UNIVERSITY OF CALIFORNIA
Los Angeles

Energy Models for Signal Processing and Matrix Factorization

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Travis Robert Meyer

2017

ABSTRACT OF THE DISSERTATION

Energy Models for Signal Processing
and Matrix Factorization

by

Travis Robert Meyer
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2017
Professor Andrea L. Bertozzi, Chair

In this work, we present a variety of energy-based methods that are solutions to problems in the fields of microscopy, hyperspectral and medical imaging, and data mining. These solutions are formulated from the perspective of extremization an energy function capturing deviation of the solution from observations and desirable properties. First we present new methods for improving imaging acquisition rates of atomic force microscopes. We propose and experimentally demonstrate image inpainting as a way to liberate scanner position limitations thereby enabling faster scans. Traditionally the scanner takes measurements in a raster pattern; in this work, we demonstrate that high-quality surface reproduction is attainable by sampling with non-raster patterns using variational image inpainting. With non-raster scan patterns existing thermomechanical drift error removal approaches no longer can be used. We propose a solution to this task with a highly effective corrective technique that utilize points of self-intersection. Our model only requires a few points of self-intersection that have minimal impact on scan time. Our correction model is potentially numerically unstable in some special, though easy to produce, cases. We propose a fitness based on analysis of the model energy that quantifies how well our method will perform for a given scan path. With minor experimental design modifications, often resulting simply from uncertainties in the scanner positioning, this fitness can be drastically increased and issues thereby alleviated. Due to its desirable properties, we focus specifically on improving the Archimedean spiral

scan. By considering basic limitations of the scanner's tip speed and resonant frequency, we derive the parametrization that exactly obeys limitations while minimizing total scan time. With small and reasonable approximations the form of this scan becomes analytically simple to state and easy to implement in practice. We defend this optimal parameterization against other choices from the perspectives of scan time, scanner limitations, and sampling distribution uniformity.

In the area of medical imaging we address the issue of signal cleaning for simultaneous electroencephalographic and functional magnetic resonance imaging. During acquisition dominant signals are produced through the ballistocardiographic effects that have challenge variability over time. Noting some properties of the signals, we propose applying an existing model known as low-rank + sparse matrix decomposition. We performed experiments with twenty individuals in simultaneous capture to observe decreases in alpha-band neural activity following Gabor flashes and find that the proposed method improves signal cleaning results considerably when compared to an existing method known as independent component analysis. In the domain of hyperspectral unmixing we address the problem of unmixing with spectral variability. We propose and study using social sparsity to enforce sparsity assumptions in the context of existing models that extract per-material endmember bundles. In a trio of experiments, two quantitative and one qualitative, we demonstrate that social sparsity - in particular group lasso - improves the solution.

In the final chapter of this work we investigate the recently popular machine learning problem of topic modeling. We present two models for solving this problem - latent Dirichlet allocation and non-negative matrix factorization - in their original forms, review the literature, and present what is known about the analytic relationship they share. In practice, because the problems are non-convex, the inference or optimization technique plays a role in solution quality. We therefore also summarize three popular algorithms for these models and frame the algorithms themselves in a common variational setting specific to the topic modeling problem. In addition to contributing this perspective for the models and algorithms together, we experimentally demonstrate differences in performance for the methods

as well as practical topic model results. The final contribution of this work is two metrics for studying the distributional properties of topics extracted from documents with additional information e.g. time or location. We study these metrics with a geotagged Twitter data set taken from Madrid throughout 2011 and find that these simple metrics provide a useful summary for topics and can significantly simplify the initial process of studying topic model results when the number of topics is large.

The dissertation of Travis Robert Meyer is approved.

Luminita A. Vese

Stanley J. Osher

Mark S. Cohen

Andrea L. Bertozzi, Committee Chair

University of California, Los Angeles

2017

for Mary and Dennis

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work was accomplished through significant contributions from many collaborators, advisers, family, and friends. For my early introduction to applied mathematics I am grateful for the patience of and help from Fred Park, Todd Wittman, and David Gieseker who provided me with early experiences in research during my undergraduate career. Todd Wittman, in particular, always kept me busy experimenting with mathematical models. His availability, discussions, and feedback allowed me to leave my undergraduate program with considerable experience in interesting areas for which I am extremely grateful.

My first experience in a large collaborative research effort was during the summer program in 2011. There I had the pleasure of working with Rodrigo Farnham and Nen Huynh who both worked with great dedication to develop our models. We three worked under the guidance of Jen-Mei Chang, Christoph Brune, and Alex Chen. In addition to being highly skilled mathematicians they were also kind colleagues, open to discussion, and dedicated to helping in any way at any time. This project would also have not met with success without the extraordinarily hard-working and skilled collaborators at Lawrence Berkeley National Laboratory: Paul Ashby, Dominik Ziegler, and Andreas Amrein. They certainly are masters of their field and go above and beyond to acquire the results we needed to demonstrate the emphasis of our papers and effectiveness of our models.

My work evaluating the LR+SD model for hybrid medical imaging would not be possible without the inspiration from and guidance of Jerome Gilles, as well as the dedicated work of Pamela Douglas collecting experimental data we needed and applying the existing ICA technique. Additionally, her knowledge of the field she shared through our discussions was invaluable for me to correctly analyse the data. To achieve the hyperspectral imaging work, Jocelyn Chanussot came to me with the idea of using social sparsity and Lucas Drumetz with plenty of synthetic data from his own research to use for quantifying our model's performance. Without their support starting the project, and their very detailed and expansive

explanations every step of the way, my work developing and evaluating our model would certainly never have come to be.

For helping me understand better topic models and the algorithms used I would like to recognize Lawrence Carin for making himself and his group available to me during my visit. In particular David Carlson was helpful, knowledgeable, and provided me with key insights for bridging the gap into the statistics literature during our tea breaks.

The contributions in this work applying topic models to Madrid are the result of work involving a few people. Miguel Camacho-Collados has been a wonderful person to discuss the many initial results we obtained with, an invaluable resource of knowledge about Spain, and the source for collecting our Twitter data. Hao Li has contributed a great deal of his time and energy to producing topic modeling results with my code on the Twitter data, as well as producing a great deal of data analysis that got us to our pending publication. Both Daniel Balague and Katie Khuu are responsible for creating the user interface included in our manuscript that makes the result of our study significantly easier to interact with for practitioners. Jeff Brantingham contributed his knowledge and time to help clean up and compose our manuscript. I am grateful for all of their hard work making the topic models I implemented and the metrics investigated by myself and Andrea Bertozzi applicable in a context and approachable with an interface for the world to use.

I would like to thank Olga Radko for the extraordinary work she does organizing the Los Angeles Math Circle I was able to participate in. The program is valuable for everyone involved and provides many humbling experiences, now memories, spreading the joy of mathematics to the next generations.

Among everyone I have encountered in my academic career I would like to acknowledge above all Andrea Bertozzi as both an interesting and highly knowledgeable colleague and an unbelievably supportive person. If not for her seemingly endless desire to be actively involved in helping others achieve their goals I would certainly not be where I am. Through her indirectly I was able to begin in research as an undergraduate, and through her directly I was able to continue doing research to this day.

To my committee, Luminita Vese, Stan Osher, Mark Cohen, and Andrea Bertozzi I am grateful for the contribution of time to review my work and provide feedback. Their willingness to follow my work across the many areas it touches comes with great appreciation on my part.

Contributions in this work were made possible by a multitude of funding sources: the National Science Foundation Cyber Enabled Discovery and Innovation contract no. 940417, the W M Keck Foundation, the German Research Foundation DFG, project BU 2327/6-1, the Office of Science and Office of Basic Energy Sciences, US Department of Energy contract no. DE-AC02-05CH11231, the European Research Council (ERC) grant ERC AdG-2012-320684 CHESS, NSF grants DMS-1118971 and DMS-1417674, ONR grant N000141210838, UC Lab Fees Research grant 12-LR-236660, DGA 2015 60 0012 00.470.75.01, Department of Defense ONR grant N00014-16-1-2119, and NSF grant DMS-1417674.

VITA

2008	A.A. Mathematics Bakersfield College
2011	B.S. Applied Mathematics, *Daus Prize Recipient* B.A. Physics University of California, Los Angeles
2011	Student UCLA Applied Mathematics REU *Atomic Force Microscopy*
2013 - 2015	Assistant UCLA Los Angeles Math Circle
2014	Teaching Assistant PIC 10A *Introduction to Programming*
2014	Co-Instructor Math 199 *Machine Learning*
2015	Teaching Assistant Math 191 *Machine Learning*
2015	Graduate Student Mentor UCLA Appled Mathematics REU *Large Data II*
2017	*(Expected)* Ph.D. Mathematics University of California, Los Angeles

PUBLICATIONS

Height Drift Correction in Non-Raster Atomic Force Microscopy by T. R. Meyer, D. Ziegler, C. Brune, A. Chen, R. Farnham, N. Huynh, J-M. Chang, A. Bertozzi, P.D. Ashby; Ultramicroscopy Volume 137, February 2014, Pages 48-54.

Improved accuracy and speed in scanning probe microscopy by image reconstruction from non-gridded position sensor data by D. Ziegler, T. Meyer, R. Farnham, Ch. Brune, A. L. Bertozzi, P. Ashby; Nanotechnology. 2013 Aug 23;24(33):335703.

Coastal Bathymetry from Sparse Level Curves by T. Meyer, T. Wittman; IEEE IGARSS 2014.

Leveraging Sparsity: A Low-Rank + Sparse Decomposition (LR+SD) Method for Automatic EEG Artifact Removal by J. Gilles, T. Meyer, P. K. Douglas; STMI 2014.

Hyperspectral Unmixing with Material Variability using Social Sparsity by T. R. Meyer, L. Drumetz, J. Chanussot, A. L. Bertozzi, C. Jutten; IEEE ICIP 2016.

Ideal Scan Path for High-Speed Scanning Atomic Force Microscopy by D. Ziegler, T. R. Meyer, A. Amrein, A. Bertozzi, P. D. Ashby; IEEE/ASME Trans. Mechatronics, 2016.

A year in Madrid as described through the analysis of geotagged Twitter data by T. R. Meyer, D. Balagu, M. Camacho-Collados, H. Li, K. Khuu, P. J. Brantingham, A. L. Bertozzi; in preparation.

