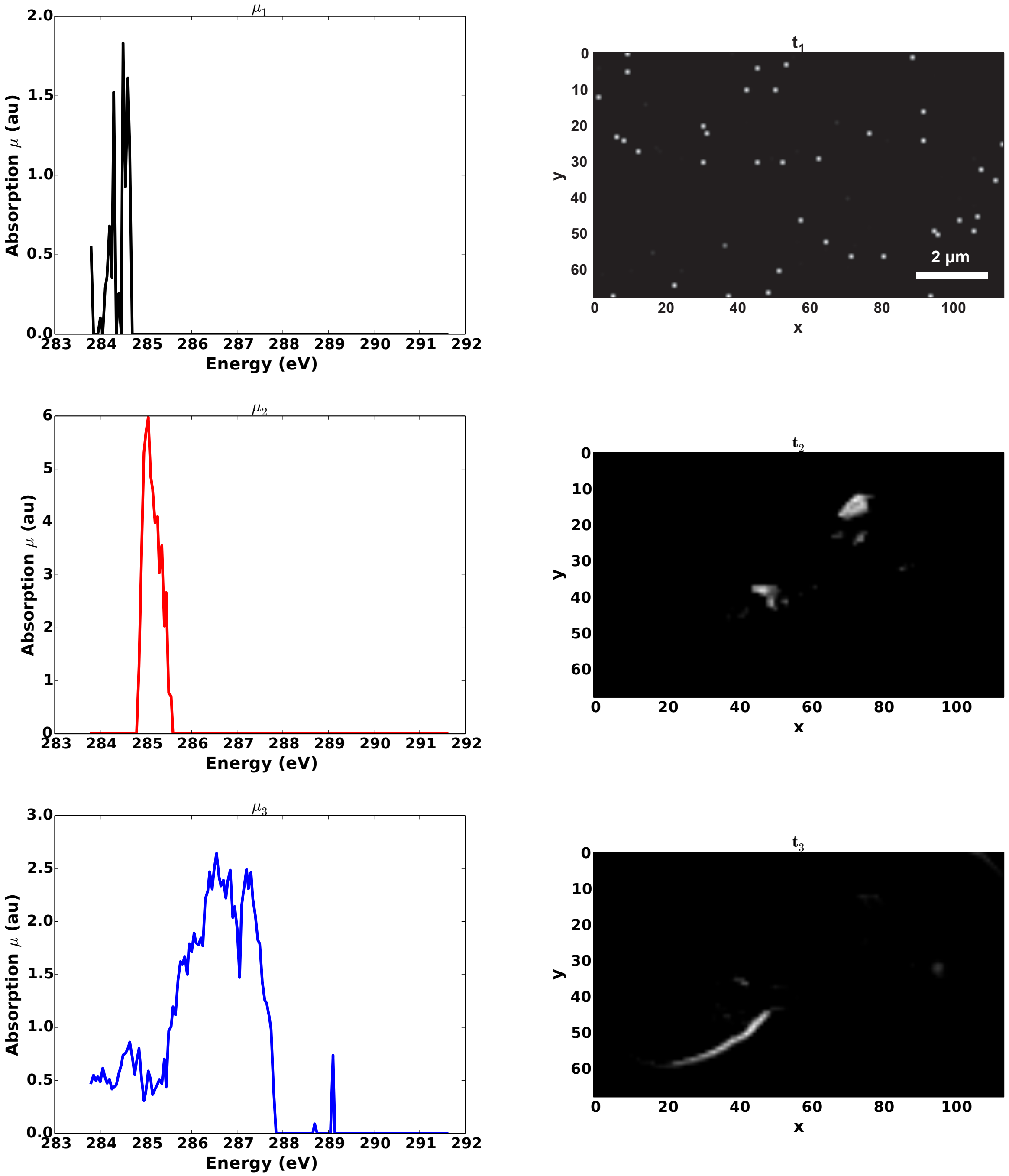


(a) Reconstructions for components $s = 1, 2, 3$ are shown here, while $s = 4, 5$ are shown in the next figure.

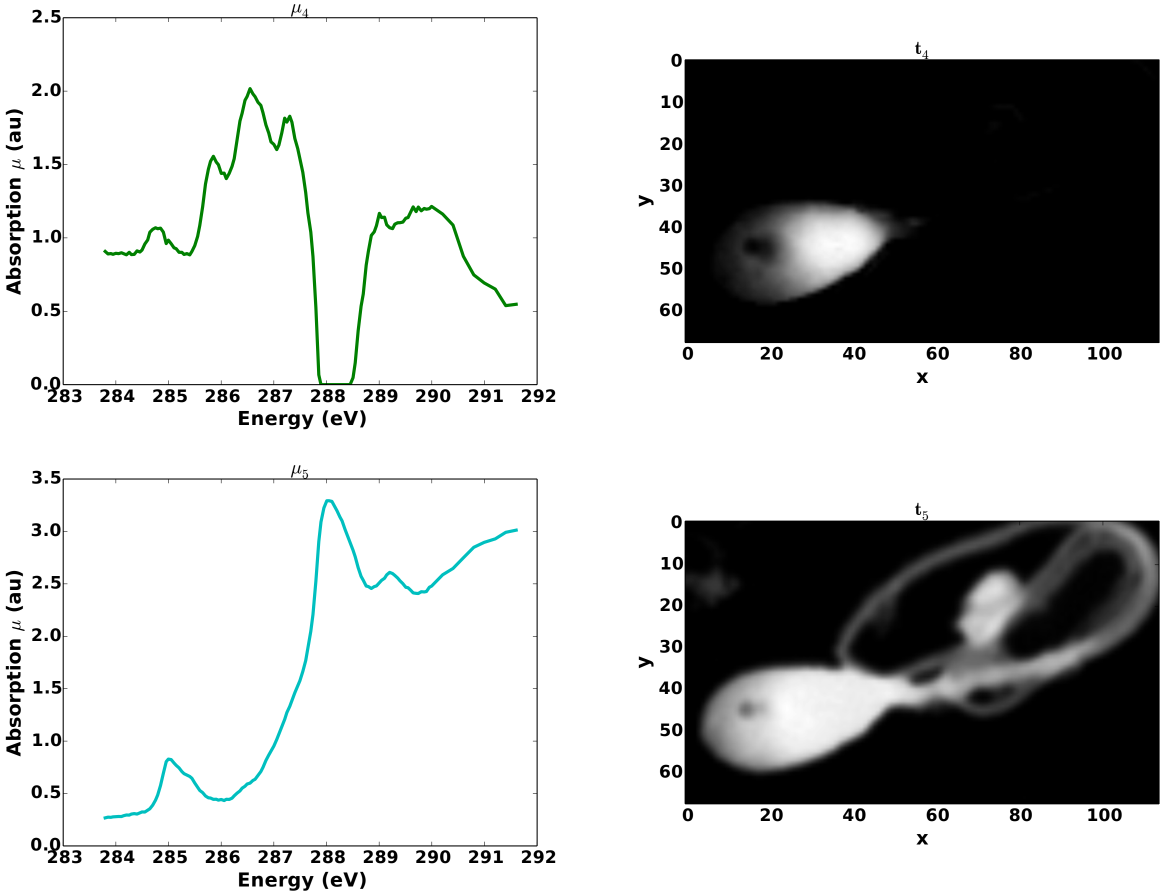


(b) *(Continued from Figure 3.17(a).)* Reconstructions for components $s = 4, 5$.

Figure 3.18. Not all data are modelled well by high sparseness. Setting too high a penalty for sparseness can lead to bad feature selection (here, $\lambda_{\mathbf{t}} = 5.0$, with other parameters being the same as those used in Figure 3.17.).



(a) Reconstructions for components $s = 1, 2, 3$ are shown here, while $s = 4, 5$ are shown in the next figure.



(b) *(Continued from Figure 3.18(a).)* Reconstructions for components $s = 4, 5$.

3.6.2. Spectra similarity regularization

We saw in Sections 3.4.3 and 3.4.4 that NNMA on its own does not always yield a set of smooth absorption spectra $\boldsymbol{\mu}$. On the other hand, we saw in Section 2.3 that cluster analysis could produce a relatively good set of absorption spectra $\boldsymbol{\mu}_{\text{cluster}}$, even though there might be negative values in the corresponding thickness maps. One idea then is to combine the advantages of each approach in the hope that this will lead to an improved set of spectra. We could do this by trying to make the NNMA spectra similar to those from cluster analysis. To achieve this, we minimize the ℓ_2 norm of the difference between the two sets of spectra by introducing a "spectra similarity" regularization term to the cost function F, given by

$$\lambda_{\boldsymbol{\mu}_{\text{sim}}} J_{\boldsymbol{\mu}_{\text{sim}}}(\boldsymbol{\mu}) = \lambda_{\boldsymbol{\mu}_{\text{sim}}} ||\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}}||_2^2 \tag{3.14}$$

$$= \lambda_{\boldsymbol{\mu}_{\text{sim}}} \sum_{k=1}^{S} \sum_{n=1}^{N} (\boldsymbol{\mu}_{n,k} - \boldsymbol{\mu}_{\text{cluster}\,n,k})^2, \tag{3.15}$$

where $J_{\boldsymbol{\mu}_{\text{sim}}}$ is the spectra similarity regularizer, and $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ is the regularization parameter for tuning how similar we desire the two sets of spectra to be (the higher $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ is, the more similar they will be, at the cost of worse data matching).

The addition to the update rule for $\boldsymbol{\mu}$ is

$$\lambda_{\boldsymbol{\mu}_{\text{sim}}} \frac{\partial J_{\boldsymbol{\mu}_{\text{sim}}}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \lambda_{\boldsymbol{\mu}_{\text{sim}}} 2(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}}), \tag{3.16}$$

so that the $\boldsymbol{\mu}$ update rule in Eq. (3.8) becomes modified to

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} \, \frac{\left(\mathbf{D}\mathbf{t}^{\mathrm{T}}\right)}{\left(\boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}} + 2\lambda_{\boldsymbol{\mu}_{\text{sim}}}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}})\right)} \,. \tag{3.17}$$

Spectra similarity regularization on the letters dataset. We apply NNMA with this new update rule to the letters dataset, setting $\lambda_{\boldsymbol{\mu}_{\text{sim}}} = 10.0$ and $\lambda_{\mathbf{t}} = 0$ (without sparseness regularization). The resulting NNMA spectra and thickness maps are shown in Figure 3.19, from which we can see that the spectra are improved from those in Figures 3.9 and 3.14, in that they more closely resemble the reference spectra. We also observe from the thickness maps that each letter is associated mainly with just one component (very little mixtures), and thus that the reconstructed $\mathbf{t}$ is sparse, even though we did not regularize sparseness in this case.

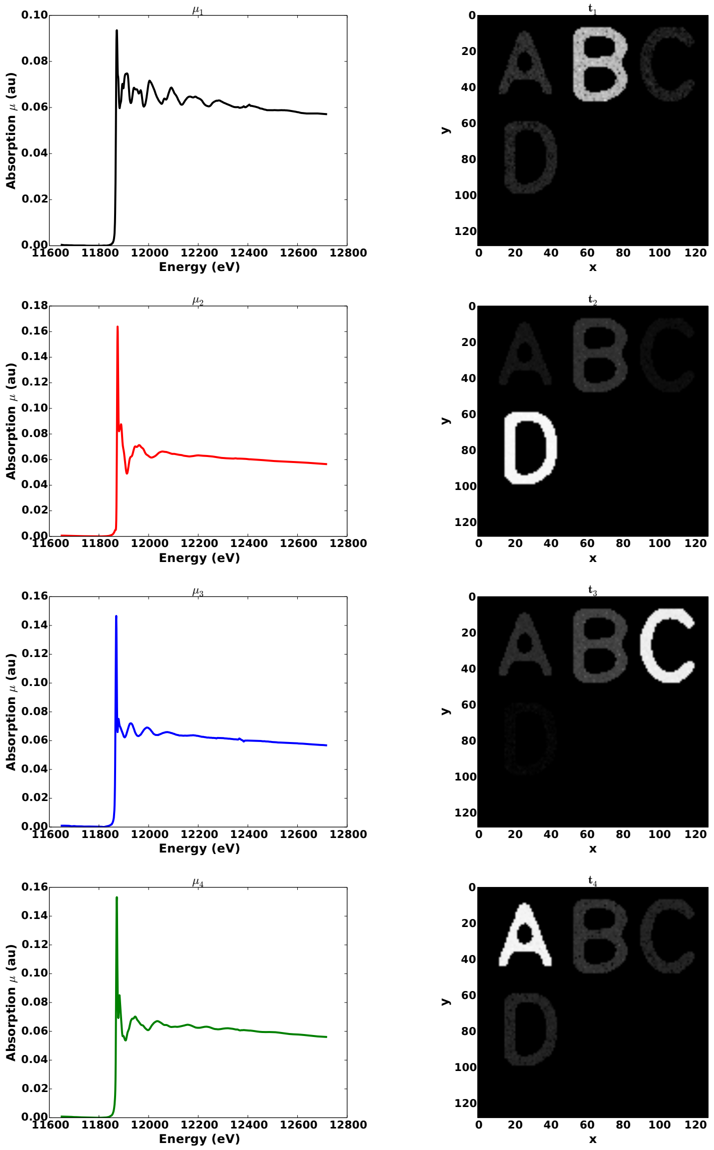


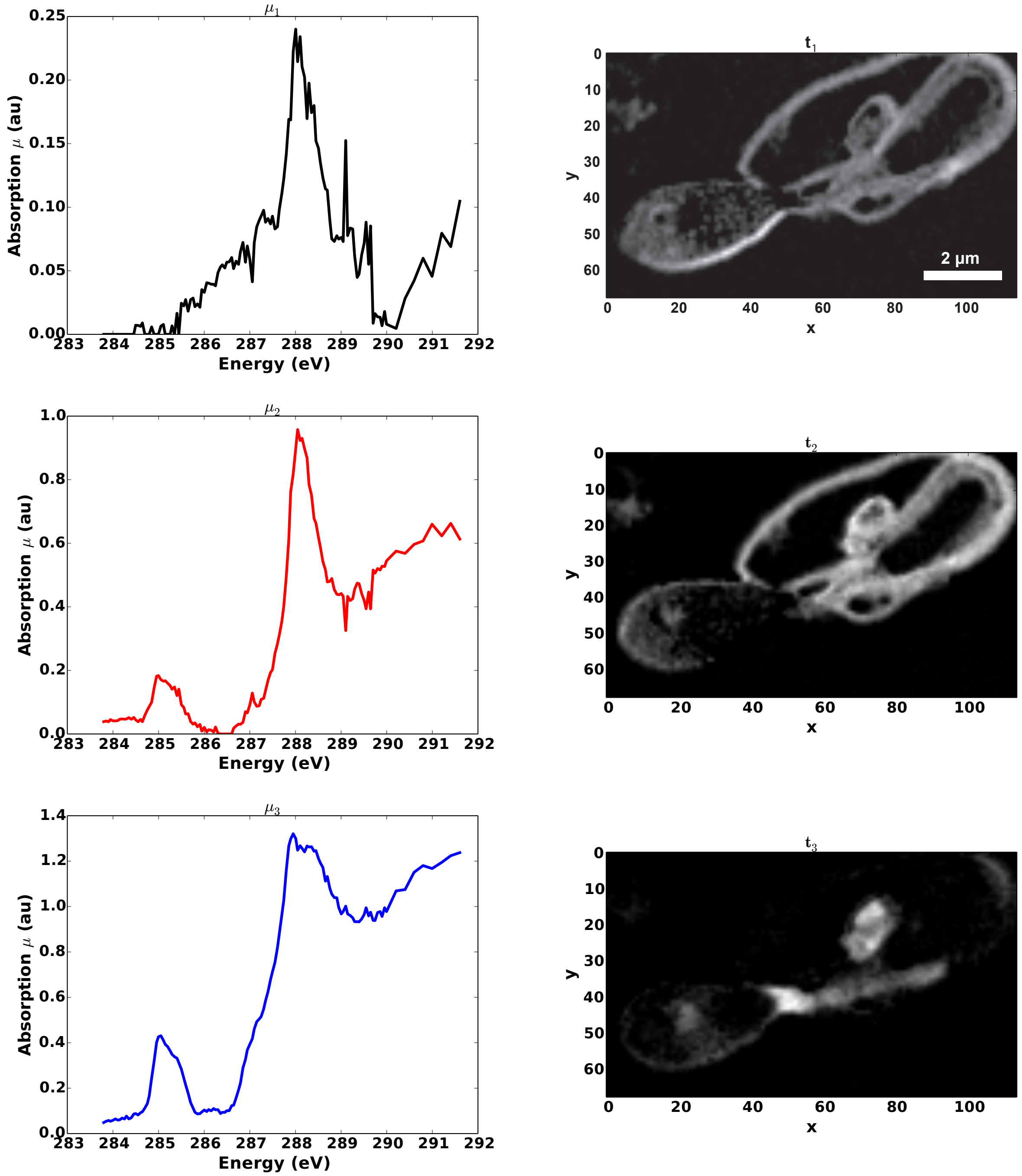
Figure 3.19. NNMA reconstructed spectra and thickness maps for the letters dataset, with cluster spectra similarity regularization. Parameters are $\lambda_{\boldsymbol{\mu}_{\text{sim}}} = 10.0$, $\lambda_{\mathbf{t}} = 0$, and $S = 4$, and number of iterations = 1000. The spectra are more similar to the reference (and cluster) spectra compared to those in Figures 3.9 and 3.14. Although we did not regularize sparseness, it can be seen from the thickness maps that $\mathbf{t}$ is mostly sparse – each pixel contains just one predominant component (though there are still some small mixtures in, for example, the pixels of letter ‘B’, which appear in faint amounts in all of the maps in addition to t_1). As discussed in the text, this is not surprising since cluster analysis inherently selects for maximum sparseness.

This is not surprising, because – in a way – clustering can be thought of as an extreme example of NNMA with maximum sparseness: in cluster analysis, each pixel is assigned exclusively to one cluster – a "winner-takes-all" approach. When coaxing the NNMA spectra to be similar to cluster spectra (by increasing $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$), we would expect the NNMA thickness maps to begin to resemble cluster component maps, as is the case in Figure 3.19.

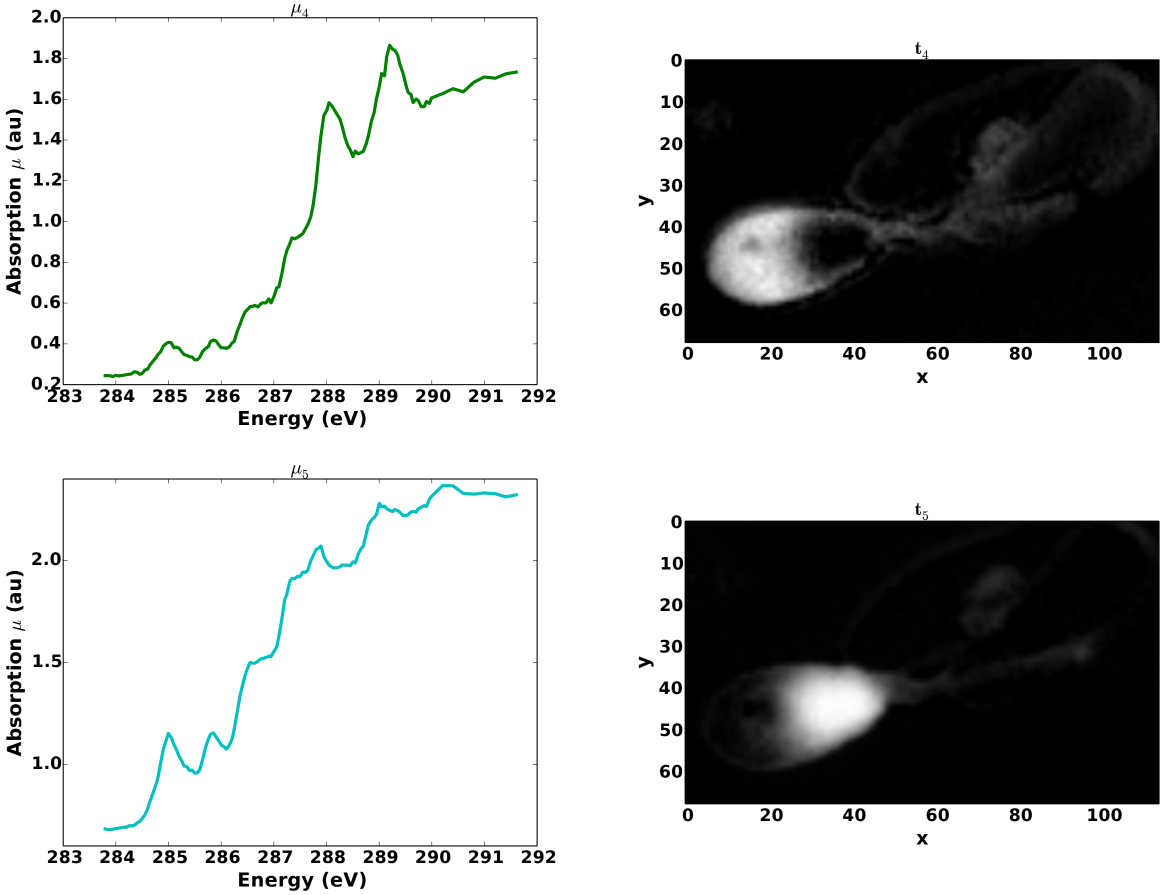
In NNMA, the sparseness regularization parameter $\lambda_{\mathbf{t}}$ can be tuned to control how many clusters a particular pixel can belong to simultaneously (the data could be highly sparse as in the letters dataset; or every pixel could be a mixture of components, as in the two-wedge prism dataset). A very large value of $\lambda_{\mathbf{t}}$ (very high sparseness) would try to force each pixel to belong to only one cluster; a lower value of $\lambda_{\mathbf{t}}$ allows for more overlap between clusters, so that each pixel would be more likely to contain more than one component.

Spectra similarity regularization on the sperm dataset. For the sperm dataset, NNMA reconstructions with spectra similarity (but without sparseness) regularization are shown in Figure 3.20, with $\lambda_{\boldsymbol{\mu}_{\text{sim}}} = 5.0$. The reconstructed spectra are improved from those without any regularizations (Figure 3.10) and those with only sparseness regularization (Figure 3.17). As expected, the component maps highlight features that are similar to the cluster components (the higher $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ is, the more similar they will be).

Figure 3.20. NNMA reconstructions of the sperm dataset, with cluster spectra similarity but no sparseness regularization. Parameters are $\lambda_{\boldsymbol{\mu}_{\text{sim}}} = 5.0$, $\lambda_{\mathbf{t}} = 0$, $S = 5$, number of iterations = 10,000. The features identified are comparable to the ones from cluster analysis (Figure 3.11), as expected.



(a) Reconstructed components $s = 1, 2, 3$ are shown here, while $s = 4, 5$ are shown in the next figure.



(b) *(Continued from Figure 3.20(a).)* Reconstructed components $s = 4, 5$.

3.6.3. Combining regularizations

We have seen in Section 3.6.1 that sparseness regularization can provide good feature selection or well-differentiated thickness maps in NNMA reconstructions, but may not yield good spectra, while Section 3.6.2 shows that cluster spectra similarity regularization can provide more realistic spectra with some inevitable sparseness built in (since clustering is a maximum sparseness feature selection method). Of course, not all datasets are described well by a sparse model. Suppose that we would like to keep cluster spectra similarity, while maintaining control over the sparseness of the model. We could combine the two regularizations by simply adding them to the cost function simultaneously [2]. During the iterative updates, we would just update $\boldsymbol{\mu}$ and $\mathbf{t}$ according to their respective update rules.

We note that the basic multiplicative updates of Eqs. (3.7) and (3.8) yield non-increasing cost functions, since they are derived by taking steps in the direction of the steepest negative gradient with respect to the cost function F. Also, since

$$\left[\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D} - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t}\right]_{ij} > 0 \tag{3.18}$$

$$\left[\mathbf{D}\mathbf{t}^{\mathrm{T}} - \boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}}\right]_{ij} > 0 \tag{3.19}$$

(from Eqs. (3.3) and (3.4)), then if we initialize with positive $\mathbf{t}$ and $\boldsymbol{\mu}$, we have

$$t_{ij}^{k+1} = t_{ij}^{k} \left[\frac{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D}}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t}}\right]_{ij} > 0 \tag{3.20}$$

$$\mu_{ij}^{k+1} = \mu_{ij}^{k} \left[\frac{\mathbf{D}\mathbf{t}^{\mathrm{T}}}{\boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}}}\right]_{ij} > 0 \tag{3.21}$$

for the $(k+1)^{\mathrm{th}}$ update. With the modified update of $\mathbf{t}$ with sparseness regularization in Eq. (3.13), $t_{ij}^{k+1} > 0$ remains true since $\lambda_{\mathbf{t}} > 0$. However, the same may not hold true with the modified update of $\boldsymbol{\mu}$ with spectra similarity regularization in Eq. (3.17) – because $(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}})$ can contain negative values, μ_{ij}^{k+1} can become negative. We could set any negative values in $\boldsymbol{\mu}$ to zero after each update, but the problem with this is that once a value has been set to zero, it stays zero according to the multiplicative update rules, and the algorithm becomes pinned to a possibly non-stationary point. Instead, we set negative values to some small, positive value: *e.g.*, 10^{-5}.

[2] Because cluster analysis automatically makes the reconstruction sparse (each pixel is assigned to one and only one cluster), it follows that cluster spectra similarity regularization is not independent of sparseness. This issue will be discussed in Section 4.3.

```
NNMA with sparseness and cluster spectra similarity regularizations

(1) Estimate the number of components $S$.
(2) Initialize $\boldsymbol{\mu}_{N\times S}$, $\mathbf{t}_{S\times P}$ with random, non-negative entries.
(3) While (iters < maxIters) and
        (($\Delta F(\boldsymbol{\mu},\mathbf{t}) < -10^{-6}$) or ($0 < \Delta F(\boldsymbol{\mu},\mathbf{t}) < 10^{-3}$)) do:
    (i) Update $\mathbf{t} : \mathbf{t}' \leftarrow \mathbf{t} \cdot \frac{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D}}{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t} + \lambda_{\mathbf{t}} + \epsilon}$;
    (ii) Set zero entries in $\mathbf{t}'$ to $10^{-5}$;
    (iii) Update $\boldsymbol{\mu} : \boldsymbol{\mu}' \leftarrow \boldsymbol{\mu} \cdot \frac{\mathbf{D}\mathbf{t}'^{\mathrm{T}}}{\boldsymbol{\mu}\mathbf{t}'\mathbf{t}'^{\mathrm{T}} + 2\lambda_{\boldsymbol{\mu}_{\mathrm{sim}}}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{cluster}}) + \epsilon}$;
    (iv) Set zero entries in $\boldsymbol{\mu}'$ to $10^{-5}$;
    (v) For each $k \in \{1 \ldots S\}$, normalize $\boldsymbol{\mu}'_k$ by $\int \boldsymbol{\mu}_{\mathrm{cluster},k}(E)\, dE$;
    (vi) Calculate $F(\boldsymbol{\mu}', \mathbf{t}')$.
```

Figure 3.21. The NNMA algorithm with modified update rules to include sparseness and cluster spectra similarity regularizations. In the updates, the symbol $\cdot$ denotes element-wise matrix multiplication; division is also element-wise. A small $\epsilon \sim 10^{-9}$ is added to the denominator in the update rules to prevent division by zero. Note that the cost function F is not guaranteed to be non-increasing under the modified update rules, as discussed in Section 3.6.3. In one of the stopping conditions, we extend the interval for ΔF into slightly positive territory to take into account that the cost function could now become slightly positive.

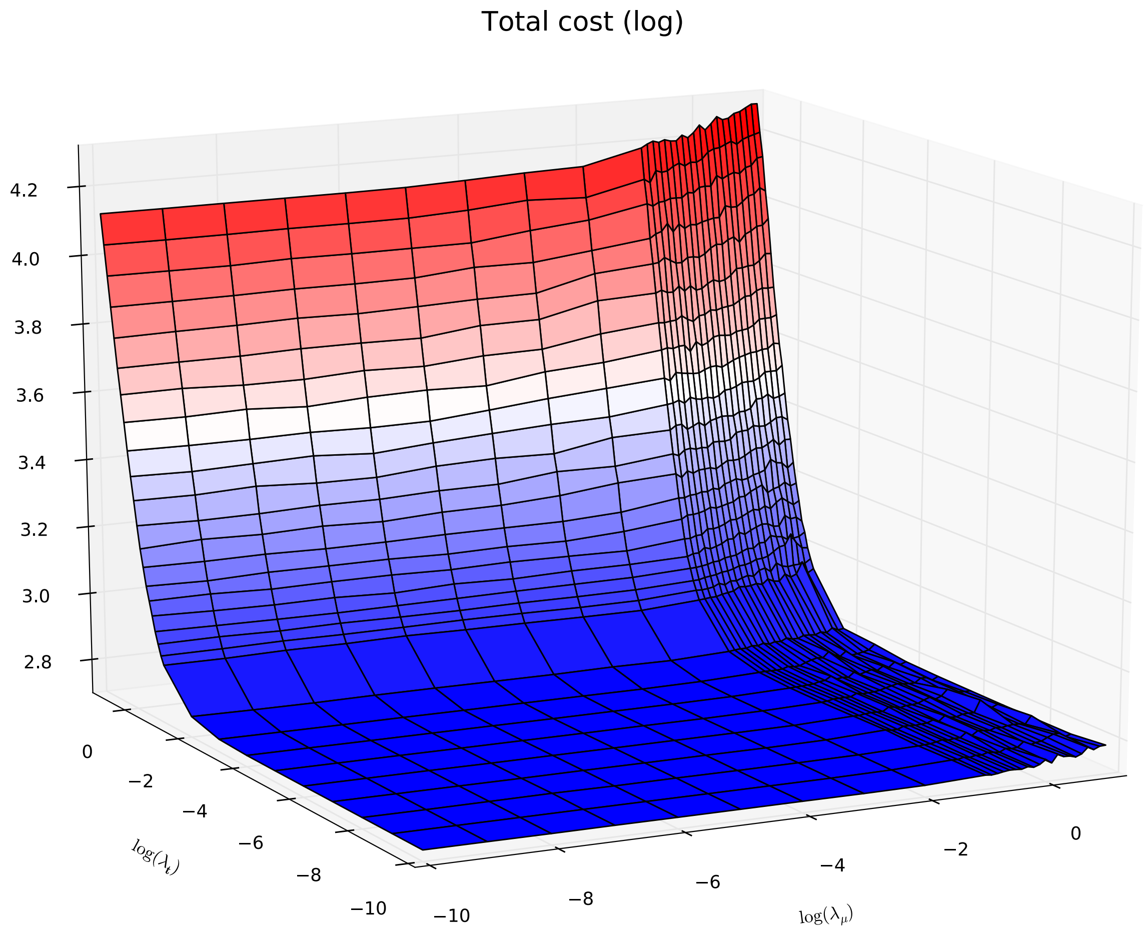
The procedure for NNMA with both sparseness and spectra similarity regularizations is summarized in Figure 3.21.

Combining the two regularizations, the form of the cost function F to be minimized becomes

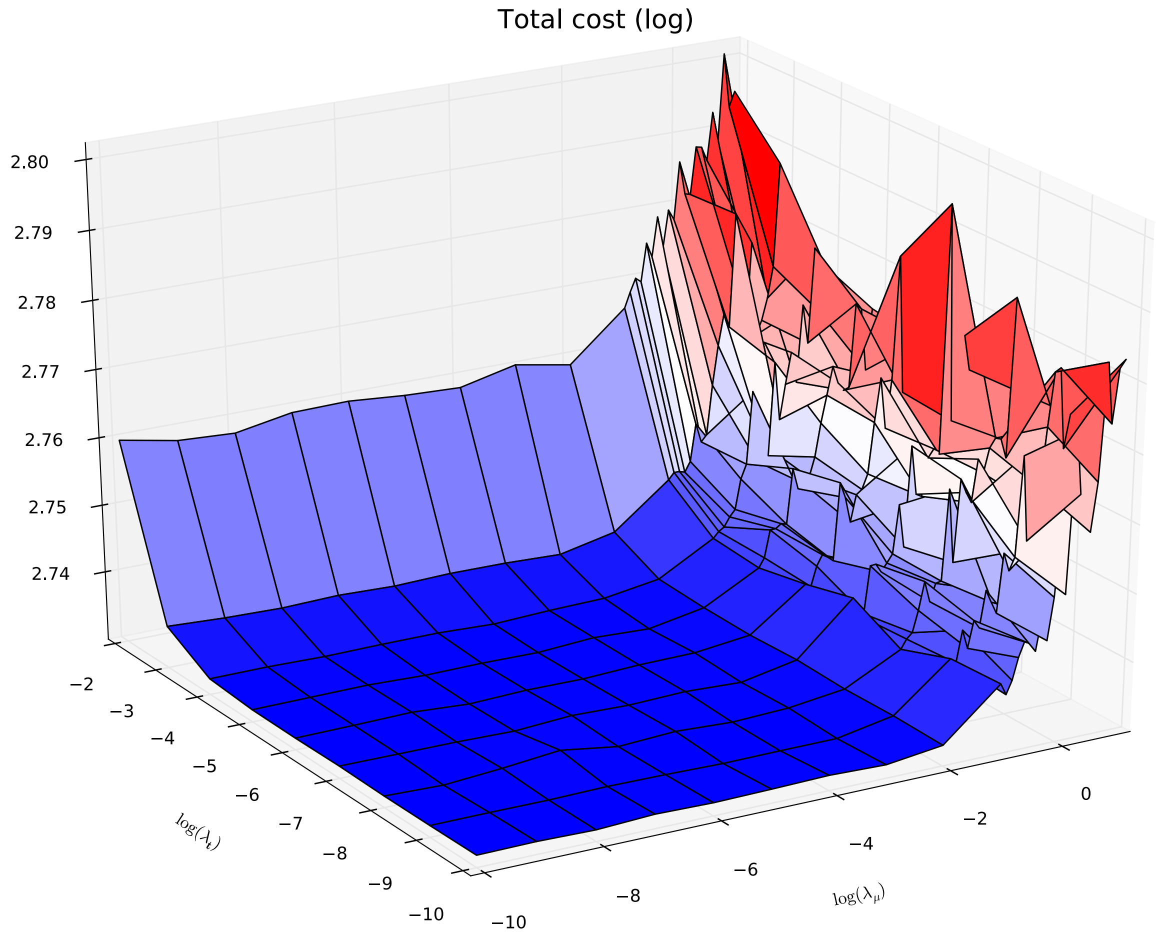
$$F(\boldsymbol{\mu}, \mathbf{t}) = \frac{1}{2}||\mathbf{D} - \boldsymbol{\mu}\mathbf{t}||_2^2 + \lambda_{\boldsymbol{\mu}_{\text{sim}}}||\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}}||_2^2 + \lambda_{\mathbf{t}}||\mathbf{t}||_1, \tag{3.22}$$