Approaches to Topic Modeling by T. R. Meyer, A. L. Bertozzi; in preparation.

CHAPTER 1

Introduction

In this work, we present a variety of energy-based methods that are solutions to problems in the fields of microscopy, hyperspectral and medical imaging, and data mining. Energy methods, or variational methods, are based around the construction of an energy functional that can be applied to any potential solution to measure how well the solution models observed data. The solution that best minimizes this functional is sought as it represents, depending on the setting, a cleaned version of or optimal representation for the data. In this work such techniques are employed at every step. With problems ranging from filling in missing information in images to finding trends in a book collection this manuscript demonstrates a variety of successful and original ways to design, build on, or employ energy-based models for general signal processing and data mining tasks.

This manuscript is divided into three parts. First, in chapter 2 we present results for fast imaging with an atomic force microscope (AFM) [BQG86]. AFMs typically capture images on the order of minutes which makes it time consuming [HSF06] and difficult to observe dynamic processes [KYI10]. The AFM captures images serially by measuring one point on a sample surface at a time through physical contact with a nano-scale tip known as the cantilever. This sequence of point measurements must then be used to complete a grid-based image of the sample. In chapter 2 we demonstrate that speed can be improved by departing from standard raster, or grid, acquisition with tight controller feedback; instead, we propose using sensor information with weaker control of the cantilever and image inpainting techniques [BVS03, AK06, BSC00, CMS98, GO09, CWT11] to generate a topograph from sensor data. This allows for new scan patterns to be used with more desirable characteristics

for the scanner such as finite tip acceleration, thereby allowing the scanner to perform well at higher capture rates.

Due to thermal changes in the AFM during acquisition the measurement taken by the AFM drifts with time [CSD01]. This drift, due to thermomechanical changes, is typically corrected using approaches specific to raster data. With the non-raster paths a more general approach is needed to correct this drift. To extract the contribution of this signal we propose using a small number of points of self-intersection in a non-raster scan path to measure the drift component. Without drift, measuring the same location twice produces the same measurement. Any difference, therefore, can be used in the variational problem we propose to solve for the smooth drift. In some situations the proposed model fails. This results from path self-intersection points that are invariant to smooth perturbations of the AFM signal. To quantify the susceptibility to this problem and guarantee that our model can correct drift present in a scan acquired with a particular path, we also present a fitness quantity that, when large, guarantees smooth drift components will be corrected by our approach.

In the remainder of this first chapter we investigate a scan pattern, the Archimedean spiral, due to its desirable surface coverage and frequency profile. Using this scan pattern there remains a decision to be made about the parameterization of the curve. In essence, the spiral pattern determines the path along which the AFM will take measurements but the speed at each point is yet to be determined. Two particular solutions - one using constant angular velocity and another constant linear velocity - are options each obeying one physical limitation of the AFM yet violating another. We investigate advantages and disadvantages then propose an alternative scan that is optimal in the sense that it completes the scan in as little time as possible while obeying all relevant AFM limitations. All techniques described in this chapter are presented with various experimental results capturing the efficacy of non-raster scanning with inpanting and the proposed drift correction, as well as intricate experiments demonstrating the impacts of different spiral scan parameterizations.

Next, in chapter 3 we propose solutions to problems in both the fields of medical and hyperspectral imaging based on entirely different data but utilizing very similar underlying

matrix factorization models. In the first half of this chapter we apply an existing technique known as low rank + sparse decomposition (LR+SD) [LCM10] to separate undesirable signals from an electroencephalogram (EEG) [DBM13] captured during a functional magnetic resonance imaging. The specific signal contribution we consider that is challenging to separate is due to the ballistocardiogram (BCG) [DSS07]. This signal results from the motion of EEG electrodes in the strong MRI fields as a result of blood flow in the patient. We demonstrate experimentally that the LR+SD model effectively captures the BCG component in the low-rank component and the brainwave activity in the sparse component when compared to another technique known as independent component analysis.

In the second half of chapter 3 we consider the challenge of hyperspectral unmixing [KM02] with material variability [ZH14, HDT14]. Hyperspectral images are images captured with many tens to hundreds of color, or spectral, bands. The unmixing problem seeks to determine what materials are present in the image and in what quantity, or abundance, each material is present in each pixel. Variability is an additional complication that exists when the same material may present with different spectral signatures in the same image. In this final half of chapter 3 we take an existing pipeline for solving this problem and study the impact of social sparsity terms [KSD13] as an enforcement mechanism for desirable sparsity properties in the solution. Using two synthetic data sets with known abundance solutions and one qualitative data set we investigate the impact three types of social, or group, norms have on the final result.

In the final chapter of this manuscript, chapter 4, we thoroughly study a popular data mining task known as topic modeling [Ble12]. Topic models are tools for extracting latent, or hidden, trends in large text document collections. We investigate two popular techniques, latent Dirichlet allocation [BNJ03] and non-negative matrix factorization [LS99], that reside in different fields with seemingly disparate formulations. First we review existing literature around these models including their relationship when formulated in the energy framework. Next, we consider three popular models for finding solutions to these models and present a general setting within which the algorithms may be compared. The topic modeling problem

is generally non-convex and therefore algorithm decisions can be as important as model decisions. We demonstrate this fact with numerical results that exemplify implications of our analysis and demonstrate performance differences for various models and algorithms.

For large document collections a large number of topics may be extracted, potentially numbering in the hundreds. In such situations a challenge that arises is further characterizing the information captured by each topic. We propose two metrics that apply to text document collections for which each document has additional information such as time or location in space. These metrics can be used to quantify automatically the distributional properties of each topic in this additional information space. We demonstrate these metrics using a collection of Tweets that have both known location in space and creation time. This implies for each topic four values - one for each proposed metric in time and in space. Chapter 4 concludes with a discussion of topic model and metric results extracted from all geo-located tweets in the Madrid area during 2011, a year of significant protests and elections.

In this work, matricies are denoted with bold capital letters such as $\mathbf{X}$, the matrix transpose with a dagger such as $\mathbf{X}^\dagger$, and the Frobenius inner product by

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i,j} \mathbf{X}_{i,j} \mathbf{Y}_{i,j}.$$

Finally, where relevant differentiation with respect to time is denoted by at dot such as $\dot{x}(t)$.

CHAPTER 2

Non-Raster Atomic Force Microscopy

2.1 Preliminaries

2.1.1 Experimental Apparatus

An atomic force microscope (AFM) [BQG86] is a device used to measure the surface topology and properties at the nano-scale, or even smaller scales [GMM09]. The device consists of a surface onto which is placed a sample – for example nano-structures or strands of DNA. This sample rests typically on a device that both insulates the sample from vibrations originating in the surrounding environment and that, through piezoelectric actuators, is capable of being translated in the plane perpendicular to gravity. Whereas a standard photon microscope observes the sample using light, the atomic force microscope analyses the sample using physical contact between the sample and a small probe known as the "cantilever". This cantilever, nearly invisible to the naked eye, is composed of a horizontal "beam" at one end attached to the larger apparatus and on the other a suspended "tip" that forms a needle-like point below to contact the sample. During the collection process, this cantilever is lowered until the downward-pointing needle tip comes in contact with the sample. The result of the interaction is a bending of the cantilever beam. This bending is observed by changing in direction of laser light reflected off the top of the cantilever beam. At the end of this process the height at the point of contact is known and a single point of data has been collected. By retracting the cantilever upwards and repeating this process at other points on the sample surface a complete picture of the surface topology is revealed. Another mode of operation – the mode with which we are concerned – involves vibrating the cantilever

at a high frequency. This resonance is of the form causing the largest motion of the tip to and from the sample surface with the beam end opposite the tip most stationary. Each cycle of the oscillation the tip contacts the sample surface, resulting in a measurement, after which the tip quickly retracts from the sample to complete the cycle. This "tapping" mode is advantageous because it results in fast measurements while preventing damage that occurs when simply dragging the tip along the sample. As this process generates many thousands of samples per second (relating to the resonant frequency of the cantilever beam) the piezoelectric actuators drive movement of the sample in the place perpendicular to the cantilever tip. Henceforth we refer to this perpendicular direction as the z-axis while the sample plane is the xy-plane.

The experiment is described mathematically as such. Let the scanning region be given by $\Omega \subset \mathbb{R}^2$ with a path taken by the scanner along a curve

$$\gamma(t) : \mathbb{R} \to \Omega \tag{2.1}$$

over a time $t \in [0, T]$ where T is the total time taken collecting data. The height information is provided through the reflected laser information and is described by a signal $h(t) : \mathbb{R} \to \mathbb{R}$. Although this signal is being captured discretely at times corresponding to taps of the cantilever, the frequency at which information is collected is very high relative to other time scales therefore the signal can be modelled mathematically as a continuous signal.

While the cantilever travels the sample surface along a path specified by 2.1, this does not correspond to the signal sent by the AFM to the actuators. Denote by

$$\gamma_s(t) : \mathbb{R} \to \Omega \tag{2.2}$$

the ideal path the AFM would like the cantilever to follow as sent to the AFM from the software. The difference between γ and γ_s results from non-linear behaviour of piezoelectric actuators and their physical limitations. Simply speaking, asking the xy-plane piezo actuators to instantly reverse the sample's velocity is simply not possible and the influence of these types of limitations result in $\gamma(t) \neq \gamma_s(t)$. Advancements in AFM have attempted to

mechanically improve the control of this cantilever positioning when scanning at high speeds where the most distortion arises.

High-speed scanning with AFM is desirable in many situations. In cases where a single high-resolution image is desired waiting on the order of hours to produce the final result can be very inconvenient and can prevent other users from utilizing the AFM. In other cases a user may be using the AFM as an exploratory tool which means waiting minutes to see the topography and have the necessary information for adjustments before starting the next scan. In general, the time required to collect an AFM image is a serious drawback compared to other imaging techniques [HSF06]. Specific application in the semiconductor industry seek faster scanning to detect defects [KLH11]. With sufficiently fast scanning speeds the AFM could prove a valuable exploratory tool in the study of nano-scale biological and chemical dynamics [KYI10]. As a result, significant effort has been placed on reducing scan times.

2.1.2 Sources of Error

The primary sources of error when scanning with the AFM are due to positioning, thermal drift, and parachuting [BQG86, CSD01]. The most significant contribution to errors when approaching fast-scanning AFM is the inability to control precisely the position of the cantilever. In particular, excitations of natural resonant frequencies within the AFM result in potentially extremely violent and chaotic behaviour of the cantilever position. A natural characterization of this limitation is with a limiting frequency ω_L above which positioning signals can produce such a resonance.

In addition to the positioning issue resulting from the xy-plane actuators, three additional notable sources of error arise. The measured height of the sample during the scan $h(t)$, as observed using the reflected laser beam, is actually a combination of three components

$$h(t) = s(t) + d(t) + z(t). \tag{2.3}$$

The most trivial contribution $s(t)$ is known as *sample tilt* or simply *tilt*. This occurs when

the sample is mounted at an angle resulting in an offset of the sample by a function

$$s(t) = c + \gamma(t) \cdot \mathbf{n} \tag{2.4}$$

where $\mathbf{n}$ is the direction of tilt and c is some offset. The sample tilt is easy to describe mathematically and can often be corrected for by an initial calibration step to learn the unknowns.

The second significant component in h is known as *thermal drift* or simply *drift* as captured by $d(t)$. While the scan is run thermal changes in the AFM cause materials to expand and contract. Though the changes in material volume are microscopic, on the scales the AFM is concerned with the changes can be significant. The changes are very gradual in time, however, so the assumption that d is smooth is useful for developing techniques to remove this corruption.

The final source of error is *parachuting*. This arises due to the z-position limitations of the cantilever. When the cantilever travels over a high feature, for example, it must be moved away from the sample to maintain approximately a constant distance to avoid damage. Likewise, when the cantilever travels off of a high feature it must be moved closer to the sample until it again contacts the surface. Since this adjustment requires finite time, the apparent topography of a large cliff when the cantilever travels off the edge will appear as a gradual descent (as the cantilever is "parachuting" downward), rather than the sharp edge observed when the cantilever travels in the opposite direction. While the error introduced by this is typically small and appears only at prominent edges, it can cause measurements of the same feature taken in opposing directions to produce conflicting information. Furthermore, as the speed of the cantilever increases the effect is more evident. To resolve features below a specific height there is thus limit to the cantilever speed based on the capabilities of the AFM. Let v_L denote this limiting speed above which features begin to appear heavily distorted. This, in addition to ω_L, are characteristic properties of the AFM that place a lower bound on scan time.

2.1.3 Raster Scanning

In traditional AFM an image of the sample surface topography is collected by following a zig-zag pattern. An example is shown in the first column of figure 2.1 along with example signals. This pattern originates in the history of AFM and the natural format images are stored. Though there are considerable limitations to this scan pattern its dominance has resulted in considerable work to correct the afore-discussed errors. To correct positioning discrepancies, for example, mechanical improvements and high-order piezo actuator models are employed to force the cantilever to visit the locations given by γ_s [BS10, KFC04, STH08, HMH05, PBU07, THR12, ZSC10]. This is a difficult mechanical task because the raster scan position signal has significant contributions at high frequencies, including ω_L, due to turn-around points that require theoretically infinite acceleration. The terrible properties of the position signal are the first major drawback of the raster scan and are the primary motivation for this work investigating non-raster methods.

To correct the other sources of error, thermal drift and sample tilt, a common technique involves subtracting from each grid line a least-squares linear fit. Since $s(t)$ is linear and $d(t)$ is smooth (and therefore approximately linear within each grid line) both are easily enough subtracted by this process. In some cases this technique results in heavily distorted results, however, thus some supervision is often required. Indeed, in a variety of situations $s(t) = d(t) = 0$ will still result in the signal $z(t) = h(t)$ being "corrected" by this approach, resulting in errors.

The second serious limitation of the raster scan is due to parachuting and positioning errors. Information can, in theory, be collected and used from when the cantilever travels in both directions of the zig-zag pattern. While this seems very natural, the positioning issues, and parachuting effects, mean that frequently the lines are not registered correctly and features disagree depending on the direction the cantilever travelled. Some positioning discrepancies may be corrected by registering odd lines with even lines to offset lag in the system, however even slight disagreement produces seriously distorted final topographs due

to parachuting. The result, unfortunately, is often that every other line is used to form the final image, the other half of the data being discarded, thereby reducing the collection rate by a factor of two.

2.1.4 Non-Raster Scanning

There is no inherent hardware restriction within many AFMs requiring the cantilever to travel in the raster pattern. In cases where the cantilever deviates such that $\gamma \neq \gamma_s$ the actual path γ, sensors can be used to observe γ directly during acquisition. The paradigm shift herein proposed is a movement away from the strict positioning demands of γ_s toward non-raster patterns using the information γ with *inpainting algorithms* applied to the observed path γ. This new methodology, coined "sensor inpainting" [ZMF13], opens up a variety of possibilities for improving the field of high-speed and real-time AFM. This seemingly trivial difference avoids the need for challenging xy-position control to combat the distortions, resonances, and discrepancies arising from the raster pattern. Additionally the complex modelling of non-linear piezoelectric actuator behaviour that becomes more difficult as the scan speed is increased is unnecessary; γ is collected as part of the collection process and sensor inpainting is used to generate a final topograph. Provided the sample surface is sufficiently sampled an accurate topograph can be generated.

figure 2.1 shows three possible scan patterns as well as plausible signals captured during a scan. Note that shown are all components of 2.3 even though the AFM only observes $z(t)$ from which all components must be isolated. The first pattern is the traditional raster scan that covers the sample surface uniformly with the zig-zag motion, but the cantilever must make sharp turns at the sides of the scan area. The Archimedean spiral is another pattern that covers the sample rather uniformly without having any sudden cantilever acceleration, however near the center high frequencies may be excited by the tight near-circular motions. The third example path is the spirograph, generated by making two simultaneous circular motions. The spirograph has a very specific frequency profile that is useful for preventing the excitations of resonant frequencies in the scanner, however this comes at the cost of poor

sampling distribution over the sample surface. Other scan patterns such as the "cycloid" and Lissajous curves are possible alternatives to the patterns in figure 2.1 herein studied, however the consideration of these and other scan patterns is beyond the scope of this work.

The Archimedean spiral is of particular interest and is the focus of this work due to uniformity of sampling and the smoothness of the path. It is described in polar coordinates for some parametrization function $f(t) : \mathbb{R} \to [0, 1]$, by

$$\begin{aligned} \theta(t) &= 2\pi N f(t) \\ r(t) &= R f(t). \end{aligned} \tag{2.5}$$

Natural properties of an ideal scan pattern avoid the pitfalls mentioned previously, in particular those arising from the raster pattern. A good scan path, or specifically for our consideration parameterization function f, should sample the surface as uniformly as possible, respect the speed limitation v_L of the cantilever, avoid exciting high-frequency resonances in the AFM beyond a characteristic resonance limit ω_L, and complete the scan in minimal time. Finally, it is desirable for the cantilever to travel in the same direction at each point, if possible, to avoid data disagreement when parachuting effects may exist.

2.1.5 Sensor Inpainting

The raster paradigm is advantageous because it makes completing a grid-based topography image trivial – simply collect a sample at each point of the grid, row-by-row, until each pixel has been visited. Unfortunately the raster scan pattern is not well-suited to the limitations of the scanning process. In order to scan with general curves γ to scan faster there must foremost be a way to view the output from the AFM in the traditional manner on a grid as a scientist would typically view it. The field of image inpainting has precisely the tools necessary to solve such a problem. Along γ the height values are known while off the curve the values must be filled-in, or inpainted, in order to form a complete picture.

While the inpainting literature is vast [BVS03, AK06, BSC00, CMS98, GO09, CWT11], one particularly effective, simple, and fast technique herein used is H^1-regularized inpainting.