and we have to minimize F over a two-dimensional parameter space, $(\lambda_{\boldsymbol{\mu}_{\text{sim}}}, \lambda_{\mathbf{t}})$. We perform systematic calculations of F with different sets of $(\lambda_{\boldsymbol{\mu}_{\text{sim}}},\ \lambda_{\mathbf{t}})$ ranging over 11 orders of magnitude. We chose such a large range because we do not have any initial idea about the scale and relative trade-off or "exchange rate" between $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ and $\lambda_{\mathbf{t}}$; that is, $\lambda_{\mathbf{t}} = 5.0$ could impose a high degree of sparseness for a given dataset, while $\lambda_{\boldsymbol{\mu}_{\text{sim}}} = 5.0$ might correspond to only a moderate match between cluster and NNMA spectra. For the sperm sample, the surface plot of the total cost F over this parameter space from $(10^{-10}, 10^{-10})$ to $(10, 10)$ is shown in Figure 3.22. (The cost function at each parameter pair is taken as the average of 5 runs.) In addition, we can examine the individual contribution from each term in Eq. (3.22): the cost contribution from the data-matching term alone is shown in Figure 3.23, while those from each of the two regularization terms are shown in Figures 3.24 and 3.25. From these parameter plots, we can see that the costs for sparseness and cluster spectra similarity decrease as their respective regularization parameters increase (since we are placing a higher penalty on non-sparse reconstructions or non-similarity to cluster spectra), at the expense of a higher data-matching cost. Although we have not thoroughly studied the relationship between the shape of the parameter plots and the "goodness" of the spectra and thickness reconstructions (which may itself require some objective measure), we find in general that reconstructions are better with parameters within the region where the cost function just begins to increase significantly (Figure 3.22(b)). This corresponds roughly to $-1 < \log(\lambda_{\mathbf{t}}) < 0$ and $-1 < \log(\lambda_{\boldsymbol{\mu}_{\text{sim}}}) < 1$.

Figure 3.22. Minimized total cost F for the sperm sample, as a function of the two regularization parameters, $\lambda_{\mathbf{t}}$ and $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$.

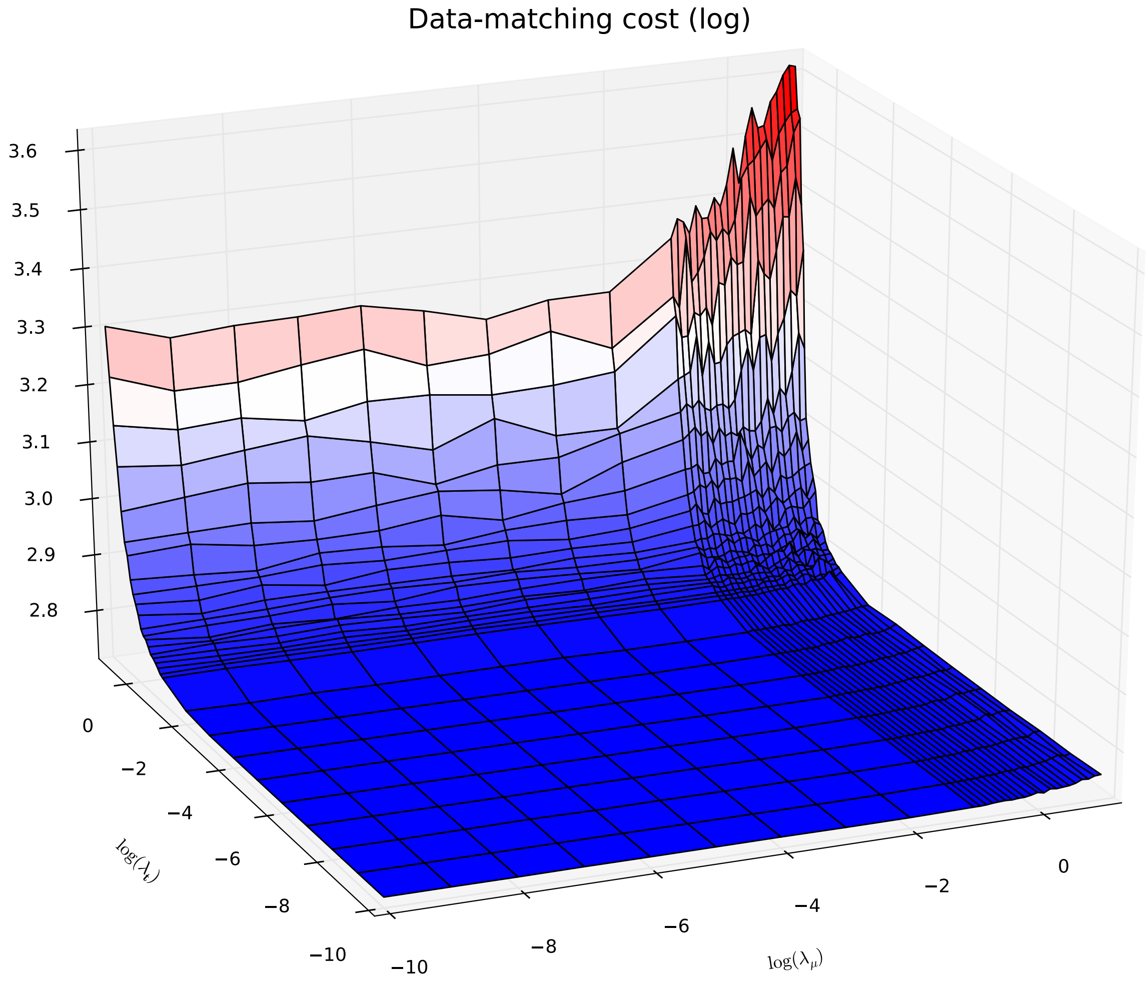


(a) For the most part, F is lowest for small parameters. This is expected since only the data-matching term is of significant contribution to F in Eq. (3.22). As the parameters increase beyond $\sim 10^{-2}$, F also begins to increase noticeably, due to increasing penalty being imposed by the regularization terms. A close-up of the region where F just begins to increase is shown in the next figure.

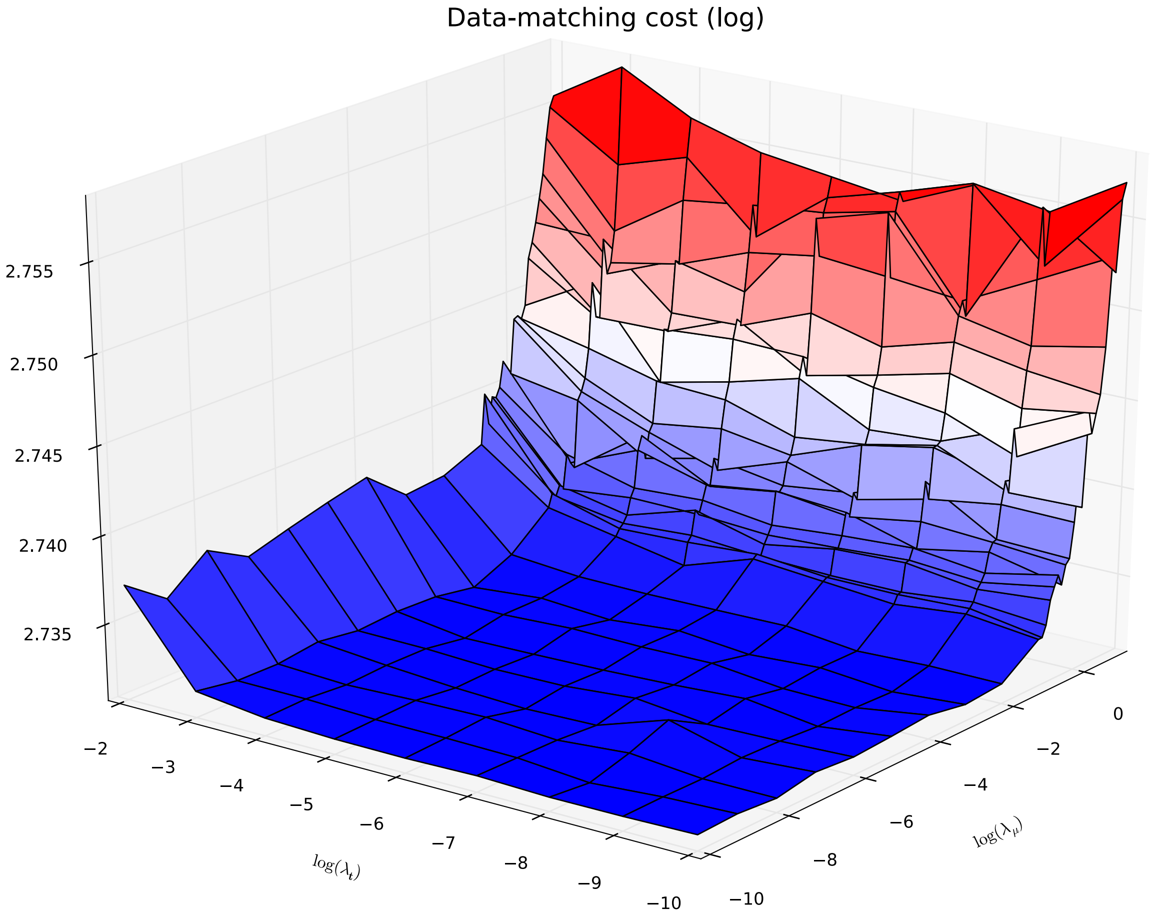


(b) A close-up of the "corner" region in Figure 3.22(a), defined by $\lambda_{\mathbf{t}}$ and $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ where the total cost function F just begins to increase.

Figure 3.23. Contribution to F from the data-matching term in Eq. (3.22) for the sperm sample, as a function of the two regularization parameters, $\lambda_{\mathbf{t}}$ and $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$.



(a) Contribution to F from the data-matching term in Eq. (3.22) for the sperm sample. A close-up of the region where the cost to data-matching just begins to increase is shown in the next figure.



(b) *(Continued from Figure 3.23(a).)* A close-up of the "corner" region defined by $\lambda_{\mathbf{t}}$ and $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ where the data-matching cost just begins to increase.

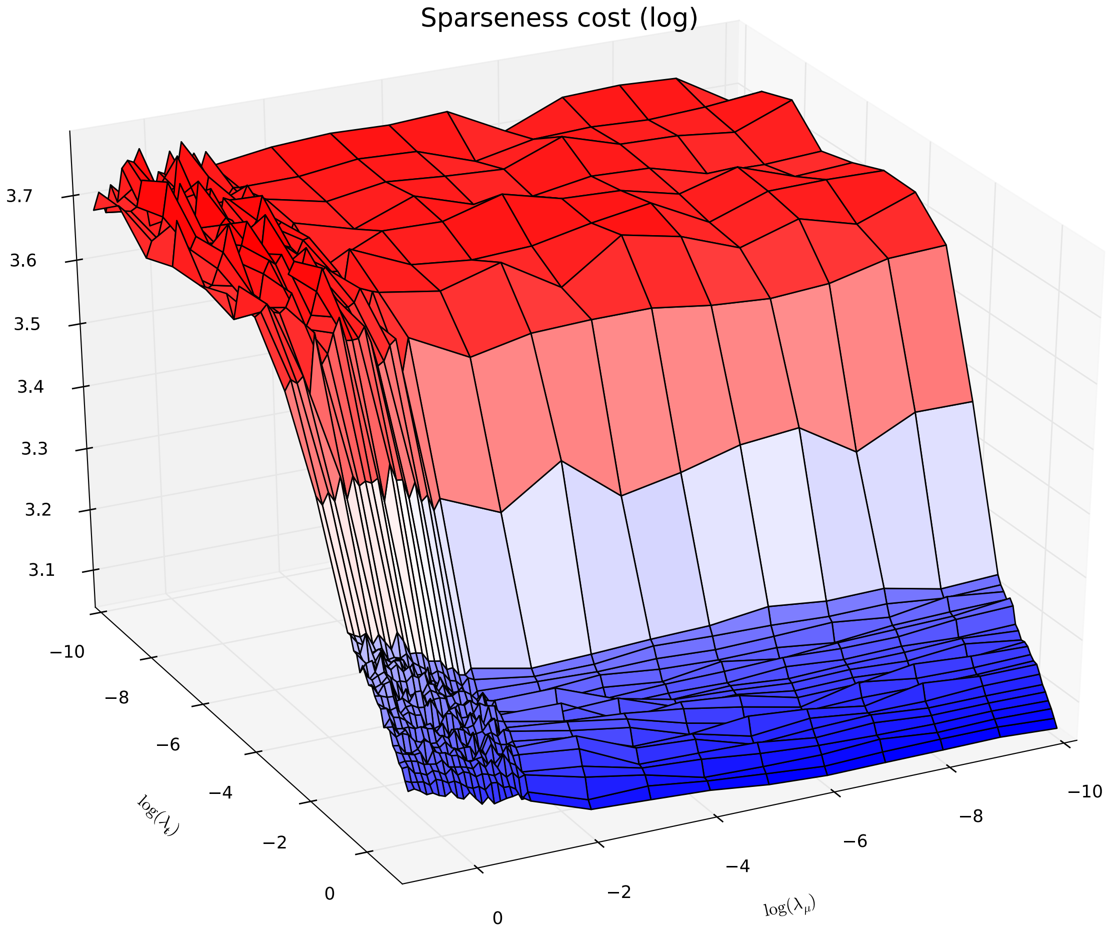


Figure 3.24. Contribution to F from the sparseness regularization term for the sperm sample, as a function of both $\lambda_{\mathbf{t}}$ and $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$. More calculations were done with $\lambda_{\mathbf{t}}$ between 0.1 and 1 (or $-1 < \log(\lambda_{\mathbf{t}}) < 0$) where the cost begins to rise significantly, hence the denser mesh in that interval. The sparseness cost is smaller as $\lambda_{\mathbf{t}}$ increases (since we are penalizing more heavily for non-sparseness), but at the expense of a higher cost to data-matching (Figure 3.23(b)).

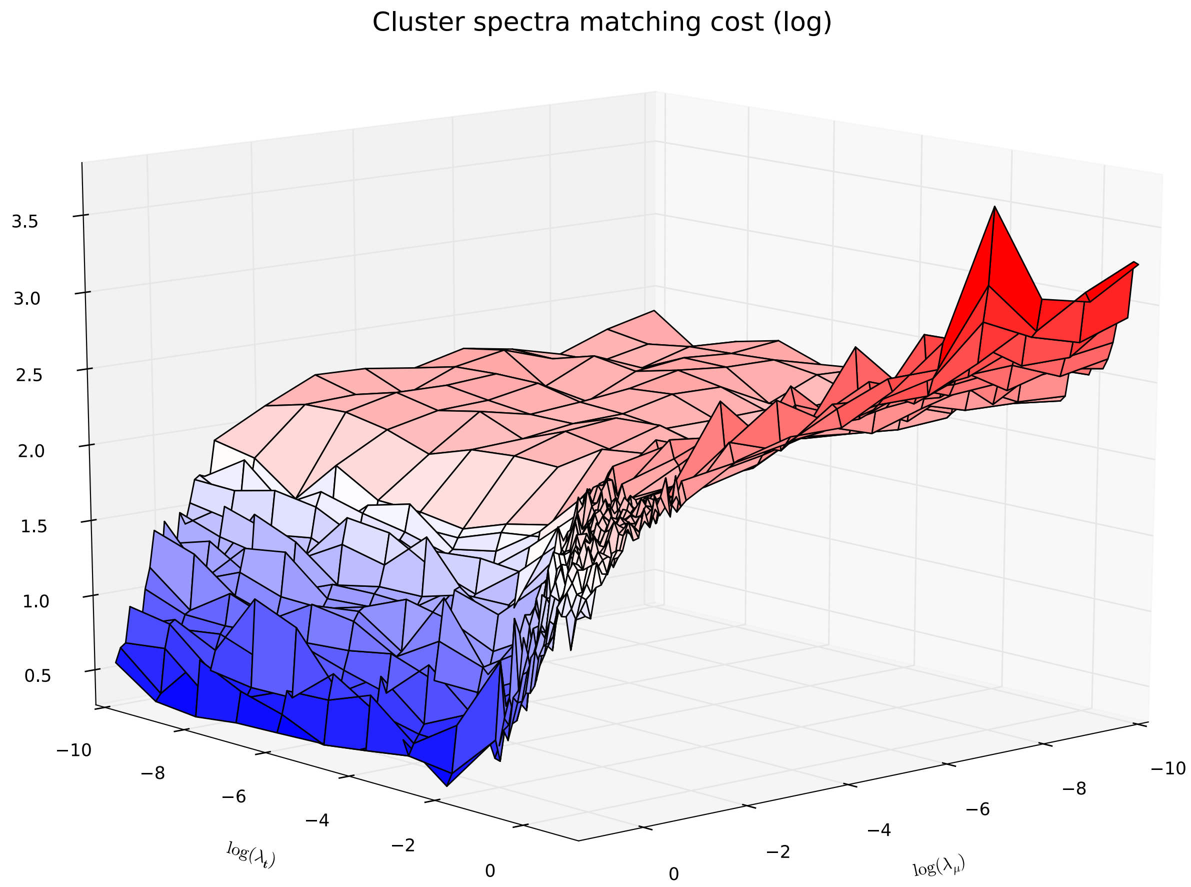


Figure 3.25. Contribution to F from the spectra similarity regularization term for the sperm sample, as a function of both $\lambda_{\mathbf{t}}$ and $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$. Similar to the sparseness cost, the cost to matching NNMA reconstructed spectra to cluster spectra becomes smaller as $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ increases, at the expense of a higher data-matching cost (Figure 3.23(b)).