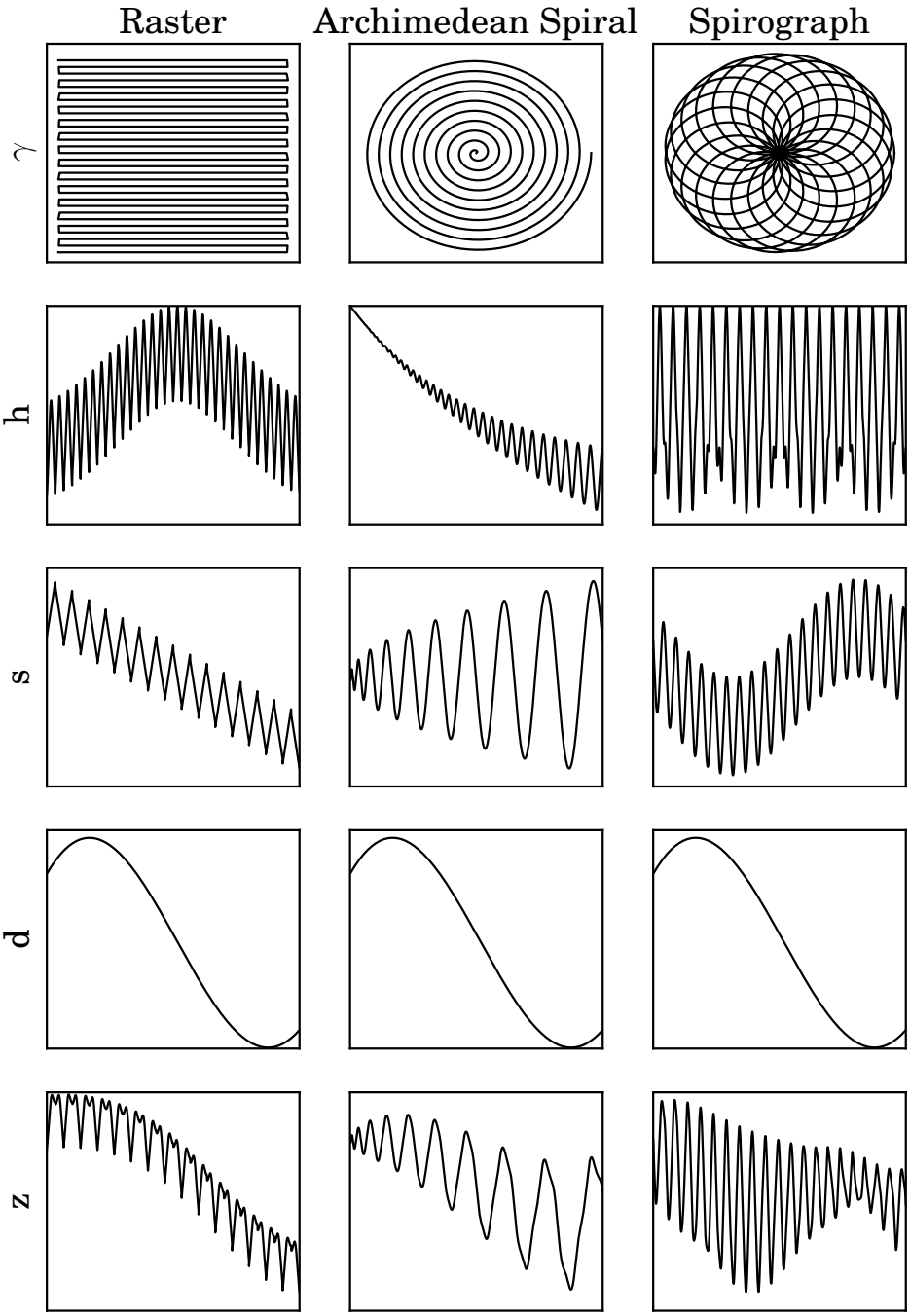

Figure 2.1: Example scan patterns. The scan path over the sample surface (first row) is travelled by the AFM cantilever. The collected signal over the scan time x (last row) is a sum of the sample surface h (second row), sample tilt s (third row), and thermal drift d (fourth row).

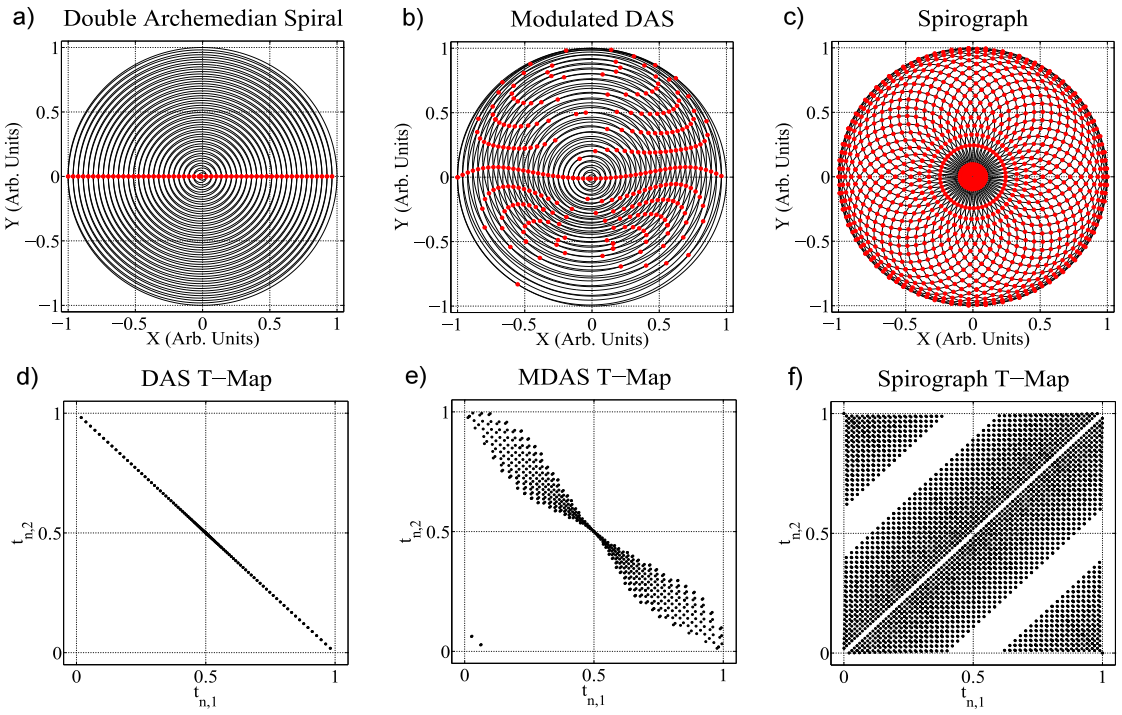

Figure 2.2: Scan patterns with self-intersections. These are three examples of non-raster self-intersecting scan patterns that can be used to discover and remove thermal drift errors. Shown are the scan patterns with red dots denoting points of self-intersection (top row) and T-maps for each scan showing times of self-intersection (bottom row). Reproduced with permission [MZB14] .

Let $D = \{\gamma(t), t \in [0,T]\}$ and $\lambda(x) : \Omega \to \mathbb{R}$. The solution to the variational problem

$$\min_{u \in H^1(\Omega)} E(u) = \int_\Omega |\nabla u|^2 + \int_0^T \lambda(x) \left[u(\gamma(t)) - z(t)\right]^2 dt \tag{2.6}$$

produces a topography function $u : \Omega \to \mathbb{R}$ that agrees with the observations along the curve γ while filling in missing information using a smooth H^1-minimal completion. Similar mathematics to this originate in thermodynamics when considering the steady-state of temperatures in a system with set boundary values. Extrema of $E(u)$ with $u \in H^1(\Omega)$ satisfy

$$\begin{aligned} 0 &= \Delta u \quad \text{on } \Omega \backslash D \\ \lambda(u - z) &= \Delta u \quad \text{on } D. \end{aligned} \tag{2.7}$$

This inpainting algorithm is attractive because it can be solved efficiently. Furthermore, the simplicity of the model means that missing information will not be completed with features that are not observed resulting in misleading information. More recent and advanced inpainting algorithms, in contrast, may potentially fill in a missing region by extrapolating patterns, not simply value. This is unreasonable for scientific observations that require accuracy above visually satisfying results. Any inpainting model used for our purpose should not add unobserved features for the sake of visual satisfaction.

To solve the problem discretely on a grid the observations in z are distributed into grid cells using bilinear weighting with inter-cell averaging. The total weight of samples for each cell are accumulated and used to produce the fidelity function in space λ. The resulting linear equation (2.7) is easily solved using stencil discretization of the Laplacian and matrix inversion, or using fast multi-scale techniques. Inpainting techniques such as this free the AFM from the strict control requirements associated with positioning the cantilever at the correct grid location at the correct times.

2.1.6 Chapter Overview

This chapter is a summary of contributions to the field of AFM pertaining to techniques for non-raster pattern AFM made available using sensor inpainting. The following section

2.2 develops a solution for correcting thermal drift in the non-raster domain using self-intersecting scan patterns. After that, section 2.3 takes the Archimedean spiral under special consideration and thoroughly investigates parametrization of the curve to optimally obey physical limitations of the AFM.

2.2 Drift Correction

2.2.1 Description

Using an AFM with non-raster scan patterns is desirable to avoid the frequency characteristics of the raster scan pattern that severely limit scan speed. The previous section demonstrated that modern inpainting techniques can take the observed cantilever path γ of arbitrary geometry over the sample surface and use this information in combination with the height signal $z(t)$ to complete a topograph of the sample surface. Recall, however, that the AFM captures a signal $h(t) = s(t)+d(t)+z(t)$ that is composed of corrupting signals s and d, respectively the result of sample tilt and thermal drift. While sample tilt can be compensated often using an initial calibration step, the drift component is more difficult to handle. The drift d is assumed to be smooth and only gradually changing throughout the scan making its removal not entirely impossible. Recall that correcting this in the raster paradigm involves subtracting a least-squares fit from each grid line with optional user supervision/intervention to prevent topograph corruption.

Extending this subtraction approach to general non-raster scan patterns is not trivial and naïve adaptations, such as subtracting least-squared fits of segments, are plagued by the same issues of data corruption and user supervision. We have found that non-raster AFM presents an interesting opportunity, described here, for using redundant observations to discover and remove drift component. Suppose that the scan path γ self-intersects at M points with intersection times $t_{n,1}, t_{n,2}$ given by $\gamma(t_{n,1}) = \gamma(t_{n,2})$. While the raster scan pattern does not self-intersect, more general patterns may and in some cases self-intersections are very common. The self-intersections represent on one hand wasted time since the same

location of the sample is being measured twice. A small number of self-intersections can be very useful for the discovery and removal of thermal drift errors, however, and when facing the challenge of removing drift from a scan taking a tiny fraction of the overall scan time to collect self-intersections is a small price to pay.

At the locations of self-intersection 2.3 implies that

$$h(t_{n,2}) - h(t_{n,1}) = z(t_{n,2}) - z(t_{n,1}) + d(t_{n,2}) - d(t_{n,1}) \tag{2.8}$$

since sample tilt is a function only of location. The component z, because it contains the effects of parachuting, may not vanish from this equation at places and times when parachuting is taking place. For appropriate scan speeds below v_L, the z contribution will be assumed vanishing in 2.8 because parachuting results from excessive in-plane tip motion. When parachuting does occur it coincides with high features and edges that often are sparse in the scanning area, so the effect is unlikely to coincide with a large percent of self-intersection points. If scanning quickly enough for parachuting to take place over a large percent of the scanning area the proposed model is unlikely to succeed due to overall poor measurement fidelity. Let $\delta_n = d(t_{n,2}) - d(t_{n,1})$ be then the observed differences in the thermal drift component.

Assuming the drift component is smooth, the following non-dimensionalized in time variational problem is proposed [MZB14] to discover the full drift contribution:

$$E(d) = \sum_{n=1}^{M} \left(d(t_{n,2}) - d(t_{n,1}) - \delta_n\right)^2 + \lambda \int_0^1 |d''(t)|^2 dt. \tag{2.9}$$

This energy is composed of contributions from two terms. The first enforces agreement with the observed differences in value between the two points in time. The second term smooths the solution and enforces the natural physical assumption that the thermal drift is gradual over time. To simplify solving 2.9 d can be restricted to linear combinations of basis functions, a natural choice for which is a basis of splines that have desirable smoothness properties. The next section outlines the least-squares solution to the energy minimization problem restricted to representation on such a basis set.

2.2.2 Solution

The goal is to solve (2.9) with representation restricted to basis of functions $\{\Phi_i(t)\}$ for $i = 1, 2, ..., N$, where $\Phi_i(t)$ is the i^{th} basis function evaluated at time t. Expanding the drift solution d using this basis with coefficients c_i, notice that the observed differences are given by a combination of differences on the basis functions

$$d(t_{n,2}) - d(t_{n,1}) = \sum_{i=1}^{N} c_i \left[\phi_i(t_{n,2}) - \phi_i(t_{n,1})\right]$$

and proceed much the same way as the classical least-squares approximation. Recall that δ_j is the error in height observed at the j^{th} intersection.

Let $\vec{d}$ and $\vec{\delta}$ denote length-M column vectors with, respectively, components $d(t_{j,2}) - d(t_{j,1})$ and δ_j for $j = 1, 2, ..., M$. Let $\vec{c}$ be the length-N column vector formed by the coefficients c_i where $i = 1, 2, ..., N$. Denote by $\mathbf{A}$ the M-by-N matrix containing the basis differences at the crossing points such that $\mathbf{A}_{ij} = \phi_j(t_{i,2}) - \phi_j(t_{i,1})$. The error on the differences is $||\vec{\delta} - \vec{d}||^2 = ||\vec{\delta} - \mathbf{A}\vec{c}||^2$. Define the following N-by-N matrix $\mathbf{M}$:

$$\mathbf{M}_{ij} = \int_0^1 \phi_i''(t)\phi_j''(t)\, dt$$

By algebraic manipulation it may be shown that:

$$\int_0^1 |d''(t)|^2 dt = \vec{c}^{\dagger}\mathbf{M}\vec{c}$$

Using these results, the functional in (2.9) may be now restated in terms of a minimization over $\vec{c}$:

$$\min_{\vec{c}} \|\vec{\delta} - \mathbf{A}\vec{c}\|^2 + \lambda \vec{c}^{\dagger}\mathbf{M}\vec{c}$$

Differentiation with respect to $\vec{c}$ leads to the optimality condition

$$L_\lambda \vec{c} = \left(\mathbf{A}^{\dagger}\mathbf{A} + \lambda\mathbf{M}\right)\vec{c} = \mathbf{A}^{\dagger}\vec{d} = \vec{h}$$

The matrix L_λ is invertible and positive definite if $\lambda > 0$, in which case the solution is additionally guaranteed to be unique. Therefore, to remove the drift component of general non-raster scan patters that have a small number of self-intersections, a least-squares fit to the difference observations produces a model for the drift component.

2.2.3 Path Fitness

The scan path chosen plays a significant role in the capacity for 2.9 to discover the drift function accurately. Indeed, the extreme case where no self-intersections exist results in the problem that all drift functions will be undiscovered. A simple argument leads to the consideration of the following quantity that enables analysis of scan path fitness for discovering $d(t)$.

Let $\vec{v}$ be the eigenvector of $\mathbf{A}^\dagger\mathbf{A} + \lambda\mathbf{M}$ corresponding to the smallest eigenvalue, the value denoted by ζ_λ. v represents the direction in which the energy, starting from the optimal solution, increases the least as the solution is manipulated. When ζ_λ is small, $\vec{v}$ is smooth and is minimally dependent on the differences δ_i indicating a drift profile unlikely to be discovered. Likewise, when ζ_λ is large deviation from any given solution to the fitting problem will significantly decrease smoothness and change the difference values. Therefore ζ_λ provides a quantitative technique for determining if a scan path's self-intersections are able to discover a smooth drift function. In general there is always a theoretical drift function under which δ_i, and our model, will be invariant. This fitness indicates approximately how variable such a function must be given the self-intersection times and therefore, for very large values, the function $d(t)$ will be extracted accurately. This fitness depends only on the times of self-intersection only, and therefore is a fitness applied to the scan path itself.

Different self-intersecting scan paths therefore will have a quantifiable difference in fitness. In order to correct for drift a scan pattern must have self-intersections. For this purpose, three scan patterns are proposed for non-raster scanning: the double Archimedean spiral (DAS), modulated DAS (MDAS), and the spirograph. Each of these patterns are shown in figure 2.2 with self-intersections and T-map, a scatter plot of the points $(t_{i,1}, t_{i,2})$.

The DAS is created simply by following an Archimedean spiral inwards and back outwards again. Unfortunately, this scan pattern performs poorly when correcting drift for reasons that are not apparent initially, through are quickly discovered in practice after brief experimentation. Indeed, the difficulties encountered using the simple DAS resulted in the

creation of the ζ_λ quantity. Taking $\lambda = 10^{-3}$ for the remainder of this work, the three scan patterns specifically shown in figure 2.2 result in the values of $\zeta_\lambda = 0.02$ for the DAS, $\zeta_\lambda = 1.2$ for the MDAS, and $\zeta_\lambda = 35$ for the spirograph.

The reason for the very small fitness for the DAS is due to symmetry of the self-intersection times. Simple inspection of the T-map results in the conclusion that any thermal drift profile that is symmetric in time about the center of the scan results in $\delta_i \approx 0$, hence such a drift profile is an invariant of the problem and paths of higher fitness are desired. Taking one coordinate of the DAS and adding a slight perturbation produces the MDAS and, as a result, more informative self-intersections and a higher fitness value. The final scan pattern - the spirograph - generates a very large number of self-intersection points and therefore a considerably higher fitness than the other two scan patterns.

2.2.4 Experimental Validation

Figure 2.3 presents three scans performed using three different scan patterns, all corrected using the proposed drift correction technique. All three scans were performed on an MFP-3D AFM by Asylum Research that was modified to permit arbitrary cantilever paths. The first scan, shown in the first column, is of an annealed gold sample using a $500nm$ diameter DAS consisting of 1700 loops and generating a fitness ζ of 0.8. The second column is a 1.4μm MDAS scan also over an annealed gold sample with 471 loops producing a fitness $\zeta = 126$. Finally, in the third column is a spirograph with a 30μm diameter, 414 loops, and an extremely high fitness of $\zeta = 447$ taken over a calibration sample with 8nm deep hexagons arranged in a grid. In all cases thermal drift component removal is performed prior to tilt correction through the subtraction of a least-squared error planar data fit.

Using H^1 sensor inpainting the highly restrictive raster pattern is not required to produce a quality image of the sample surface with these various scan patterns. In all three cases the result is a quality scan requiring no supervision to be corrected. Only in the case of the DAS, due to low self-intersection fitness, does a small amount of the thermal drift component

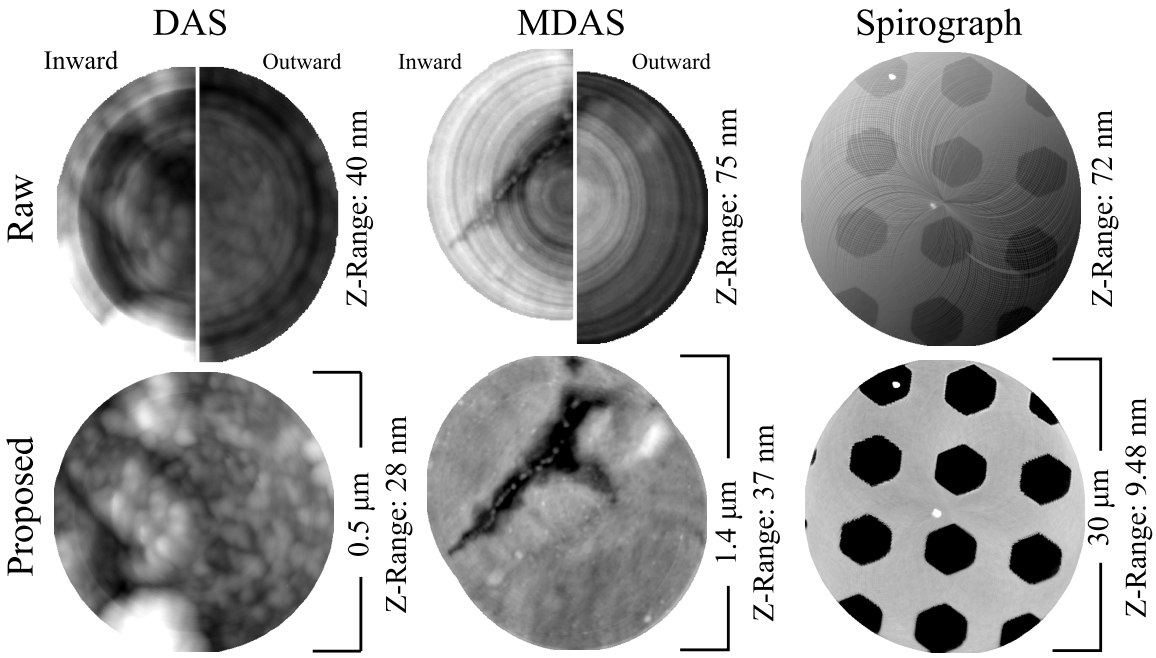

Figure 2.3: Drift-corrected and sensor inpainted AFM scans [MZB14]. The left column is the result of a DAS scan on annealed gold with significant drift. The center column is a MDAS scan also on an annealed gold sample, and the last column corresponds to a spirograph scan taken over a calibration sample. The significant thermal drift present in the raw data (top row) is removed using the proposed method in all cases (bottom row). Reproduced with permission [MZB14]

remain in the form of faint visible rings. Both the MDAS and the spirograph results, due to high ζ fitness, are easily corrected. Unlike variations on the fit subtraction technique traditionally used for thermal drift correction with raster patterns, the tilt present in the spirograph scan has no impact on the efficacy of the drift correction because tilt has no influence on the differences δ_j.

2.3 Archimedean Spiral Parametrization

The Archimedean spiral is an attractive base scan pattern due to the path γ that covers the sample area uniformly. Additionally, unlike the spirograph the cantilever travels in the

same direction at each location on the sample surface thereby minimizing the impact of parachuting errors. Unfortunately it is not clear how to parametrize this scan pattern given by 2.5 through the choice of $f(t)$. In this section the choice of $f(t)$ is studied with respect to the physical challenges posed by the AFM. The conclusion is that by considering carefully the relevant limitations a parametrization exists that is optimal in the sense of completing the scan in the shortest possible time.