3.6.4. Other types of regularization

Apart from the sparseness and cluster spectra similarity regularizations discussed above, there are other regularization schemes we could implement to better model the outcome based on knowledge we have about the data, including smoothness, continuity, closure, unimodality, orthogonality, or local rank-selectivity [**61**]. Below, we will briefly discuss two more regularizations relevant to spectromicroscopy; but for reasons due to time constraint or difficulty encountered during implementation, we did not fully test these out.

Smoothness regularization. Absorption spectra, represented by the columns of $\boldsymbol{\mu}$, are smooth functions of energy. This is not very well-reflected in the reconstructed spectra from the basic NNMA procedure on the sperm dataset, as we saw in Figure 3.10. Although the experimental data always contains noise from detectors and other instruments which contributes to fluctuations (non-smoothness), there is also an overall smoothing effect due to the monochromator response if we sample at energy intervals finer than the energy resolution. One way to introduce smoothness into $\boldsymbol{\mu}$ is to add $||\boldsymbol{\mu}||_2^2$, the square of the ℓ_2 norm of $\boldsymbol{\mu}$, as a regularization to the cost function [**44, 62**]. However, in order to have more control over the smoothing of each column in $\boldsymbol{\mu}$, and to take into account that the region around an absorption edge may not be smooth, we instead tried to regularize the second derivative of each column of $\boldsymbol{\mu}$. This should still have the effect of keeping fluctuations small, while also having the advantage of taking non-uniform energy step sizes into account. An explicit form for the smoothness regularizer is derived in Appendix C, but we have not yet been able to integrate it into our regularized NNMA scheme because the formula as it stands appears to increase (rather than decrease, as desired) the cost function as soon as the algorithm begins.

Stochasticity regularization. The thickness maps as they stand do not (yet) relate directly to the physical thicknesses in the sample – for example, it is possible for a thin but copper-rich specimen to exhibit the same absorption around the copper absorption edge as a thicker specimen containing less copper. Thus, to translate the entries in $\mathbf{t}$ into actual thicknesses, we must first measure the absorption far away from the edge, and use this value as a scaling reference when converting into physical thickness.

From a different point of view, the weightings in $\mathbf{t}$ might also be interpreted as probabilities or ratios. From Figure 2.3, we saw that for a given pixel p, its spectral composition is made up of a linear combination of spectra weighted by the corresponding thickness map weightings in the column t_p. If we assume that the S spectral components form a complete set to describe the sample composition, and that the ratios of all components sum to one, then we might interpret the t_{sp} as either the probability of component s being found in pixel p, or the fraction of pixel p that contains component s. In these interpretations, we would want to constrain each column of $\mathbf{t}$ to sum to one (in other words, making $\mathbf{t}$ column stochastic). However, we would also like to leave room for the possibility that our estimate for S may not be exact, so we don't want to place a hard constraint on $\mathbf{t}$. In this instance, we could make use of a "stochasticity regularizer", $J_{\mathbf{t}_{\mathrm{st}}}(\mathbf{t})$, to be added to the cost function [**44**], given by

$$\lambda_{\mathbf{t}_{\mathrm{st}}} J_{\mathbf{t}_{\mathrm{st}}}(\mathbf{t}) = \lambda_{\mathbf{t}_{\mathrm{st}}} ||\mathbf{t}^{\mathrm{T}}\mathbf{e_1} - \mathbf{e_2}||_2^2, \tag{3.23}$$

where $\lambda_{\mathbf{t}_{\mathrm{st}}}$ is the regularization parameter, and $\mathbf{e_1}$, $\mathbf{e_2}$ are column vectors of ones in $\mathbb{R}^S$ and $\mathbb{R}^P$ respectively. $J_{\mathbf{t}_{\mathrm{st}}}$ simply measures the deviation of the sum in each column of $\mathbf{t}$ from one – the bigger the deviation, the higher the penalty in the cost function. As usual, the addition to the $\mathbf{t}$ update rule is the derivative with respect to $\mathbf{t}$, where element kp of the derivative is given by:

$$\lambda_{\mathbf{t}_{\mathrm{st}}} \left[\frac{\partial J_{\mathbf{t}_{\mathrm{st}}}}{\partial \mathbf{t}} \right]_{kp} = 2\lambda_{\mathbf{t}_{\mathrm{st}}} \left(\sum_{s=1}^{S} t_{sp} - 1 \right) \tag{3.24}$$

As expected, this update rule depends only on the column p in $\mathbf{t}$, and not the row k.

CHAPTER 4

Discussion and future directions

4.1. Study of the human sperm cell

In Chapter 3, we considered the application of x-ray spectromicroscopy and NNMA to the study of the XANES spectra and thickness maps of the human sperm cell. Sperm are compact cells with tightly-packed and well-segregated materials in their head, and a long flagellum which allows them to move through fluid (Figure 4.1). Their density and total thickness makes them difficult to study in electron microscopy without sectioning, and their small size means that the sub-50 nm spatial resolution of x-ray spectromicroscopy is helpful for seeing compositional details.

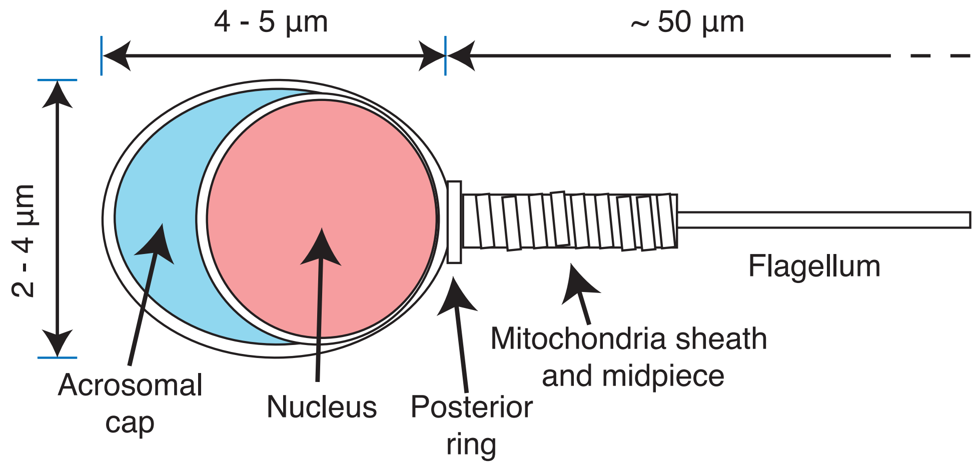


Figure 4.1. Diagram of a human spermatozoa. The enzymes involved in penetrating the egg are in the acrosomal cap, while the nucleus contains the DNA tightly packed with histone proteins. The flagellar motor is within the posterior ring.

One in four couples experience difficulties in conceiving [**63**] with a male factor contributing in more than 50% of these cases [**64**]. The andrological assessment of male infertility relies mostly on conventional criteria of sperm quality like concentration, motility and morphology. While threshold values of these metrics can be used to classify men as subfertile, of indeterminate fertility or fertile, none of the measures are reliable diagnostics of infertility [**65**]. This indicates that the sperm of infertile men may have hidden abnormalities in the composition of their nuclei. DNA damage above a certain threshold appears to impair fertilization and embryo development [**66, 67**], but little is known about the etiologies of sperm DNA damage and its full impact on human reproduction. Light microscopy does not deliver valuable information on sperm DNA or chromatin abnormalities, while bulk chemical measurements average over many morphologies and are not sensitive to individual spermatozoa. Flow cytometry can correlate sperm morphology with total DNA content [**68**], but it is still useful to visualize overall biochemical organization at higher resolution and without using a single biochemical marker. X-ray spectromicroscopy insights into the correlation between sperm morphology and abnormal DNA or protein distributions could lead one to better understand the basis for light microscopy selection of one abnormality over another for in vitro fertilization in cases where no sperm are present with normal morphology.

Several investigators have carried out high resolution x-ray microscopy studies of sperm [**69, 70, 71, 72**]. Zhang *et al.* have used carbon near-edge x-ray absorption spectromicroscopy for compositional mapping of hamster, rat, and bull sperm [**19**]. They acquired spectra of thin film standards of proteins protamine 1 and 2, and of calf thymus deoxyribonucleic acid (DNA); a species-weighted ratio of the protamine spectra was used along with the DNA spectrum to form a two-spectrum matrix $\boldsymbol{\mu}$ which was then inverted using singular value decomposition to yield thickness maps (Eq. 2.6) and estimate protein-to-DNA ratios.

While proteins and DNA make up the majority of content by mass in sperm, we have used NNMA analysis on carbon near-edge x-ray spectromicroscopy data to image the major biochemical organization of human spermatozoa without assuming a composition given by selected thin-film standards. We have used both cluster analysis [**24**] and NNMA to analyze the same dataset.

Using the cluster spectra as targets, we can use Eq. (2.6) to calculate the thickness maps for each of the components. These maps and cluster spectra are shown in comparison to results by NNMA analysis in Figure 4.2. We can readily identify the components found by NNMA as the lipid membrane and flagellum (t_2), the acrosomal cap (t_3), the nucleus (t_4) containing DNA with tightly-packed histone proteins, while t_1 appears to highlight the posterior ring in which the flagellar motor resides.

The areas of negative thickness (red) in the cluster maps indicate that the cluster spectra do not perfectly represent all chemical signatures in the sample. However, as demonstrated in Section 3.6.2, it is possible to use the cluster spectra in a spectra similarity regularization term in NNMA in order to produce a more realistic set of reconstructed spectra and non-negative thickness maps.

4.2. A note about the run-time of NNMA

As described in Section 3.6.3, we run the NNMA algorithm for a set number of iterations ($\sim 10^4$), or until the change in the cost function ΔF is below some threshold ($\sim 10^{-6}$). NNMA runs most quickly without regularizations, and an average run with 10^4 iterations typically takes on the order of 10 minutes on a laptop computer with a 2.66 GHz dual-core Intel i7 processor. Depending on the chosen regularization parameters, the same number of runs could take about 2 to 3 times longer. For very sub-optimal parameters, ΔF can start

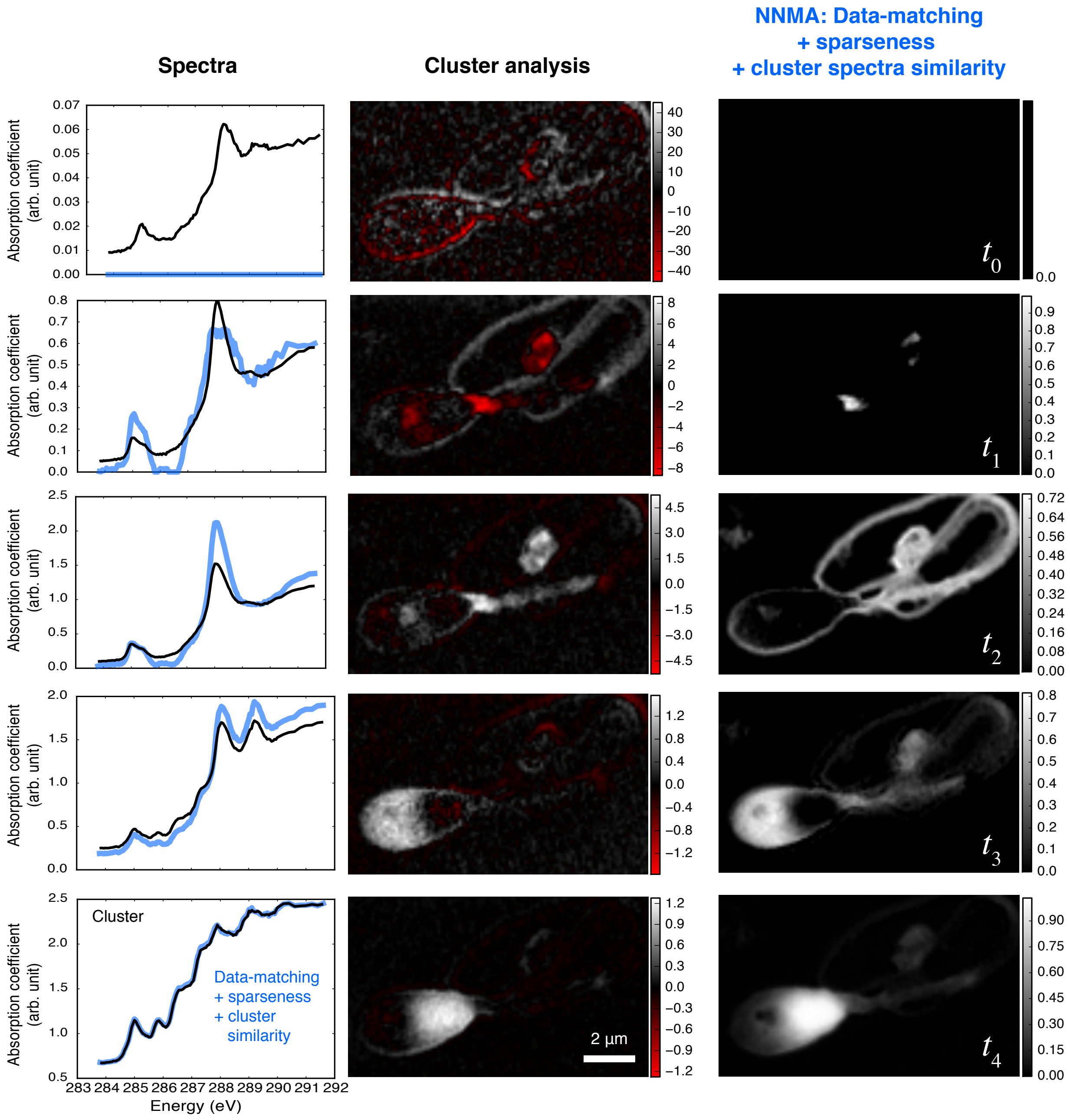


Figure 4.2. Comparison of cluster and NNMA analysis of the sperm dataset, showing their respective reconstructions of absorption spectra and corresponding thickness maps. The cluster spectra are shown in black and its corresponding thickness maps are shown in the middle column, while the NNMA spectra are shown in blue with corresponding maps in the rightmost column. The regularization parameters used are $\lambda_{\mathbf{t}} = 0.5$ and $\lambda_{\boldsymbol{\mu}_{\text{sim}}} = 10$.

increasing quickly, preventing the algorithm from iterating more than a few times before stopping. In general, the run-times of our NNMA implementation is an improvement over those reported which are on the order of days for previous NNMA algorithms tested on similarly sized spectromicroscopy datasets [**50**].

4.3. Limitations of regularized NNMA in its current form

We have shown that regularized NNMA provides an improvement in the reconstruction of interpretable spectra and thickness maps. We first applied NNMA with the basic multiplicative update rules provided by Lee and Seung [**48**], and found that because of the non-uniqueness of the factorized solutions, the reconstructions of the absorption spectra $\boldsymbol{\mu}$ and thickness maps $\mathbf{t}$ are not well-defined. We then added sparseness and cluster spectra similarity regularization terms to the cost function, leveraging prior knowledge about the data to build a model for producing more realistic spectra and maps.