In addition to obeying physical limitations of the AFM, another property of scan patterns is the distribution of samples in the area of interest. While the Archimedean spiral path covers the sample area well, different choices of $f(t)$ change the distribution of samples along the scan path. This is because the sampling rate is constant in time and therefore depends on how quickly the cantilever is moving on the scan path. To characterize this mathematically two distances are of interest: the distance between each loop of the Archimedean spiral and the distance between samples along the scan path. The first of these is the radial distance (RD)

$$\mathrm{RD}(r,\theta) = \frac{2\pi\dot{r}}{\dot{\theta}} \tag{2.10}$$

while the second is the tangential distance (TD)

$$\mathrm{TD}(r,\theta) = \frac{r\dot{\theta}}{f_s} \tag{2.11}$$

with f_s the frequency at which the AFM measures the sample. The quantities (2.10) and (2.11) can then be combined to find the density of samples at each point in the sampling area, a function of the radius

$$\delta(r) = \frac{1}{\mathrm{TD}\cdot\mathrm{RD}} = \frac{f_s}{2\pi r\dot{r}}. \tag{2.12}$$

In addition to the amount of information at each radius r it is important to know how the samples are distributed. For example, having extremely high density along the scan path isn't helpful if the scan has only five loops and therefore large gaps of no information. The second quantity proposed in addition to $\delta(r)$ is the homogeneity of samples

$$\eta(r,\theta) = \frac{\mathrm{RD}}{\mathrm{TD}} = \frac{2\pi f_s\dot{r}}{r\dot{\theta}^2}. \tag{2.13}$$

η has an ideal value of one if the radial and tangential spacing of samples is equal. Therefore while θ_L and v_L describe physical limitations on the choice of $f(t)$, the values of δ and η provide a quantitative approach to studying how uniformly a parametrization $f(t)$ samples.

2.3.1 The CAV and CLV

2.3.1.1 Derivation

The primary restrictions of the AFM arise from resonant frequencies and speed limitations. The frequency limitation is characterized by a frequency ω_L above which excitation from position actuators can produce serious signal distortions and excitation of resonance. The second limitation is described by v_L, a limiting cantilever speed above which the signal is distorted by position errors and parachuting errors resulting in poorly-resolved surface topography. To state these limitations mathematically, differentiate (2.5) to get the angular and radial velocity components

$$\begin{aligned} \dot{\theta}(t) &= 2\pi N \dot{f}(t) \\ \dot{r}(t) &= R\dot{f}(t). \end{aligned} \tag{2.14}$$

These velocities can then be used to describe the restrictions on f

$$R\dot{f}(t)\sqrt{(2\pi N f(t))^2 + 1} \leq v_L \tag{2.15}$$

$$\dot{\theta}(t) = 2\pi N \dot{f}(t) \leq \theta_L \tag{2.16}$$

with (2.15) the cantilever speed restriction and (2.16) the frequency restriction. Taking first into consideration (2.15), the 1 in the square root plays an insignificant role anywhere $Nr \gg R$ which is most of the scan. Ignoring the 1 for now and letting $f_{\mathrm{CLV}}(t)$ represent a parametrization that completes the scan as quickly as possible subject to this constraint. The fastest scan parametrization will naturally be the scan for which the speed is equal to

this limit, therefore it will obey

$$R\dot{f}_{\mathrm{CLV}}(t) = \frac{v_L}{2\pi N R f_{\mathrm{CLV}}(t)} \tag{2.17}$$

$$\Rightarrow f_{\mathrm{CLV}}(t) = \sqrt{\frac{v_L t}{\pi N R}} \tag{2.18}$$

known as the *constant linear velocity* (CLV) [MM10] spiral since it attempts to maintain a constant cantilever speed. The second constraint (2.16) also implies a parametrization that completes the scan as fast as possible subject to the angular frequency constraint

$$2\pi N\dot{f}_{\mathrm{CAV}}(t) = \theta_L \tag{2.19}$$

$$\Rightarrow f_{\mathrm{CAV}}(t) = \frac{\theta_L t}{2\pi N}. \tag{2.20}$$

This parametrization is the *constant angular velocity* (CAV) spiral [Wie01, MM09, Hun10, RPP14] and it is designed to avoid the excitation of resonant frequencies. Using the fact that $f(t) = 1$ is when the scan is completed the total scan times T_{CLV} and T_{CAV} can also be derived

$$T_{\mathrm{CLV}} = \frac{\pi N R}{v_L} \tag{2.21}$$

$$T_{\mathrm{CAV}} = \frac{2\pi N}{\theta_L}. \tag{2.22}$$

2.3.1.2 Characteristics

The CAV and CLV spirals are designed to push the limits of the AFM. The CAV finishes the scan as quickly as possible subject to frequency limitations while the CLV does the same subject to speed limits. Here the two parametrization options are studied with respect to both ω_L and v_L conditions as well as δ and η sample distribution properties.

The use of f_{CLV} results in a cantilever speed that is approximately equal to v_L. This is only approximate, in particular near $t = 0$, due to the simplification made from (2.15) but in practice discretization of the scan path means that the constraint is essentially obeyed.

The parametrization (2.18) implies an angular velocity for the CLV of

$$\dot{\theta}(t) = \sqrt{\frac{2\pi N v_L}{Rt}}$$

that as $t \to 0$ tends to infinity. Thus the CLV is not theoretically capable of satisfying the frequency requirement since in the center of the spiral loops are completed at arbitrarily high frequencies due to the constant top speed and shrinking path radius. Specifically, when t satisfies

$$t < \frac{2\pi N v_L}{R\theta_L^2}$$

the frequency constraint is violated by the CLV. The other criteria for evaluating the CLV is using η and δ. These quantities are independent of time for the CLV resulting in

$$\delta_{\mathrm{CLV}}(r) = \frac{N f_s}{R v_L} \tag{2.23}$$

$$\eta_{\mathrm{CLV}}(r) = \frac{n}{\pi N^2}. \tag{2.24}$$

Because these quantities do no change in time, the number of loops N and the radius R can be adjusted until samples are distributed ideally within the sample area. In the left half of figure 2.4 the properties of the CLV solution are visualized including the cantilever speed, angular frequency, and distribution of samples. While the theoretical sample distribution is perfect, the high angular frequency near the center results in poor positioning of the cantilever and chaotic behaviour.

The alternate parametrization for the Archimedean spiral proposed is f_{CAV}. By design this obeys the frequency limitations of the AFM, however the cantilever speed constraint must be considered as well. Using this parametrization in (2.15) produces

$$t \leq \frac{T_{\mathrm{CAV}}}{\theta_L} \sqrt{\left(\frac{v_L}{R}\right)^2 - \frac{1}{T_{\mathrm{CAV}}^2}} \tag{2.25}$$

hence there is a violation of the speed constraint near the end of the scan if

$$v_L < R\theta_L. \tag{2.26}$$

Using (2.26) with (2.21) leads to the conclusion that the CAV parametrization violates the speed constraint if the CAV spiral takes less than twice the time as the CLV spiral. That is, unless the CAV is significantly slower it will have too high of a cantilever speed on the periphery. Things look worse for the CAV spiral when considering the sampling densities

$$\delta_{\mathrm{CAV}}(r) = \frac{f_s N}{\theta_L R r} \tag{2.27}$$

$$\eta_{\mathrm{CAV}}(r) = \frac{f_s R}{N \theta_L r} \tag{2.28}$$

Unlike the CLV spiral that had uniformly distributed samples the CAV sampling density depends on the radius, thus no choice of scan parameters R and N will be able to achieve a satisfactory sampling distribution. In the next section the advantages from both the CLV and CAV spirals are combined into an *optimal scan parametrization*, where optimality is in the sense of completing an Archimedean spiral scan as quickly as possible without violating the constraints. In the right half of figure 2.4 the properties of the CAV solution are visualized where unlike the CLV solution the path is followed correctly. Unfortunately samples are not acquired uniformly and the cantilever speed exceeds the limiting value.

2.3.1.3 Results and Discussion

The results of using both the CLV and the CAV parametrizations are displayed in figure 2.4 with experimental acquisition on . On the left the CLV solution with a limiting cantilever speed (a) generates an ideal sampling distribution (c). Unfortunately, the high frequencies near the center of the scan (b) generates chaotic positioning due to resonance excitation and, therefore, disagreement between γ and γ_s (d). The final sensor inpainting result (e) therefore produces quality results near the periphery (C) and mid-scan (B), but poor results in the center (A). Alternatively, the CAV solution obeys the frequency limitation strictly (g) but violates the cantilever speed constraint by a factor of two (f). The distribution of samples is heavily center-biased (h) and the cantilever speed is extremely high on the periphery (i). As a result while the center (A) and mid-scan (B) are clear in the inpainted result (j), the periphery (C) is overly smooth due to z-piezo limitations.

Evidently the two parametrizations perform well in different regions of the scan where they obey the AFM limitations. In 2.3.2 the best of each scan is used to develop an optimal solution that completes the scan in the fastest time subject to the constraints of the AFM. This optimal parametrization is then demonstrated to consist exactly of a CAV solution near the center that transitions into a CLV solution near the perimeter.

2.3.2 Optimal Scan Parametrization (OPT)

2.3.2.1 Derivation

Both the CLV and CAV spirals risk violating the limitations of the AFM unless limited to slow scan speeds. The CLV necessarily violates the frequency limitation near the center of the scan while the CAV potentially violates the speed condition near the periphery. Furthermore, while the sampling distribution of the CLV does not depend on radius and can be therefore adjusted to the ideal values through manipulation of R and N, the CAV has a sampling density dependent on the radius with very high density near the center and low density at the periphery. An optimal scan parametrization [ZMA16] would combine the benefits of these two options and, by construction, not violate AFM limitations.

To formulate such a parametrization, first let T denote the total scan time and $t_* = t/T$ denote dimensionless scan time. The optimal parametrization of the spiral sought is some function $f(t_*)$ that completes the scan as quickly as possible while respecting both AFM constraints on speed and frequency.