The sparseness regularization $||\mathbf{t}||_1$ alone was able to give well-differentiated thickness maps with better feature selection, but the spectral reconstructions still left something to be desired (smoothness). To remedy this, we took the set of spectra from cluster analysis and used it as a basis for comparison with NNMA reconstructions via a spectra similarity regularization term $||\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}}||_2^2$ in the cost function, imposing a higher penalty for larger deviations. This helped to add smoothness to the NNMA spectra.

However, while cluster analysis does appear to provide a good basis set with which to compare NNMA spectra, there is a sense that this is somewhat of a “shortcut”. Because cluster analysis by default provides maximum sparseness by assigning each pixel in an image exclusively to one component (which is great from a feature selection point of view), it is rather inflexible and can lead to inaccurate results in pixels which contain a mixture of two or more components. In addition, the thickness maps obtained from the inverse SVD of cluster spectra contain non-physical negative regions, which in turn indicates corrections to be made to the spectra. Thus we run the risk that too much emphasis on a close spectral match can detract NNMA from finding a better solution. In addition, cluster analysis inherently selects features with maximum sparseness, since each pixel is assigned exclusively to one component. This could counteract or at least interfere with the sparseness that we specify via the sparseness regularization parameter $\lambda_{\mathbf{t}}$. Recall that the reason we included this similarity regularization in the first place was because NNMA often produced non-smoothly varying absorption spectra, without or with sparseness regularization. Since cluster analysis spectra are generally smooth functions with respect to energy, we were able to confer some smoothness to the NNMA spectra by applying a similarity regularization. However, it would be even better if we were able to achieve smoothness directly via a smoothness regularization.

There are different ways to regularize smoothness: $||\boldsymbol{\mu}||^2$ to minimize fluctuations; but this can suppress resonances or even absorption edges which contain important information for chemical identification. We could also minimize the first derivative of each column of $\boldsymbol{\mu}$, which takes neighbouring points into account and tries to minimize large deviations from one energy point to the next. But the absorption edge might again be suppressed, since we take the value E_0 to be at the point of inflection along the edge, and the first derivative at the point E_0 is not necessarily zero. Naturally then, a third option for smoothness regularization would be to minimize the second derivative of each column of $\boldsymbol{\mu}$. This would not suppress resonance peaks or absorption edges, and has the added advantage that since the numerical calculation of the second derivative is a 5-point difference algorithm, smoothing around a given μ_{ij} does not just take into account its nearest neighbours ($\mu_{i-1,j}$ and $\mu_{i+1,j}$), but also its next nearest neighbours ($\mu_{i-2,j}$ and $\mu_{i+2,j}$), thus allowing for more accurate smoothing.

4.4. The bionanoprobe: x-ray fluorescence imaging with spectroscopy

At the end of Chapter 1, we briefly mentioned that high-brilliance synchrotron x-rays that could be tuned over a wide range of energy are capable of collecting rich absorption spectra over millions of pixels in a reasonable amount of time. In addition, if we could collect fluorescence signals from the same sample and perhaps even rotate the sample and repeat measurements at different tomographic projections, this would allow us to have a more comprehensive understanding of the elemental and chemical make-up of the material under study. One such instrument capable of performing these tasks is a recently-commissioned scanning x-ray bionanoprobe located at the Advanced Photon Source at Argonne National Laboratory, equipped with fluorescence and transmission detectors for the high-resolution imaging of frozen-hydrated biological specimens at cryogenic temperatures. The segmented transmission detector also allows differential phase contrast imaging by looking at the relative difference in signals between two given segments. While scanning fluorescence using the bionanprobe has already been demonstrated [**73, 74**], spectroscopy was only recently tested [**75**], with the eventual goal of combining both fluorescence microscopy imaging and absorption spectroscopy in order to reap the benefits of spectroscopic imaging discussed in Chapter 1.

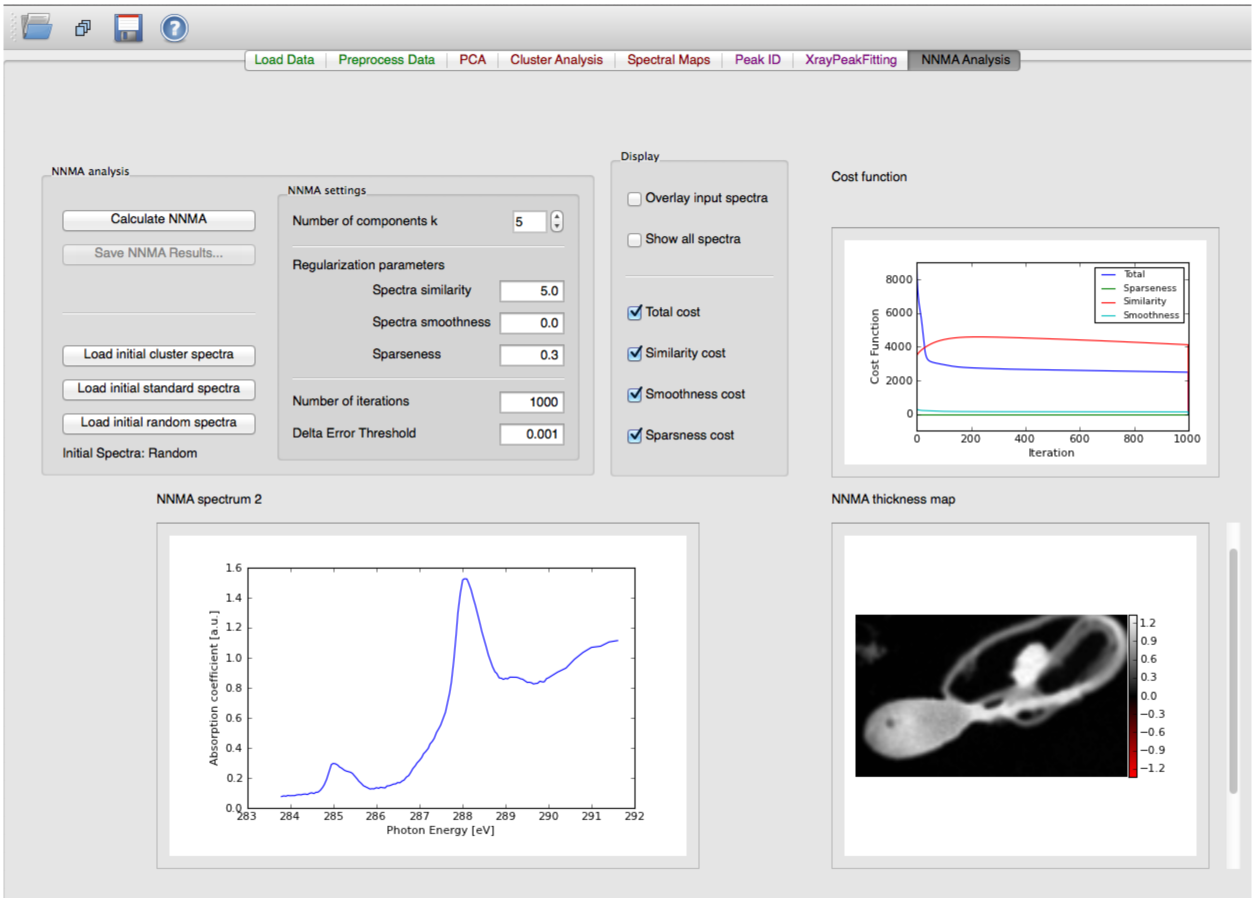
4.5. Other applications in x-ray spectromicroscopy

In the future, we would also like to be able to apply NNMA to the EXAFS of real datasets, and examine reconstructed structures in the extended energy region which we can Fourier transform in order to determine interatomic distances of different compounds as well as the edge energy shifts to identify the different valence states of the absorbing atom. One such application would be to the study of arsenic contamination in soil samples, which can contain a complex mixture of minerals and organic components [**76**]. Arsenic is known to be of environmental and health concern. The toxicity of arsenic depends on its chemical speciation – for example, arsenite (As III) is usually more toxic than arsenate (As V) [**77**]. Speciation can depend on environmental factors such as pH and oxygen level (which alter the redox potential of As), and can affect the adsorption of dissolved arsenic – an important remediation tool for the removal of arsenic from contaminated drinking water.

Another interesting application of EXAFS is to the study of bimetallic (*e.g.*, Au-Pd) nanoparticle catalysts [**78, 79**], which are used for catalysing oxidation or reduction in the process of transforming bio-renewable materials [**80**]. Quantitative information such as the nanoparticle's size, shape, metal-metal bond length, and coordination number could be obtained. NNMA could be used as a test to check the uniformity of the nanoparticles, which is difficult to achieve in mass production.

4.6. MANTiS: a graphical user interface application

Recently, an open-source application for the analysis of spectromicroscopy data has been developed in Python by Lerotic [**81**]. The application, called MANTiS (Multivariate ANalysis Tool for Spectromicroscopy), includes a graphical user interface which guides the user through a series of steps in the analysis, from data pre-processing (limiting energy ranges, choosing regions of interest, setting background threshold, etc.), to PCA and cluster analysis. We have also implemented in MANTiS the regularized NNMA algorithms discussed in Chapter 3. The NNMA module calls functions for both the basic multiplicative updates or the modified updates depending on the user's selection of regularization parameters. A screenshot of the NNMA "tab" is shown in Figure 4.3.



Initial Spectra: Random

NNMA settings

Number of components k 5

Regularization parameters

Spectra similarity 5.0

Spectra smoothness 0.0

Sparseness 0.3

Number of iterations 1000

Delta Error Threshold 0.001

☐ Overlay input spectra

☐ Show all spectra

☑ Total cost

☑ Similarity cost

☑ Smoothness cost

☑ Sparsness cost

Figure 4.3. A screenshot of the NNMA analysis tab in MANTiS.

The code is open source, and the repository could be cloned from `https://code.google.com/p/spectromicroscopy`.

References

[1] Siegfried Hofmann. *Auger- and X-Ray Photoelectron Spectroscopy in Materials Science.* Springer, 2013.

[2] H. H. Rose. Optics of high-performance electron microscopes. *Science and Technology of Advanced Materials*, 9(1), 2008.

[3] David L. Wetzel and Steven M. LeVine. Imaging molecular chemistry with infrared microscopy. *Science*, 285(5431):1224–1225, 1999.

[4] Michael C. Martin et al. 3D spectral imaging with synchrotron fourier transform infrared spectro-microtomography. *Nat. Meth.*, 10:861–864, 2013.

[5] Joachim Stöhr. *NEXAFS Spectroscopy.* Springer-Verlag, 1992.

[6] Dale E. Sayers, Edward A. Stern, and Farrel W. Lytle. New technique for investigating noncrystalline structures: Fourier analysis of the extended x-ray¯absorption fine structure. *Phys. Rev. Lett.*, 27:1204–1207, Nov 1971.

[7] P. A. Lee and J. B. Pendry. Theory of the extended x-ray absorption fine structure. *Phys. Rev. B*, 11:2795–2811, Apr 1975.

[8] J. J. Rehr. Theoretical approaches to x-ray absorption fine structure. *Reviews of Modern Physics*, 72(3):621–654, July 2000.

[9] Robert A. Neville, C. Nadeau, John Levesque, Tomas Szeredi, Karl Staenz, P. Haurr, and Gary A. Borstad. Hyperspectral imagery for imagery exploration: comparison of data from two airborne sensors. In *Proc. SPIE 3438, Imaging Spectroscopy IV, 74*, 1998.

[10] Floyd F. Sabins. Remote sensing for mineral exploration. *Ore Geology Reviews*, 14:157–183, 1999.

[11] C. J. Keith, K. S. Repasky, R. L. Lawrence, S. C. Jay, and J. L. Carlsten. Monitoring effects of a controlled subsurface carbon dioxide release on vegetation using a hyperspectral imger. *International Journal of Greenhouse Gas Control*, 3:096003–1–9, 2009.

[12] K. Hege, D. O'Connell, W. Johnson, S. Basty, and E. L. Dereniak. Hyperspectral imaging for astronomy and space surveillance. In S. S. Shen and P. E. Lewis, editors, *Proc. SPIE 5159, Imaging Spectroscopy IX*, pages 380–391, 1998.

[13] Rajinder P. Singh-Moon, Darren M. Roblyer, Irving J. Biggio, and Shailendra Josha. Spatial mapping of drug delivery to brain tissue using hyperspectral spatial frequency-domain imaging. *Journal of Biomedical Optics*, 19:626–632, 2014.

[14] Jiewen Zhao, Quansheng Chen, Jianrong Cai, and Qin Ouyang. Automated tea quality classification by hyperspectral imaging. *Applied Optics*, 48:3557–3564, 2009.

[15] N Bonnet. Multivariate statistical methods for the analysis of microscopic image series: applications in material science. *Journal of Microscopy*, 190(1–2):2–18, 1998.

[16] N Bonnet, N Brun, and C Colliex. Extracting information from sequences of spatially resolved eels spectra using multivariate statistical analysis. *Ultramicroscopy*, 77:97–112, 1999.

[17] Peter Lasch, Wolfgang Wäsche, W J McCarthy, Gerhard J Müller, and Dieter Naumann. Imaging of human colon carcinoma thin sections by FT-IR microspectrometry. *Proc. SPIE 3257, Infrared Spectroscopy: New Tool in Medicine*, 3257:187–198, 1998.

[18] P Lasch, M Boese, A Pacifico, and M Diem. FT-IR spectroscopic investigations of single cells on the subcellular level. *Vibrational Spectroscopy*, 28:147–157, 2002.

[19] X Zhang, R Balhorn, J Mazrimas, and J Kirz. Mapping and measuring DNA to protein ratios in mammalian sperm head by XANES imaging. *Journal of Structural Biology*, 116(3):335–344, May 1996.

[20] I N Koprinarov, A P Hitchcock, C T McCrory, and R F Childs. Quantitative mapping of structured polymeric systems using singular value decomposition analysis of soft x-ray images. *Journal of Physical Chemistry B*, 106(21):5358–5364, 2002.

[21] P L King, R Browning, P Pianetta, I Lindau, M Keenlyside, and G Knapp. Image processing of multispectral x-ray photoelectron spectroscopy images. *Journal of Vacuum Science and Technology A*, 7(6):3301–3304, 1989.

[22] A Osanna and Chris Jacobsen. Principal component analysis for soft x-ray spectromicroscopy. In T Warwick, W Meyer-Ilse, and David T Attwood, editors, *X-ray Microscopy: Proceedings of the Sixth International Conference (American Institute of Physics Conference Proceedings)*, pages 350–357, 2000.

[23] Donald F. Morrison. *Multivariate Statistical Methods*. McGraw-Hill, second edition, 1976.

[24] M. Lerotic, C. Jacobsen, T. Schäfer, and S. Vogt. Cluster analysis of soft x-ray spectromicroscopy data. *Ultramicroscopy*, 100(1–2):35–57, 2004.

[25] M. Lerotic, C. Jacobsen, J.B. Gillow, A.J. Francis, S. Wirick, S. Vogt, and J. Maser. Cluster analysis in soft x-ray spectromicroscopy: finding the patterns in complex specimens. *Journal of Electron Spectroscopy and Related Phenomena*, 144–147:1137–1143, 2005.

[26] Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *Computer Journal*, 26:354–359, 1983.

[27] Paul D. McNicholas and Thomas Brendan Murphy. Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26:2705–2712, 2010.

[28] K.-L. Du. Clustering: A neural network approach. *Neural Networks*, 23:89–107, 2010.

[29] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.

[30] J. Goldberger and S. T. Roweis. Hierarchical clustering of a mixture model. In *Advances in Neural Information Processing Systems*, pages 505–512, 2004.

[31] G Mitrea, J Thieme, P Guttmann, S Heim, and S Gleber. X-ray spectromicroscopy with the scanning transmission x-ray microscope at BESSY II. *Journal of Synchrotron Radiation*, 15(1):26–35, December 2007.

[32] Johannes Lehmann, Dawit Solomon, James Kinyangi, Lena Dathe, Sue Wirick, and Chris Jacobsen. Spatial complexity of soil organic matter forms at nanometre scales. *Nature Geoscience*, 1(4):238–242, 2008.

[33] Bhavin J. Shastri and Martin D. Levine. Face recognition using localized features based on non-negative sparse coding. *Machine Vision and Applications*, 18:101–122, 2007.

[34] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmentrics*, 5:111–126, 1994.

[35] Srilakshmi Inuganti and Veerraju Gampala. Image compression using constrained non-negative matrix factorization. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3:498–503, 2013.

[36] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[37] N. Rao, S. J. Shepherd, and D. Yao. Extracting characteristic patterns from genome-wide expression data by non-negative matrix factorization. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, pages 556–562, august 2004.

[38] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[39] Amy N. Langville and Carl D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2006.

[40] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42:30–37, 2009.

[41] M. S. Karoui, Y. Deville, S. Hosseini, A. Ouamri, and D. Ducrot. Contribution of non-negative matrix factorization to the classification of remote sensing images. *Proc. SPIE 7109, Image and Signal Processing for Remote Sensing XIV*, 7109, 2008.

[42] P. Paatero and U. Tapper. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[43] P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37:23–35, 1997.

[44] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, September 2007.

[45] Andrzej Cichocki and Rafal Zdunek. Multilayer nonnegative matrix factorization using projected gradient approaches. *International Journal of Neural Systems*, 17(6):431–446, 2007.

[46] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, Oct 1999.

[47] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[48] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In T K Leen, T G Dietterich, and V Tresp, editors, *Advances in Neural Information Processing Systems*, pages 556–562, 2001. 14th Annual Neural Information Processing Systems Conference (NIPS), Denver, CO, Nov 27-Dec 02, 2000.

[49] Albert Thompson et al., editors. *X-ray Data Booklet*. Lawrence Berkeley National Laboratory, University of California, third edition, September 2009.

[50] Holger Fleckenstein. *High resolution chemical mapping via scanning transmission x-ray microscopy*. PhD thesis, Stony Brook University, August 2008.

[51] Michael W. Berry and Murray Browne. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11:249–264, 2005.

[52] Farial Shahnaz, Michael W. Berry, V. Paul Pauca, and Robert J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42:373–386, 2006.

[53] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[54] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[55] Günther Palm. Neural associative memories and sparse coding. *Neural Networks*, 37:165–171, 2013.

[56] Michael C. Wu, Lingsong Zhang, Zhaoxi Wang, David C. Christiani, and Xihong Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.

[57] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodology)*, 58:267–288, 1996.

[58] Morten Arendt Rasmussen and Rasmus Bro. A tutorial on the lasso approach to sparse modeling. *Chemometrics and Intelligent Laboratory Systems*, 119:21–31, 2012.

[59] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[60] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Methodology)*, 67:301–320, 2005.

[61] Rafal Zdunek and Andrzej Cichocki. Blind image separation using non-negative matrix factorization with gibbs smoothing. In *ICONIP 2007, Part II, LNCS 4985*, pages 519–528. Springer-Verlag, 2008.

[62] L. Taslaman and B. Nilsson. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PLOS ONE*, 7(11):e46331, 2012.

[63] L. Schmidt, K. Munster, and P. Helm. Infertility and the seeking of infertility treatment in a representative population. *British Journal of Obstetric Gynecology*, 102:978–984, 1995.

[64] D. S. Irvine. Epidemiology and aetiology of male infertility. *Human Reproduction*, 13(suppl. 1):33–44, 1996.

[65] D. S. Guzick, J. W. Overstreet, P. Factor-Litvak, et al. Sperm morphology, motility, and concentration in fertile and infertile men. *New England Journal of Medicine*, 345(19):1388–1393, 2001.

[66] A. Ahmadi and S. C. Ng. Fertilizing ability of DNA-damaged spermatozoa. *Journal Of Experimental Zoology*, 284(6):696–704, 1999.

[67] Chunghee Cho, Haesook Jung-Ha, William D Willis, Eugenia H Goulding, Paula Stein, Zhe Xu, Richard M Schultz, Norman B Hecht, and Edward M Eddy. Protamine 2 deficiency leads to sperm DNA damage and embryo death in mice. *Biology of Reproduction*, 69(1):211–217, 2003.

[68] U B Hacker-Klom, W Göhde, E Nieschlag, and H M Behre. DNA flow cytometry of human semen. *Human Reproduction*, 14(10):2506–2512, October 1999.

[69] B. W. Loo, Jr., S. Williams, S. Meizel, and S. S. Rothman. X-ray stereomicroscopy: high resolution 3-D imaging of human spermatozoa in aqueous suspension with natural contrast. *Journal of Microscopy*, 166:RP5–RP6, 1992.

[70] R. Balhorn, R. E. Braun, B. Breed, J. T. Brown, D. Evenson, J. M. Heck, J. Kirz, I. McNulty, W. Meyer-Ilse, and X. Zhang. Applications of x-ray microscopy to the analysis

of sperm chromatin. In J. Thieme, G. Schmahl, E. Umbach, and D. Rudolph, editors, *X-ray Microscopy and Spectromicroscopy*, pages II–29–46, Berlin, 1998. Springer-Verlag.

[71] T. Vorup-Jensen, T. Hjort, J. V. Abraham-Peskir, P. Guttmann, J. C. Jensenius, E. Uggerhøj, and R. Medenwaldt. X-ray microscopy of human spermatozoa shows change of mitochondrial morphology during capacitation. *Human Reproduction*, 14:880–884, 1998.

[72] J. Abraham-Peskir, E. Chantler, C. McCann, R. Medenwaldt, and E. Ernst. Ultrastructure of human sperm using x-ray microscopy. *Medical Science Research*, 26:663–667, 1998.

[73] S. Chen, J. Deng, Y. Yuan, C. Flachenecker, R. Mak, et al. The bionanoprobe: hard x-ray fluorescence nanoprobe with cryogenic capabilities. *Journal of Synchrotron Radiation*, 21:66–75, 2014.

[74] Y. Yuan, S. Chen, T. Paunesku, S. C. Gleber, W. Liu, C. Doty, R. Mak, et al. Epidermal growth factor receptor targeted nuclear delivery and high-resolution whole cell x-ray imaging of Fe_3O_4@TiO_2 nanoparticles in cancer cells. *ACS Nano*, 7:10502–10517, 2013.

[75] J. Glass. Nano-scale elemental imaging of microbes and minerals from deep sea methane seeps. `http://goldschmidt.info/2014/abstracts/abstractView?abstractId=3757`, 2014.

[76] Khalid H. Al-Assaf, F. Tyson Julian, and Peter C. Uden. Determination of four arsenic species in soil by sequential extraction and high performance liquid chromatography with post-column hydride generation and inductively coupled plasma optical emission spectrometry detection. *J. Anal. At. Spectrom*, 24:376–384, 2009.

[77] K. Sharma, Virender and Mary Sohn. Aquatic arsenic: Toxicity, speciation, transformations, and remediation. *Environment International*, 35:743–759, 2009.

[78] Naoki Toshima and Tetsu Yonezawa. Bimetallic nanoparticles – novel materials for chemical and physical applications. *New Journal of Chemistry*, 22:1179–1201, 1998.

[79] Anatoly I. Frenkel, Oded Kleifeld, Stephen R. Wasserman, and Irit Sagi. Phase speciation by extended x-ray absorption fine structure spectroscopy. *J. Chem. Phys*, 116:9449, 2002.

[80] M. Sankar, N. Dimitratos, P. J. Miedziak, P. P. Wells, C. J. Kiely, and G. J. Hutchings. Designing bimetallic catalysts for a green and sustainable future. *Chemical Society Reviews*, 41:8099–8139, 2012.

[81] Mirna Lerotic, Rachel Mak, Sue Wirick, Florian Meirer, and Chris Jacobsen. Mantis: a program for the analysis of x-ray spectromicroscopy data. *Journal of Synchrotron Radiation*, 21, 2014.

[82] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

[83] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society Series B (Methodology)*, 73:273–282, 2011.

APPENDIX A

Matrix norms

In Chapter 3, our goal was to find a balance between the minimization of various ℓ_2 and ℓ_1 in order to yield different characteristics from the data. For completeness, we include here the explicit form of each norm, and a brief qualitative description of how each norm can be used to tease out different features.

A.1. Error minimization using the ℓ_2 norm

When applying NNMA to our x-ray absortpion spectromicroscopy data, the basic criterion is to minimize the difference or error between the measured optical density $\mathbf{D}$, and the approximation obtained from the NNMA reconstruction $\boldsymbol{\mu}\mathbf{t}$. One way to do this is to minimize the ℓ_2 norm of this difference, *i.e.*, the basic cost function

$$F(\boldsymbol{\mu}, \mathbf{t}) = \frac{1}{2}||\mathbf{D} - \boldsymbol{\mu}\mathbf{t}||_F^2 \tag{A.1}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{P}\left(D_{ij} - \sum_{k=1}^{S}\mu_{ik}t_{kj}\right)^2 . \tag{A.2}$$

This is a least squares problem and is commonly used for error minimization. The ℓ_2 norm is also known as the Euclidean or Frobenius norm. In addition to the basic cost function, we also used the ℓ_2 norm as a regularization to minimize the difference between the NNMA and cluster analysis absorption spectra $\boldsymbol{\mu}$.

A.2. Sparseness and the ℓ_0 "norm"

If we would like the $\mathbf{t}$ matrix to be sparse, we would ideally minimize its number of non-zeros (though we don't want any component (row) or any pixel (column) to contain all zeros as it would then carry no information and be redundant), *i.e.*, we would minimize

$$||\mathbf{t}||_0 = \#\{(i,j) \,|\, t_{ij} \neq 0\} \tag{A.3}$$

or the ℓ_0 "norm" of $\mathbf{t}$ (as mentioned in the text, this is not a true norm as it does not satisfy all the properties of a norm). However, the minimization of the ℓ_0 norm is a combinatorial (counting) problem known to be NP-hard, and so cannot be solved in polynomial time [**82**].

A.3. Sparseness regularization using the ℓ_1 norm

As an approximation to the NP-hard problem of minimizing $||\mathbf{t}||_0$ to regularize sparseness in $\mathbf{t}$, we instead minimize its ℓ_1 norm:

$$||\mathbf{t}||_1 = \sum_{i,j} |t_{ij}|. \tag{A.4}$$

We mentioned in Section 3.6.1 that the ℓ_1 norm is a popular approximation to the ℓ_0 norm for regularizing sparseness. To understand this visually, suppose that we are only interested in two components, so that $rank(\mathbf{t}) = 2$. Then the component space of $\mathbf{t}$ can be represented on the 2D plane, where each axis represents one of the components. The ℓ_0 norm is depicted as the bolded line segments (non-negative region along the axes) in Figure A.1, and each point (or pixel) is located somewhere on the S_1S_2 plane depending on its component composition. In the ideal case of maximum sparseness, each pixel would contain just one (pure) component, and would lie along one the bolded line segments. Minimization of $||\mathbf{t}||_0$ would then result the the clustering of pixels along these two line segments. However, since the minimization

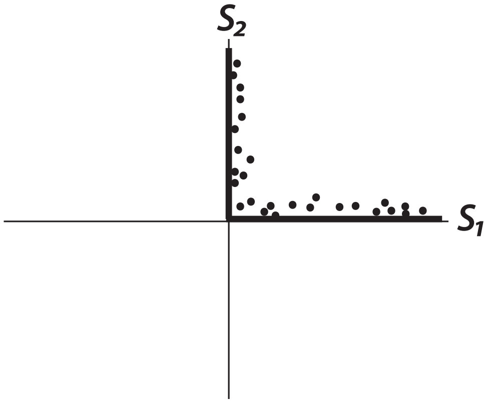


Figure A.1. A representation of the component space of $\mathbf{t}_{2\times P}$ in the 2D plane. In the ideal case where $||\mathbf{t}||_0$ is minimized for maximum sparseness, all of the P pixels would cluster along the bolded line segments that lie on one of the two axes. In other words, each pixel would contain just one (pure) component, not a linear combination – the solution we desire if we wanted to look for maximum component separation. (In reality, if we were able to minimize $||\mathbf{t}||_0$, some pixels may contain predominantly one component, with some small mixture of the second, as depicted in this figure.)

of $||\mathbf{t}||_0$ is computationally difficult to solve, what about minimizing

$$||\mathbf{t}||_{1/p} = \left(\sum_{i,j} t_{ij}^p\right)^{1/p} \quad \text{for} \quad 0 < p < 1\,? \tag{A.5}$$

For example, Figure A.2 shows what $||\mathbf{t}||_{1/4}$ would conceptually look like. This appears to still be a good approximation to a feature selection model (matching each pixel to one component), in that many of the pixel points lie close to one of the component axes. However, this turns out to be also a non-convex problem. In fact, feature selection becomes more powerful as $p \to 0$, but global minimization of $||\mathbf{t}||_p$ for $0 < p < 1$ is a non-convex problem [**83**].

The next best option, then, would be to try to minimize the ℓ_1 norm, shown in Figure A.3(a). At first glance, it may not look like a very good feature selection model as many of the pixels appear as mixtures of components. However, as we increase the significance

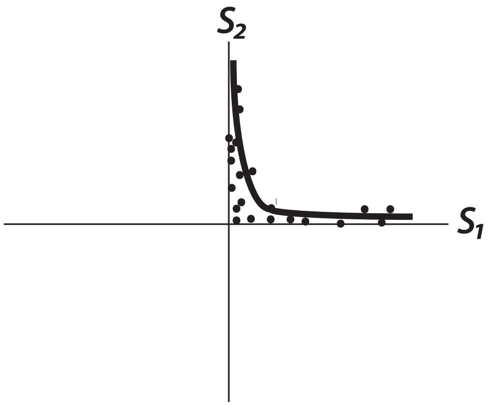
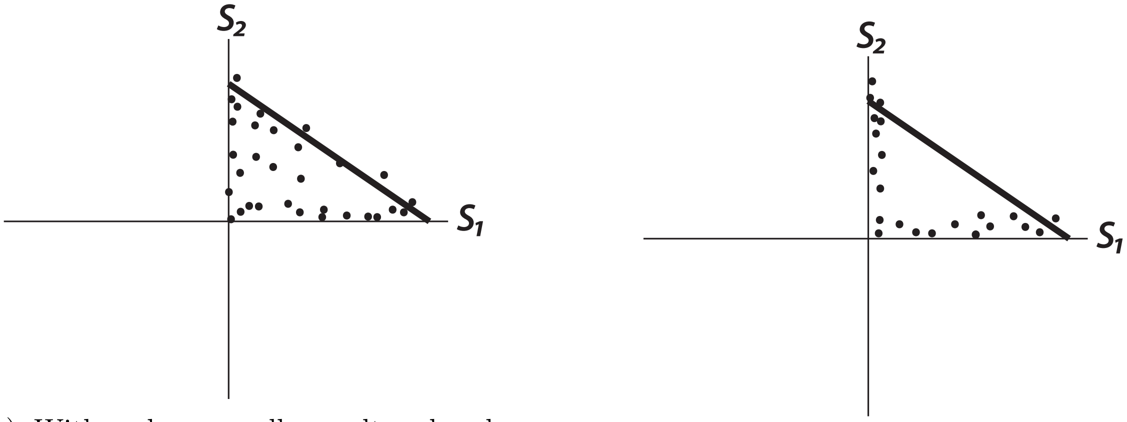


Figure A.2. Minimizing $||\mathbf{t}||_{1/4}= \left(\sum_{i,j} t_{ij}^{1/4}\right)^4$ looks like it might be a good approximation to the NP-hard counting problem of minimizing $||\mathbf{t}||_0$. However, this problem, too, turns out to be non-convex, as does the more general case for $||\mathbf{t}||_p$ where $0 < p < 1$.

of the ℓ_1 penalty term by increasing the sparseness regulariztion parameter $\lambda_{\mathbf{t}}$, we find that minimizing $||\mathbf{t}||_1$ does produce better feature selection, as shown in Figure A.3(b).



(a) With only a small penalty placed on the minimization $||\mathbf{t}||_1$ (small value for the sparseness regularization parameter $\lambda_{\mathbf{t}}$), the model produces some pixels with mixtures of components.

(b) As we increase the penalty placed on the minimization of $||\mathbf{t}||_1$ by increasing $\lambda_{\mathbf{t}}$, it becomes a better feature selection model, yielding more pixels with pure components.

Figure A.3. Comparison between small and large penalties imposed on the minimization of $||\mathbf{t}||_1$.

APPENDIX B

Partial derivatives of the cost function F

First, we write down some identities related to partial derivatives of matrix traces, which we will make use of when calculating partial derivatives of the cost function F:

$$\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A}) = \mathbf{A} \tag{B.1}$$

$$\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}) = \mathbf{A}\mathbf{X} + \mathbf{A}^{\mathrm{T}}\mathbf{X} \tag{B.2}$$

$$\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{B}) = \mathbf{B}\mathbf{A} \tag{B.3}$$

$$\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{B}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{B}) = \mathbf{X}\mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{X}\mathbf{B}\mathbf{B}^{\mathrm{T}} \tag{B.4}$$

Next, we expand the cost function F:

$$\begin{aligned} F &= \frac{1}{2}||\mathbf{D} - \boldsymbol{\mu}\mathbf{t}||_F^2 \\ &= \frac{1}{2}\mathrm{Tr}\left[(\mathbf{D} - \boldsymbol{\mu}\mathbf{t})^{\mathrm{T}}(\mathbf{D} - \boldsymbol{\mu}\mathbf{t})\right] \\ &= \frac{1}{2}\mathrm{Tr}(\mathbf{D}^{\mathrm{T}}\mathbf{D}) - \mathrm{Tr}(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D}) + \frac{1}{2}\mathrm{Tr}(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t}) \end{aligned} \tag{B.5}$$

Then, we can calculate the partial derivatives:

$$\begin{aligned} \frac{\partial F}{\partial \mathbf{t}} &= -\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D} + \frac{1}{2}\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t}\right) \\ &= -\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D} + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t} \end{aligned} \tag{B.6}$$

where we have used the identities (B.1) and (B.2);

$$\begin{aligned}\frac{\partial F}{\partial \boldsymbol{\mu}} &= -\mathbf{D}\mathbf{t}^{\mathrm{T}} + \frac{1}{2}\left(\boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}} + \boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}}\right) \\ &= -\mathbf{D}\mathbf{t}^{\mathrm{T}} + \boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}} \qquad \text{(B.7)}\end{aligned}$$

where we have used the identities (B.3) and (B.4).