Theorem 1. *The optimal parametrization of the Archimedean spiral, $f_{OPT}(t)$, is the CAV solution near $t = 0$. If $v_L < R\theta_L$ there is a single transition at some time to a CLV solution for the remainder of the spiral.*

Proof. The scanning path for the Archimedean spiral is determined in polar coordinates by the angle $\theta(t) = 2\pi N g(t)$ and the radius $r(t) = Rg(t)$ for some parametrization g. The optimal parametrization is a function g that completes the scan in the least time subject to

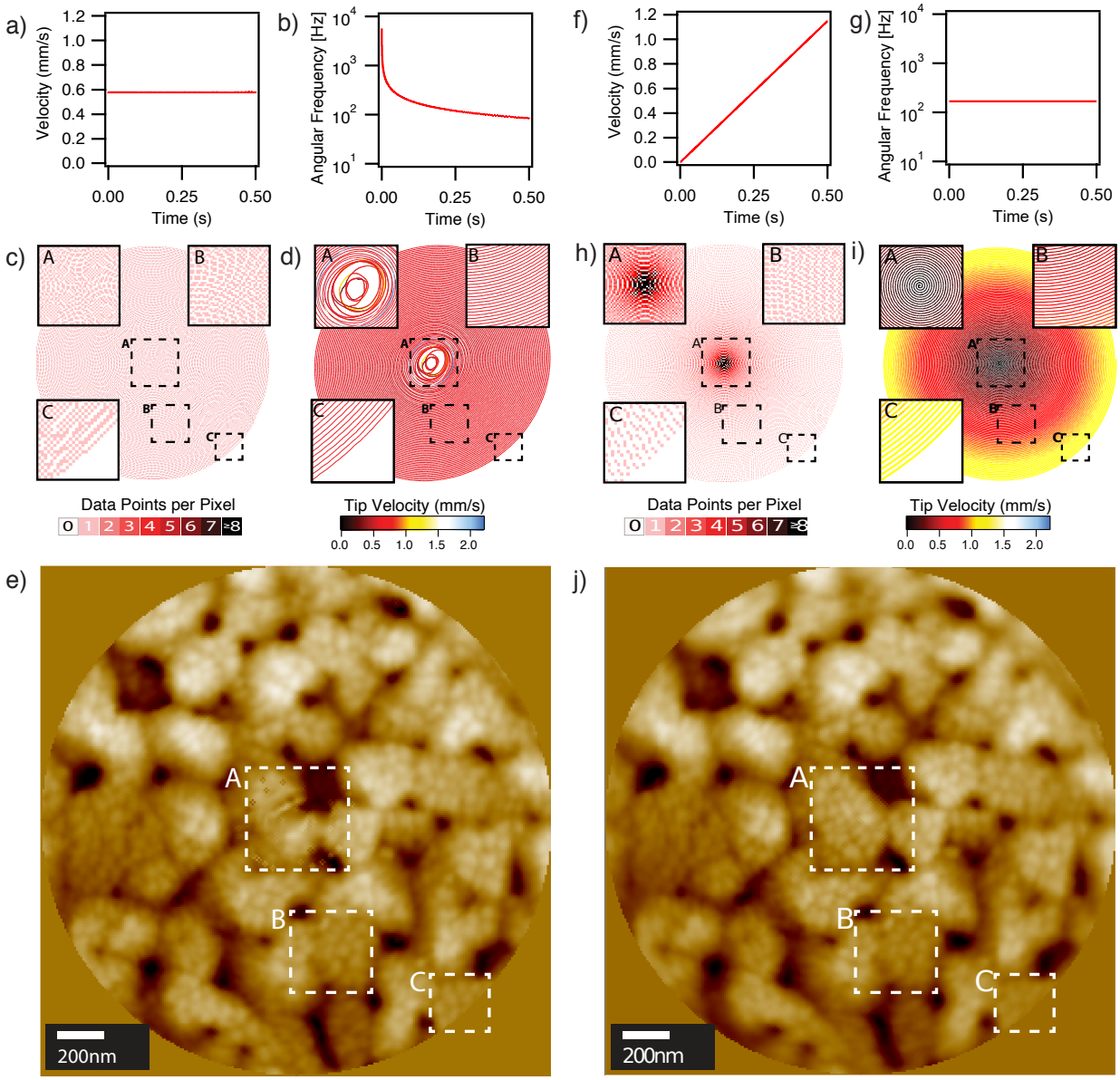

Figure 2.4: AFM scans and properties using CLV (left side) and CAV (right side) parametrizations [ZMA16]. Shown in the top row is the cantilever speed (a/f) and angular frequency (b/g) as functions of time. In the middle row is the sampling density expected using γ_s (c/h) and path observed γ travelled with color representing instantaneous cantilever speed (d/i). Finally, the sensor inpainted AFM scan of an annealed gold sample is in the bottom row (e/j). The AFM used was a Cypher ES by Oxford Instruments. Figure is ©2016 IEEE.

the physical constraints of the device. The constraints are of the form $|g'| < l(g)$

$$|g'| \leq \min\left(\frac{v_L}{R\sqrt{1+(2\pi N g)^2}}, \frac{\omega_L}{2\pi N}\right) \equiv l(g). \tag{2.29}$$

Then the optimal g minimize the scan time. The scan is finished when $g = 1$ when scanning counter clockwise or $g = -1$ when scanning clockwise. Taking the counter clockwise scenario, define the scan completion time by

$$T[g] = \min_{t\geq 0, g(t)=1} t.$$

The problem is to find a function g which, subject to the constraints, minimizes this quantity

$$g = \arg\min_{\hat{g}\in F} T[\hat{g}]$$

where F is the set of all continuously differentiable functions satisfying the constraint l,

$$F = \left\{h \in C^1([0,\infty]) : h(0) = 0, h' \leq l(h)\right\}.$$

Define f_{OPT} to be the solution to the differential equation $f'_{\mathrm{OPT}} = l(f_{\mathrm{OPT}})$ with initial condition $f_{\mathrm{OPT}}(0) = 0$. Because l is autonomous, uniformly Lipschitz, and bounded, the solution exists, is unique, and resides in F.

The parametrization given by f_{OPT} is fastest in the sense of $T[\cdot]$. To see this, suppose $h \in F$ is a another parametrization. Let $I = (a, b]$ be an interval such that $h(a) = g(a)$ and $h > g$ on I. If such an a and b do not exist it must be that $h \leq f_{\mathrm{OPT}}$ for all time, so $T[h] \geq T[f_{\mathrm{OPT}}]$ and h is not faster. Assume therefore a and b can be chosen. Within I there must be a point s at which $h'(s) > f'_{\mathrm{OPT}}(s) \Rightarrow l(f_{\mathrm{OPT}}(s)) < l(h(s))$, but this is impossible since $h(s) > f_{\mathrm{OPT}}(s)$ and l decreases monotonically. No such interval I can exist, and therefore $T[h] \geq T[f_{\mathrm{OPT}}]$. Because h was arbitrary there exists no strictly faster parametrization than f_{OPT}. □

The analytic form for f_{OPT} is given by simple linear growth until the frequency constraint becomes less restrictive than the speed constraint when

$$\frac{v_L}{R\sqrt{1+(2\pi N f_{\mathrm{OPT}})^2}} = \frac{\omega_L}{2\pi N}.$$

Because of monotonic growth there is a single time t_L at which this occurs, from which point onward the solution belongs to a family of solutions to the CLV problem. These solutions take the form of a class of functions implicitly solving

$$\nu + \frac{v_L t}{R} = \frac{g(t)}{2}\sqrt{1 + (2\pi N g(t))^2} + \frac{\sinh^{-1}(2\pi N g(t))}{4\pi N}$$

for ν a constant determined by boundary conditions at $f_{\text{OPT}}(t_L)$. Provided that the approximations

$$N \gg 1$$

and

$$N f_{\text{OPT}}(t_L) \gg 1$$

hold, the 1 in the square root and the hyperbolic sine terms can be ignored thereby producing a solution of the form

$$f_{\text{OPT}}(t) \approx \frac{1}{\pi N}\sqrt{\nu + \frac{v_L t}{R}}.$$

Therefore the optimal scan parametrization, f_{OPT}, is a CAV solution until a transition time when a CLV solution is used.

Although this verifies the existence and form of the optimal parametrization f_{OPT} it does not clearly define a functional form for explicit use. To construct this, let f again be some parametrization of the Archimedean spiral and T the total scan time. The CLV is produced approximately when $f(t_*) = \sqrt{t_*}$ and the CAV is produced when $f(t_*) = t_*$. Define $a \equiv \frac{2\pi N}{T\omega_L}$. To push the angular frequency limit initially the composite spiral's f must be of the form $f(t_*) = \frac{t_*}{a}$ as this results in $\frac{d\theta}{dt} = \omega_L$. Using the CAV up to some time t_{*L} then transitioning to a CLV spiral with parameters C_1 and C_2 means the OPT solution is of the form

$$f_{\text{OPT}}(t_*) \approx \begin{cases} \frac{t_*}{a} & \text{if} \quad t_* < t_{*L} \\ \sqrt{C_1 t_* + C_2} & \text{if} \quad t_* \geq t_{*L} \end{cases} \tag{2.30}$$

To find the parameters, t_{*L}, C_1, and C_2, three natural properties should be enforced. The scan should be finished at time $t_* = 1$ hence $f(1) = 1$ and f_{OPT} and f'_{OPT} should be continuous at t_{*L}.