APPENDIX C

Smoothness regularizer $J_{\boldsymbol{\mu}_{\mathrm{sm}}}(\boldsymbol{\mu})$

We desire our absorption spectra to be smooth functions with respect to energy E. Each spectrum is represented by a column in the $\boldsymbol{\mu}$ matrix. We can regularize the smoothness over each column by adding a regularization term $J_{\boldsymbol{\mu}_{\mathrm{sm}}}(\boldsymbol{\mu})$ (we will drop the subscript "sm" in the rest of this appendix for simplicity) to minimize the second derivative of each column in $\boldsymbol{\mu}$.

Since any given column μ_k in $\boldsymbol{\mu}$ is a measurement at N discrete energy points, we will use finite difference (shown below) to calculate the second derivative. Moreover, the energy points are not necessarily evenly spaced: during XANES or EXAFS data collection, more measurements need to be collected to resolve the steep rise of an absorption edge, and (especially in the case of EXAFS), more measurements also need to be collected in the extended energy region (beyond the absorption edge) to resolve the fine modulations arising from short-range interactions.

To find the finite difference second derivative of a column vector μ_k with respect to E (with unequal subintervals), we first write the Taylor series expansions for $\mu_{n+1,k}$ (forward expansion) and $\mu_{n-1,k}$ (backward expansion) around $\mu_{n,k}$ (n is the index label for energy, and hence also indexes the rows of $\boldsymbol{\mu}$):

$$\mu_{n+1,k} = \mu_{n,k} + \mu'_{n,k}(E_{n+1} - E_n) + \frac{1}{2}\mu''_{n,k}(E_{n+1} - E_n)^2 + \mathcal{O}((E_{n+1} - E_n)^3) \tag{C.1}$$

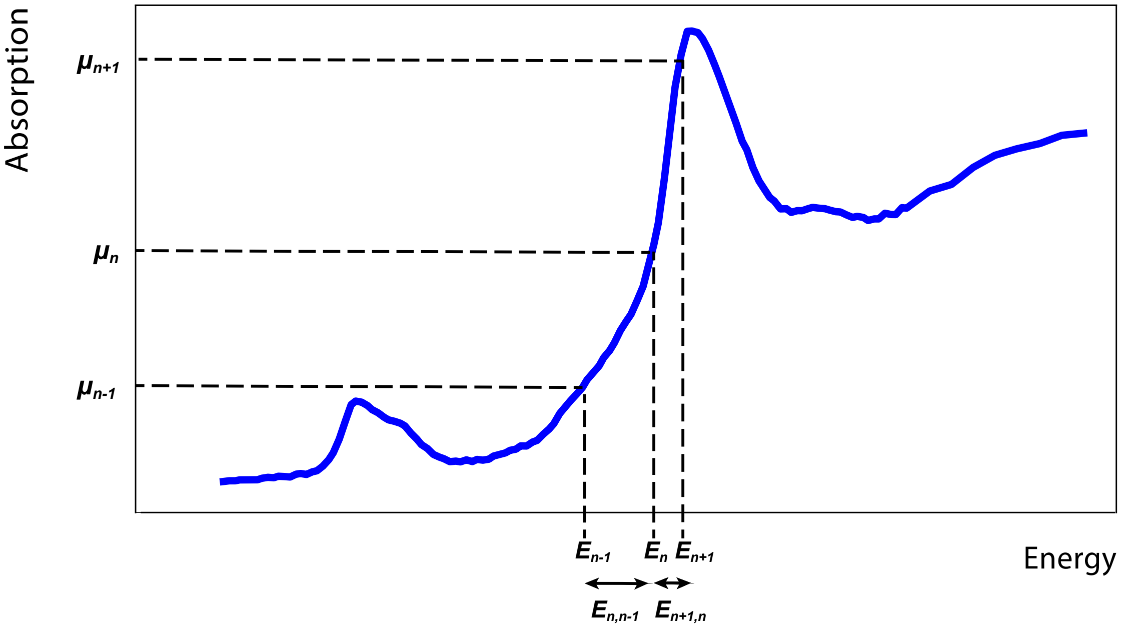


Figure C.1. The energy intervals in a measured absorption spectrum need not be divided evenly.

and

$$\mu_{n-1,k} = \mu_{n,k} - \mu'_{n,k}(E_n - E_{n-1}) + \frac{1}{2}\mu''_{n,k}(E_n - E_{n-1})^2 + \mathcal{O}((E_n - E_{n-1})^3), \tag{C.2}$$

where $\mu'_{n,k} = \left[\frac{\partial \mu_k}{\partial E}\right]_{n,k}$, and $\mu''_{n,k} = \left[\frac{\partial^2 \mu_k}{\partial E^2}\right]_{n,k}$.

We multiply Eq. (C.1) by $(E_n - E_{n-1})$, and Eq. (C.2) by $(E_{n+1} - E_n)$, then add the two equations to obtain:

$$\begin{aligned}\mu''_{n,k} = 2\Bigg[&\frac{\mu_{n+1,k}}{(E_{n+1} - E_n)(E_{n+1} - E_{n-1})} - \frac{\mu_{n,k}}{(E_{n+1} - E_n)(E_n - E_{n-1})} \\ &+ \frac{\mu_{n-1,k}}{(E_n - E_{n-1})(E_{n+1} - E_{n-1})}\Bigg]\end{aligned} \tag{C.3}$$

Eq. (C.3) is the centred difference second derivative for $\mu_{n,k}$, where $2 \leq n \leq N - 1$.

Assuming that indexing begins at 1 and ends at N, we should treat the cases $n = 1$ and $n = N$ separately.

For $n = 1$, we use the forward difference to find $\mu''_{1,k}$. First, we write the forward Taylor series expansions for $\mu_{2,k}$ and $\mu_{3,k}$:

$$\mu_{2,k} = \mu_{1,k} + \mu'_{1,k}(E_2 - E_1) + \mu''_{1,k}\frac{(E_2 - E_1)^2}{2} + \mathcal{O}((E_2 - E_1)^3) \tag{C.4a}$$

$$\mu_{3,k} = \mu_{1,k} + \mu'_{1,k}(E_3 - E_1) + \mu''_{1,k}\frac{(E_3 - E_1)^2}{2} + \mathcal{O}((E_3 - E_1)^3) \tag{C.4b}$$

We multiply Eq. (C.4a) by $\left(\frac{E_3 - E_1}{E_2 - E_1}\right)$, then subtract Eq. (C.4b) to obtain:

$$\mu''_{1,k} = 2\left[\frac{\mu_{3,k}}{(E_3 - E_2)(E_3 - E_1)} - \frac{\mu_{2,k}}{(E_3 - E_2)(E_2 - E_1)} + \frac{\mu_{1,k}}{(E_2 - E_1)(E_3 - E_1)}\right]. \tag{C.5}$$

For $n = N$, we use the backward difference to find $\mu''_{N,k}$. First, we write the backward Taylor series expansions for $\mu_{N-1,k}$ and $\mu_{N-2,k}$:

$$\mu_{N-1,k} = \mu_{N,k} - \mu'_{N,k}(E_N - E_{N-1}) + \mu''_{N,k}\frac{(E_N - E_{N-1})^2}{2} + \mathcal{O}((E_N - E_{N-1})^3) \tag{C.6a}$$

$$\mu_{N-2,k} = \mu_{N,k} - \mu'_{N,k}(E_N - E_{N-2}) + \mu''_{N,k}\frac{(E_N - E_{N-2})^2}{2} + \mathcal{O}((E_N - E_{N-2})^3) \tag{C.6b}$$

We multiply Eq. (C.6a) by $\left(\frac{E_N - E_{N-2}}{E_N - E_{N-1}}\right)$, then subtract Eq. (C.6b) to obtain:

$$\begin{split}\mu''_{N,k} = 2\Bigg[\frac{\mu_{N,k}}{(E_N - E_{N-1})(E_N - E_{N-2})} - \frac{\mu_{N-1,k}}{(E_N - E_{N-1})(E_{N-1} - E_{N-2})} \\ + \frac{\mu_{N-2,k}}{(E_{N-1} - E_{N-2})(E_N - E_{N-2})}\Bigg].\end{split} \tag{C.7}$$

For each column μ_k in $\boldsymbol{\mu}$, we have the second derivative $\mu''_{n,k} = \left[\frac{\partial^2 \mu_k}{\partial E^2}\right]_{n,k}$ calculated at each point n, for $1 < n < N$. To strive for a smooth spectral function, we would like to minimize this second derivative. We simply sum the absolute values of each entry and call this our smoothness regularizer $J_{\boldsymbol{\mu}}(\boldsymbol{\mu})$:

$$J_{\boldsymbol{\mu}}(\boldsymbol{\mu}) = \sum_{k=1}^{S} \sum_{n=1}^{N} \left| \left[\frac{\partial^2 \mu_k}{\partial E^2} \right]_{n,k} \right| . \tag{C.8}$$

Now, the addition to the NNMA update algorithm requires us to calculate the derivative of Eq. (C.8) with respect to $\mu_{i,j}$, *i.e.*, $\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{i,j}}$.

For $4 \leq i \leq N-3$ (we will treat the cases $i \in \{1, 2, 3, N-2, N-1, N\}$ separately):

$$\frac{\partial J_{\mu}(\boldsymbol{\mu})}{\partial \mu_{i,j}} = 2 \left[\underbrace{\frac{-1}{(E_{i+1} - E_i)(E_i - E_{i-1})}}_{\text{contribution when } n=i} + \underbrace{\frac{1}{(E_i - E_{i-1})(E_i - E_{i-2})}}_{n=(i-1)} + \underbrace{\frac{1}{(E_{i+1} - E_i)(E_{i+2} - E_i)}}_{n=(i+1)} \right] . \tag{C.9}$$

For $i = 1$:

$$\begin{aligned} \frac{\partial J_{\mu}(\boldsymbol{\mu})}{\partial \mu_{1,j}} &= 2 \left[\underbrace{\frac{1}{(E_2 - E_1)(E_3 - E_1)}}_{n=1} + \underbrace{\frac{1}{(E_2 - E_1)(E_3 - E_1)}}_{n=2} \right] \\ &= \frac{4}{(E_2 - E_1)(E_3 - E_1)} . \end{aligned} \tag{C.10}$$

For $i = 2$:

$$\begin{aligned} \frac{\partial J_{\mu}(\boldsymbol{\mu})}{\partial \mu_{2,j}} &= 2 \left[\underbrace{\frac{-1}{(E_2 - E_1)(E_3 - E_2)}}_{n=1} + \underbrace{\frac{-1}{(E_3 - E_2)(E_2 - E_1)}}_{n=2} + \underbrace{\frac{1}{(E_3 - E_2)(E_4 - E_3)}}_{n=3} \right] \\ &= 2 \left[\frac{-2}{(E_3 - E_2)(E_2 - E_1)} + \frac{1}{(E_3 - E_2)(E_4 - E_3)} \right] . \end{aligned} \tag{C.11}$$

For $i = 3$:

$$\begin{aligned}\frac{\partial J_\mu(\boldsymbol{\mu})}{\partial \mu_{3,j}} = 2 \Bigg[& \underbrace{\frac{1}{(E_3 - E_2)(E_3 - E_1)}}_{n=1} + \underbrace{\frac{1}{(E_3 - E_2)(E_3 - E_1)}}_{n=2} + \underbrace{\frac{-1}{(E_4 - E_3)(E_3 - E_2)}}_{n=3} \\ & + \underbrace{\frac{1}{(E_4 - E_3)(E_5 - E_3)}}_{n=4} \Bigg] \\ = 2 \Bigg[& \frac{2}{(E_3 - E_2)(E_2 - E_1)} + \frac{-1}{(E_4 - E_3)(E_3 - E_2)} + \frac{1}{(E_4 - E_3)(E_5 - E_3)} \Bigg].\end{aligned} \tag{C.12}$$

For $i = N - 2$:

$$\begin{aligned}\frac{\partial J_\mu(\boldsymbol{\mu})}{\partial \mu_{N-2,j}} = 2 \Bigg[& \underbrace{\frac{1}{(E_{N-2} - E_{N-3})(E_{N-2} - E_{N-4})}}_{n=N-3} + \underbrace{\frac{-1}{(E_{N-1} - E_{N-2})(E_{N-2} - E_{N-3})}}_{n=N-2} \\ & + \underbrace{\frac{1}{(E_{N-1} - E_{N-2})(E_N - E_{N-2})}}_{n=N-1} + \underbrace{\frac{1}{(E_{N-1} - E_{N-2})(E_N - E_{N-3})}}_{n=N} \Bigg] \\ = 2 \Bigg[& \frac{1}{(E_{N-2} - E_{N-3})(E_{N-2} - E_{N-4})} + \frac{-1}{(E_{N-1} - E_{N-2})(E_{N-2} - E_{N-3})} \\ & + \frac{2}{(E_{N-1} - E_{N-2})(E_{N-2} - E_{N-3})} \Bigg].\end{aligned} \tag{C.13}$$

For $i = N-1$:

$$\begin{aligned}\frac{\partial J_{\mu}(\boldsymbol{\mu})}{\partial \mu_{N-1,j}} &= 2\Bigg[\underbrace{\frac{1}{(E_{N-1}-E_{N-2})(E_{N-1}-E_{N-3})}}_{n=N-2} + \underbrace{\frac{-1}{(E_N-E_{N-1})(E_{N-1}-E_{N-2})}}_{n=N-1} \\ &\qquad + \underbrace{\frac{-1}{(E_N-E_{N-1}(E_{N-1}-E_{N-2})}}_{n=N}\Bigg] \\ &= 2\left[\frac{1}{(E_{N-1}-E_{N-2})(E_{N-1}-E_{N-3})} + \frac{-2}{(E_N-E_{N-1})(E_{N-1}-E_{N-2})}\right].\end{aligned} \tag{C.14}$$

For $i = N$:

$$\begin{aligned}\frac{\partial J_{\mu}(\boldsymbol{\mu})}{\partial \mu_{N,j}} &= 2\Bigg[\underbrace{\frac{1}{(E_N-E_{N-1})(E_N-E_{N-2})}}_{n=N-1} + \underbrace{\frac{1}{(E_N-E_{N-1})(E_N-E_{N-2})}}_{n=N}\Bigg] \\ &= \frac{4}{(E_N-E_{N-1})(E_N-E_{N-2})}.\end{aligned} \tag{C.15}$$

However, it seems that the smoothness regularizer Eq. (C.8) does not always result in a monotonically decreasing cost function. This may be due to the non-differentiable nature of the function at values around zero. Instead, we will sum the squares of each element in Eq. (C.8) and use this as our smoothness regularizer:

$$J_{\boldsymbol{\mu}}(\boldsymbol{\mu}) = \sum_{k=1}^{S}\sum_{n=1}^{N}\left[\frac{\partial^2 \mu_k}{\partial E^2}\right]_{n,k}^2. \tag{C.16}$$

Again, the addition to the NNMA update algorithm requires us to calculate the derivative of Eq. (C.16) with respect to $\mu_{i,j}$. First, we write out the expansion of $\left[\frac{\partial^2 \mu_k}{\partial E^2}\right]_{n,k}^2$; to simplify notation, we denote $E_{n+1,n} = (E_{n+1} - E_n)$, etc.

For $2 \leq n \leq N-1$ (expanding Eq. (C.3)):

$$\left[\frac{\partial^2 \mu_k}{\partial E^2}\right]^2_{n,k} = 4\left[\frac{\mu^2_{n+1,k}}{E^2_{n+1,n}E^2_{n+1,n-1}} + \frac{\mu^2_{n,k}}{E^2_{n+1,n}E^2_{n,n-1}} + \frac{\mu^2_{n-1,k}}{E^2_{n,n-1}E^2_{n+1,n-1}} - \frac{2\mu_{n+1,k}\mu_{n,k}}{E^2_{n+1,n}E_{n,n-1}E_{n+1,n-1}} \right.$$
$$\left. + \frac{2\mu_{n+1,k}\mu_{n-1,k}}{E_{n+1,n}E_{n,n-1}E^2_{n+1,n-1}} - \frac{2\mu_{n,k}\mu_{n-1,k}}{E_{n+1,n}E^2_{n,n-1}E_{n+1,n-1}}\right] \tag{C.17}$$

For $n = 1$ (expanding Eq. (C.5)):

$$\left[\frac{\partial^2 \mu_k}{\partial E^2}\right]^2_{1,k} = 4\left[\frac{\mu^2_{1,k}}{E^2_{2,1}E^2_{3,1}} + \frac{\mu^2_{2,k}}{E^2_{3,2}E^2_{2,1}} + \frac{\mu^2_{3,k}}{E^2_{3,2}E^2_{3,1}} - \frac{2\mu_{1,k}\mu_{2,k}}{E_{3,2}E^2_{2,1}E_{3,1}} + \frac{2\mu_{1,k}\mu_{3,k}}{E_{3,2}E_{2,1}E^2_{3,1}} - \frac{2\mu_{2,k}\mu_{3,k}}{E^2_{3,2}E_{2,1}E_{3,1}}\right] \tag{C.18}$$

For $n = N$ (expanding Eq. (C.7)):

$$\left[\frac{\partial^2 \mu_k}{\partial E^2}\right]^2_{N,k} = 4\left[\frac{\mu^2_{N,k}}{E^2_{N,N-1}E^2_{N,N-2}} + \frac{\mu^2_{N-1,k}}{E^2_{N,N-1}E^2_{N-1,N-2}} + \frac{\mu^2_{N-2,k}}{E^2_{N-1,N-2}E^2_{N,N-2}} \right.$$
$$- \frac{2\mu_{N,k}\mu_{N-1,k}}{E^2_{N,N-1}E_{N-1,N-2}E_{N,N-2}} + \frac{2\mu_{N,k}\mu_{N-2,k}}{E_{N,N-1}E_{N-1,N-2}E^2_{N,N-2}} \tag{C.19}$$
$$\left. - \frac{2\mu_{N-1,k}\mu_{N-2,k}}{E_{N,N-1}E^2_{N-1,N-2}E_{N,N-2}}\right]$$

Now, we sum the contributions from Eqs. (C.17), (C.18), and (C.19) to give $J_{\boldsymbol{\mu}}(\boldsymbol{\mu})$ and differentiate with respect to $\mu_{i,j}$.

For $4 \leq i \leq N-3$ (we will treat the cases $i \in \{1,2,3,N-2,N-1,N\}$ separately):

$$\begin{aligned}
\frac{\partial J_\mu(\boldsymbol{\mu})}{\partial \mu_{i,j}} = 4 \Bigg[& \underbrace{\frac{2\mu_{i,j}}{E_{i+1,i}^2 E_{i,i-1}^2} - \frac{2\mu_{i+1,j}}{E_{i+1,i}^2 E_{i,i-1} E_{i+1,i-1}} - \frac{2\mu_{i-1,j}}{E_{i+1,i} E_{i,i-1}^2 E_{i+1,i-1}}}_{\text{contribution when } n=i} \\
& + \underbrace{\frac{2\mu_{i,j}}{E_{i,i-1}^2 E_{i,i-2}^2} - \frac{2\mu_{i-1,j}}{E_{i,i-1}^2 E_{i-1,i-2} E_{i,i-2}} + \frac{2\mu_{i-2,j}}{E_{i,i-1} E_{i-1,i-2} E_{i,i-2}^2}}_{n=(i-1)} \\
& + \underbrace{\frac{2\mu_{i,j}}{E_{i+1,i}^2 E_{i+2,i}^2} + \frac{2\mu_{i+2}}{E_{i+2,i+1} E_{i+1,i} E_{i+2,i}^2} - \frac{2\mu_{i+1}}{E_{i+2,i+1} E_{i+1,i}^2 E_{i+2,i}}}_{n=(i+1)} \Bigg] . \qquad \text{(C.20)} \\
= 8 \Bigg[& \mu_{i-2,j} \left(\frac{1}{E_{i,i-1} E_{i-1,i-2} E_{i,i-2}^2} \right) \\
& + \mu_{i-1,j} \left(\frac{-1}{E_{i+1,i} E_{i,i-1}^2 E_{i+1,i-1}} + \frac{-1}{E_{i,i-1}^2 E_{i-1,i-2} E_{i,i-2}} \right) \\
& + \mu_{i,j} \left(\frac{1}{E_{i+1,i}^2 E_{i,i-1}^2} + \frac{1}{E_{i,i-1}^2 E_{i,i-2}^2} + \frac{1}{E_{i+1,i}^2 E_{i+2,i}^2} \right) \\
& + \mu_{i+1,j} \left(\frac{-1}{E_{i+1,i}^2 E_{i,i-1} E_{i+1,i-1}} + \frac{-1}{E_{i+2,i+1} E_{i+1,i}^2 E_{i+2,i}} \right) \\
& + \mu_{i+2,j} \left(\frac{1}{E_{i+2,i+1} E_{i+1,i} E_{i+2,i}^2} \right) \Bigg]
\end{aligned}$$

For $i = 1$:

$$\begin{aligned}
\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{1,j}} = 4 \Bigg[& \underbrace{\frac{2\mu_{1,j}}{E_{2,1}^2 E_{3,1}^2} - \frac{2\mu_{2,j}}{E_{3,2} E_{2,1}^2 E_{3,1}} + \frac{2\mu_{3,j}}{E_{3,2} E_{2,1} E_{3,1}^2}}_{n=1} \\
& + \underbrace{\frac{2\mu_{1,j}}{E_{2,1}^2 E_{3,1}^2} + \frac{2\mu_{3,j}}{E_{3,2} E_{2,1} E_{3,1}^2} - \frac{2\mu_{2,j}}{E_{3,2} E_{2,1}^2 E_{3,1}}}_{n=2} \Bigg] \qquad \text{(C.21)} \\
= 16 & \left[\frac{\mu_{1,j}}{E_{2,1}^2 E_{3,1}^2} - \frac{\mu_{2,j}}{E_{3,2} E_{2,1}^2 E_{3,1}} + \frac{\mu_{3,j}}{E_{3,2} E_{2,1} E_{3,1}^2} \right] .
\end{aligned}$$

For $i = 2$:

$$\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{2,j}} = 4 \left[\underbrace{\frac{2\mu_{2,j}}{E_{3,2}^2 E_{2,1}^2} - \frac{2\mu_{1,j}}{E_{3,2} E_{2,1}^2 E_{3,1}} - \frac{2\mu_{3,j}}{E_{3,2}^2 E_{2,1} E_{3,1}}}_{n=1} \right.$$

$$+ \underbrace{\frac{2\mu_{2,j}}{E_{3,2}^2 E_{2,1}^2} - \frac{2\mu_{3,j}}{E_{3,2}^2 E_{2,1} E_{3,1}} - \frac{2\mu_{1,j}}{E_{3,2} E_{2,1}^2 E_{3,1}}}_{n=2} \tag{C.22}$$

$$\left. + \underbrace{\frac{2\mu_{2,j}}{E_{3,2}^2 E_{4,2}^2} + \frac{2\mu_{4,j}}{E_{4,3} E_{3,2} E_{4,2}^2} - \frac{2\mu_{3,j}}{E_{4,3} E_{3,2}^2 E_{4,2}}}_{n=3} \right]$$

$$= 8 \left[\mu_{1,j} \left(\frac{-2}{E_{3,2} E_{2,1}^2 E_{3,1}} \right) + \mu_{2,j} \left(\frac{2}{E_{3,2}^2 E_{2,1}^2} + \frac{1}{E_{3,2}^2 E_{4,2}^2} \right) \right.$$

$$\left. + \mu_{3,j} \left(\frac{-2}{E_{3,2}^2 E_{2,1} E_{3,1}} + \frac{-1}{E_{4,3} E_{3,2}^2 E_{4,2}} \right) + \mu_{4,j} \left(\frac{1}{E_{4,3} E_{3,2} E_{4,2}^2} \right) \right]$$

For $i = 3$:

$$
\begin{aligned}
\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{3,j}} = 4 \Bigg[& \underbrace{\frac{2\mu_{3,j}}{E_{3,2}^2 E_{3,1}^2} + \frac{2\mu_{1,j}}{E_{3,2} E_{2,1} E_{3,1}^2} - \frac{2\mu_{2,j}}{E_{3,2}^2 E_{2,1} E_{3,1}}}_{n=1} \\
& + \underbrace{\frac{2\mu_{3,j}}{E_{3,2}^2 E_{3,1}^2} - \frac{2\mu_{2,j}}{E_{3,2}^2 E_{2,1} E_{3,1}} + \frac{2\mu_{1,j}}{E_{3,2} E_{2,1} E_{3,1}^2}}_{n=2} \\
& + \underbrace{\frac{2\mu_{3,j}}{E_{4,3}^2 E_{3,2}^2} - \frac{2\mu_{4,j}}{E_{4,3}^2 E_{3,2} E_{4,2}} - \frac{2\mu_{2,j}}{E_{4,3} E_{3,2}^2 E_{4,2}}}_{n=3} \\
& + \underbrace{\frac{2\mu_{3,j}}{E_{4,3}^2 E_{5,3}^2} + \frac{2\mu_{5,j}}{E_{5,4} E_{4,3} E_{5,3}^2} - \frac{2\mu_{4,j}}{E_{5,4} E_{4,3}^2 E_{5,3}}}_{n=4} \Bigg] \\
= 8 \Bigg[& \mu_{1,j} \left(\frac{2}{E_{3,2} E_{2,1} E_{3,1}^2} \right) + \mu_{2,j} \left(\frac{-2}{E_{3,2}^2 E_{2,1} E_{3,1}} + \frac{-1}{E_{4,3} E_{3,2}^2 E_{4,2}} \right) \\
& + \mu_{3,j} \left(\frac{2}{E_{3,2}^2 E_{3,1}^2} + \frac{1}{E_{4,3}^2 E_{3,2}^2} + \frac{1}{E_{4,3}^2 E_{5,3}^2} \right) + \mu_{4,j} \left(\frac{-1}{E_{4,3}^2 E_{3,2} E_{4,2}} + \frac{-1}{E_{5,4} E_{4,3}^2 E_{5,3}} \right) \\
& + \mu_{5,j} \left(\frac{1}{E_{5,4} E_{4,3} E_{5,3}^2} \right) \Bigg]
\end{aligned}
\tag{C.23}
$$

For $i = N - 2$:

$$
\begin{aligned}
&\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{N-2,j}} \\
&= 4 \Bigg[\underbrace{\frac{2\mu_{N-2,j}}{E^2_{N-2,N-3}E^2_{N-2,N-4}} - \frac{2\mu_{N-3,j}}{E^2_{N-2,N-3}E_{N-3,N-4}E_{N-2,N-4}} + \frac{2\mu_{N-4,j}}{E_{N-2,N-3}E_{N-3,N-4}E^2_{N-2,N-4}}}_{n=N-3} \\
&\quad + \underbrace{\frac{2\mu_{N-2,j}}{E^2_{N-1,N-2}E^2_{N-2,N-3}} - \frac{2\mu_{N-1,j}}{E^2_{N-1,N-2}E_{N-2,N-3}E_{N-1,N-3}} - \frac{2\mu_{N-3,j}}{E_{N-1,N-2}E^2_{N-2,N-3}E_{N-1,N-3}}}_{n=N-2} \\
&\quad + \underbrace{\frac{2\mu_{N-2,j}}{E^2_{N-1,N-2}E^2_{N,N-2}} + \frac{2\mu_{N,j}}{E_{N,N-1}E_{N-1,N-2}E^2_{N,N-2}} - \frac{2\mu_{N-1,j}}{E_{N,N-1}E^2_{N-1,N-2}E_{N,N-2}}}_{n=N-1} \\
&\quad + \underbrace{\frac{2\mu_{N-2,j}}{E^2_{N-1,N-2}E^2_{N,N-2}} + \frac{2\mu_{N,j}}{E_{N,N-1}E_{N-1,N-2}E^2_{N,N-2}} - \frac{2\mu_{N-1,j}}{E_{N,N-1}E^2_{N-1,N-2}E_{N,N-2}}}_{n=N} \Bigg] \\
&= 8 \Bigg[\mu_{N-4,j} \left(\frac{1}{E_{N-2,N-3}E_{N-3,N-4}E^2_{N-2,N-4}} \right) \\
&\quad + \mu_{N-3,j} \left(\frac{-1}{E^2_{N-2,N-3}E_{N-3,N-4}E_{N-2,N-4}} + \frac{-1}{E_{N-1,N-2}E^2_{N-2,N-3}E_{N-1,N-3}} \right) \\
&\quad + \mu_{N-2,j} \left(\frac{1}{E^2_{N-2,N-3}E^2_{N-2,N-4}} + \frac{1}{E^2_{N-1,N-2}E^2_{N-2,N-3}} + \frac{2}{E^2_{N-1,N-2}E^2_{N,N-2}} \right) \\
&\quad + \mu_{N-1,j} \left(\frac{-1}{E^2_{N-1,N-2}E_{N-2,N-3}E_{N-1,N-3}} + \frac{-2}{E_{N,N-1}E^2_{N-1,N-2}E_{N,N-2}} \right) \\
&\quad + \mu_{N,j} \left(\frac{2}{E_{N,N-1}E_{N-1,N-2}E^2_{N,N-2}} \right) \Bigg]
\end{aligned}
\tag{C.24}
$$

For $i = N - 1$:

$$\begin{aligned}
&\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{N-1,j}} \\
&= 4\left[\underbrace{\frac{2\mu_{N-1,j}}{E^2_{N-1,N-2}E^2_{N-1,N-3}} - \frac{2\mu_{N-2,j}}{E^2_{N-1,N-2}E_{N-2,N-3}E_{N-1,N-3}} + \frac{2\mu_{N-3,j}}{E_{N-1,N-2}E_{N-2,N-3}E^2_{N-1,N-3}}}_{n=N-2}\right. \\
&\qquad + \underbrace{\frac{2\mu_{N-1,j}}{E^2_{N,N-1}E^2_{N-1,N-2}} - \frac{2\mu_{N,j}}{E^2_{N,N-1}E_{N-1,N-2}E_{N,N-2}} - \frac{2\mu_{N-2,j}}{E_{N,N-1}E^2_{N-1,N-2}E_{N,N-2}}}_{n=N-1} \\
&\qquad \left. + \underbrace{\frac{2\mu_{N-1,j}}{E^2_{N,N-1}E^2_{N-1,N-2}} - \frac{2\mu_{N,j}}{E^2_{N,N-1}E_{N-1,N-2}E_{N,N-2}} - \frac{2\mu_{N-2,j}}{E_{N,N-1}E^2_{N-1,N-2}E_{N,N-2}}}_{n=N}\right] \\
&= 8\left[\mu_{N-3,j}\left(\frac{1}{E_{N-1,N-2}E_{N-2,N-3}E^2_{N-1,N-3}}\right)\right. \\
&\qquad + \mu_{N-2,j}\left(\frac{-1}{E^2_{N-1,N-2}E_{N-2,N-3}E_{N-1,N-3}} + \frac{-2}{E_{N,N-1}E^2_{N-1,N-2}E_{N,N-2}}\right) \\
&\qquad \left. + \mu_{N-1,j}\left(\frac{1}{E^2_{N-1,N-2}E^2_{N-1,N-3}} + \frac{2}{E^2_{N,N-1}E^2_{N-1,N-2}}\right) + \mu_{N,j}\left(\frac{-2}{E^2_{N,N-1}E_{N-1,N-2}E_{N,N-2}}\right)\right]
\end{aligned} \tag{C.25}$$

For $i = N$:

$$\begin{aligned}
\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{N,j}} &= 4\left[\underbrace{\frac{2\mu_{N,j}}{E^2_{N,N-1}E^2_{N,N-2}} - \frac{2\mu_{N-1,j}}{E^2_{N,N-1}E_{N-1,N-2}E_{N,N-2}} + \frac{2\mu_{N-2,j}}{E_{N,N-1}E_{N-1,N-2}E^2_{N,N-2}}}_{n=N-1}\right. \\
&\qquad \left. + \underbrace{\frac{2\mu_{N,j}}{E^2_{N,N-1}E^2_{N,N-2}} - \frac{2\mu_{N-1,j}}{E^2_{N,N-1}E_{N-1,N-2}E_{N,N-2}} + \frac{2\mu_{N-2,j}}{E_{N,N-1}E_{N-1,N-2}E^2_{N,N-2}}}_{n=N}\right] \\
&= 16\left[\frac{\mu_{N-2,j}}{E_{N,N-1}E_{N-1,N-2}E^2_{N,N-2}} - \frac{\mu_{N-1,j}}{E^2_{N,N-1}E_{N-1,N-2}E_{N,N-2}} + \frac{\mu_{N,j}}{E^2_{N,N-1}E^2_{N,N-2}}\right].
\end{aligned} \tag{C.26}$$

Now, to recap: we (finally!) have an explicit form for $J_{\boldsymbol{\mu}}(\boldsymbol{\mu})$ given by Eq. (C.16), along with Eqs. (C.17), (C.18), and (C.19). This is the smoothness regularizer to be added to the cost function F. We also have an explicit form for $\frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \mu_{i,j}}$, given by Eqs. (C.20 - C.26). This is the addition to the NNMA update rule for $\boldsymbol{\mu}$, which is now given by:

$$\boldsymbol{\mu}_{i,j} \leftarrow \boldsymbol{\mu}_{i,j} \left[\frac{(\mathbf{D}\mathbf{t}^T)_{i,j}}{(\boldsymbol{\mu}\mathbf{t}\mathbf{t}^T)_{i,j} + \lambda_{\boldsymbol{\mu}} \frac{\partial J_{\boldsymbol{\mu}}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}_{i,j}}} \right] \tag{C.27}$$