The three conditions imply, in order, the equations

$$1 = \sqrt{C_1 + C_2} \tag{2.31}$$

$$\frac{t_{*L}}{a} = \sqrt{C_1 t_{*L} + C_2} \tag{2.32}$$

$$\frac{1}{a} = \frac{C_1}{2}(C_1 t_{*L} + C_2)^{-\frac{1}{2}}. \tag{2.33}$$

The first equation implies $C_2 = 1 - C_1$, which after substitution into the remaining equations produces

$$\frac{t_{*L}}{a} = \sqrt{C_1(t_{*L} - 1) + 1} \tag{2.34}$$

$$\frac{1}{a} = \frac{C_1}{2}(C_1(t_{*L} - 1) + 1)^{-\frac{1}{2}}. \tag{2.35}$$

Squaring both sides of (2.34) and solving for C_1 implies that

$$C_1 = \frac{\frac{t_{*L}^2}{a^2} - 1}{t_{*L} - 1} \tag{2.36}$$

which after substituting into (2.35) produces an equation only in t_{*L}

$$\frac{1}{a} = \frac{\frac{t_{*L}^2}{a^2} - 1}{\frac{2t_{*L}}{a}(t_{*L} - 1)} \tag{2.37}$$

$$\Rightarrow \frac{1}{a} = \frac{\frac{t_{*L}^2}{a^2} - 1}{\frac{2t_{*L}}{a}(t_{*L} - 1)} \tag{2.38}$$

$$\Rightarrow 1 = \frac{2t_{*L}}{a^2} - \frac{t_{*L}^2}{a^2} \tag{2.39}$$

$$\Rightarrow 0 = t_{*L}^2 - 2t_{*L} + a^2 \tag{2.40}$$

$$\Rightarrow t_{*L} = 1 \pm \sqrt{1 - a^2}. \tag{2.41}$$

The discriminant is positive provided $a < 1$, which is violated only when the scan cannot be completed in the given time subject to the given angular frequency limit. As the transition must take place in the scan time $t_{*L} \in [0, 1]$ the negative sign is the natural solution hence

$$t_{*L} = 1 - \sqrt{1 - a^2} \tag{2.42}$$

is the transition time t_{*L}. Substituting again (2.34) into (2.35) provides simple statements of the quantities C_1 and C_2 in terms of t_{*L}:

$$C_1 = \frac{2t_{*L}}{a^2} \tag{2.43}$$

$$C_2 = 1 - C_1. \tag{2.44}$$

The instantaneous speed of the tip for this f is

$$v(t_*) = \sqrt{\dot{r}^2 + \left(r\dot{\theta}\right)^2} \tag{2.45}$$

and in the case $t_* = t_{*L}$ this is

$$v(t_{*L}) = \sqrt{\left(\frac{R}{aT}\right)^2 + \left((Rt_{*L})\frac{2\pi N}{Ta^2}\right)^2} \tag{2.46}$$

$$v(t_{*L}) = \frac{R}{aT}\sqrt{1 + \left(\frac{2\pi N}{a}\right)^2 t_{*L}^2} \approx \frac{\pi NR}{T}\frac{2t_{*L}}{a^2}. \tag{2.47}$$

One may now arrive at the explicit solution through specification of ω_L and T, thereby implying a, t_{*L} using 2.41, and the maximum achieved cantilever velocity using 2.47.

2.3.2.2 Characteristics and Comparison

Shown in figure 2.5 are the various properties of the OPT scan, as well as the result of scanning and inpanting in a real scan setting. Both the cantilever speed (a) and the frequency (b) remain under the limitations of the AFM. Furthermore, while the samples have a bias towards the center of the scan the majority of the scan area is sampled uniformly (c). As the frequency limit ω_L decreases, however, this area of higher sampling density will increase until the limiting case of the CAV scan. Specifically, when $v_L = \omega_L R$ the optimal scan transitions to a CLV spiral at approximately half radius. Furthermore, because the angular frequency constraint is obeyed for the OPT scan the travelled scan path (d) agrees well with γ_s. The inpainted scan result (e) has well-resolved features in all three regions (A-C).

Comparing this scan parametrization to the CLV and CAV solution presented in figure 2.4 it is clear that the OPT solution is superior. The insets from all three inpainted results

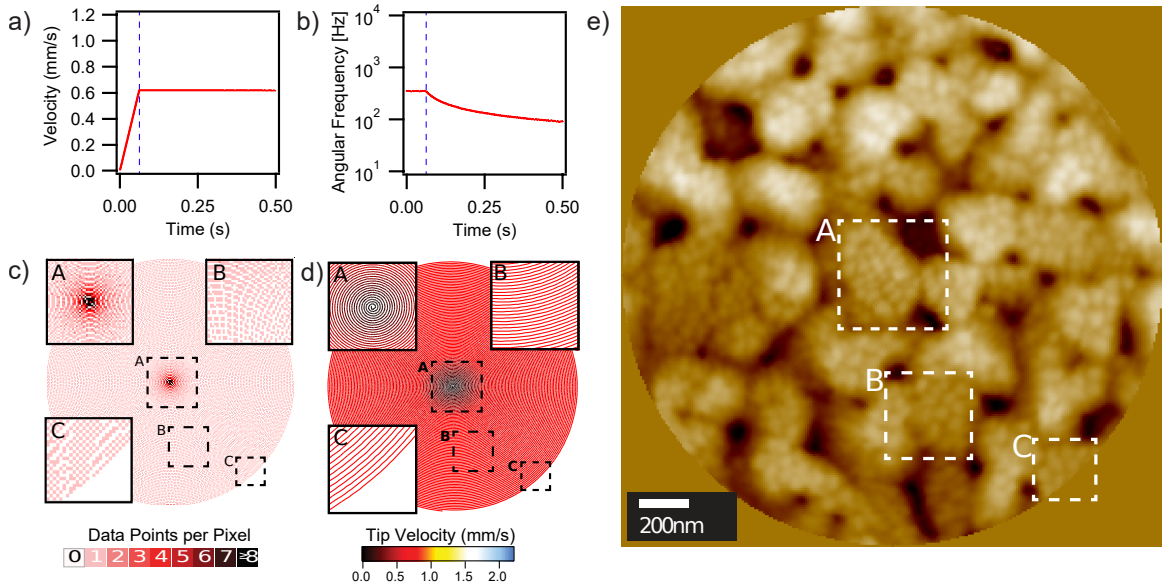

Figure 2.5: AFM scan and properties using the OPT parametrization [ZMA16]. Shown is the cantilever speed (a) and angular frequency (b) versus time, as well as sampling distribution (c) and observed scan path with color indicating cantilever speed (d). On the right is the inpainted result of the scan over an annealed gold sample collected using a Cypher ES AFM by Oxford Instruments. Figure is ©2016 IEEE.

are shown in figure 2.6 for easy comparison. When comparing quality, the outliers are the CLV's failure in the center (A) and the CAV's performance on the periphery (C). The OPT scan manages to resolve features well across all three regions through the simple transition that is provably the fastest solution.

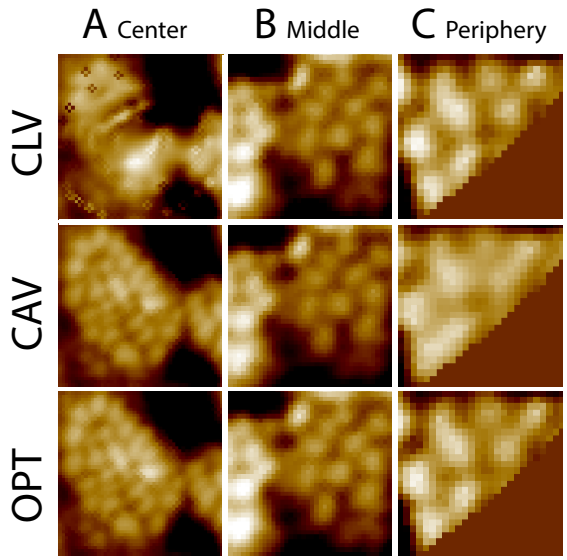

Figure 2.6: Insets from inpainted results for different parametrizations [ZMA16]. The three rows represent the use of CLV, CAV, and OPT to capture the same sample area with the same AFM in the same time. The three columns are the three regions highlighted in figure 2.4 and figure 2.5. Figure is ©2016 IEEE.

CHAPTER 3

Matrix Factor Models

3.1 EEG + fMRI Error Correction

3.1.1 Problem Overview

Electroencephalography (EEG) [DBM13] is an experimental technique that is capable of acquiring information about neural activity with very high temporal resolution but with a very low spatial resolution. The EEG consists of a network with on the order of one hundred potentiometers contacting a patient's scalp at known locations. The electrical signals received originate in the environment, muscles, or most desirably in the brain from neurological activity. Although the signal typically contains considerable noise and convolution due to the skull and environment, the information being captured is instantly available and can be recorded hundreds of times per second to monitor brain waves.

An entirely different technique for monitoring a subject's brain activity is functional magnetic resonance imaging (fMRI). The fMRI captures volumetric images of brain activity in three dimensions by measuring local changes in blood flow. While fMRI can provide a view into the structure of the brain, the captured activity is of a significantly lower temporal resolution requiring seconds to capture a frame. The advantage of the fMRI over EEG is the ability to see precisely where activity is taking place. Because of the relative strengths and weaknesses of the two techniques, a natural goal is to simultaneously capture both sources of information and combine the results into a temporally and spatially high resolution composite [DLF09, VSS09, TVR11].

Simultaneously capturing EEG and fMRI leads to distortions of the EEG signal resulting

from fMRI electromagnetic interaction. The most significant contribution of distortion arises from the switching magnetic field that produces a signal with orders of magnitude higher strength than the desired neural waveforms. This magnetic field gradient artifact, however, is well-understood and corrective techniques exist to subtract the primary distortion [GSE02].

A variety of non-nerual signals contribute to the EEG. Slight facial muscle movements, for example, clearly register in the data. A very challenging contribution to distinguish is the ballistocardiogram (BCG) [DSS07]. This is generated by the cardiac cycle in the subject producing blood flow and motion of the electrodes on the scalp. The effect is quasi-periodic yet difficult to remove due to variability in blood motion profile and frequency. Current methods exist to remove this artifact including cardiac r-wave timing [APK98], filtering [MAF07], independent component analysis (ICA) [JMH00], optimal basis sets (OBS) [NBI05], clustering [ZHJ12], and combined methods [DSS07]. These techniques can often be made automatic but frequently require user feedback during the filtering process, and in some cases require the acquisition of extra EEG information for a period of time to characterize the signal to be removed. This is undesirable because of the time requirements for both subject and the operator as well as the need to determine periods that can be used for artifact characterization.

Although there is not a single signature over time representing the BCG contribution to the EEG signal, we make the assumption that the BCG contribution is a linear combinations of a few characteristic profiles over the sensors. Common correction techniques currently in use involve signal decomposition into components followed by the selection and removal of those corresponding to the BCG. In the next section the proposed assumption that the BCG contributes only a few components to the signal is used to automatically discover and extract the BCG artifact using a low rank + sparse decomposition (LR+SD) [LCM10]. This is an approach for decomposing a matrix into the sum of two components, one being of low rank and the other having a small entry-wise 1-norm. In the subsequent section the application of such a decomposition to EEG data is presented followed by empirical results demonstrating automatic correction of the BCG artifact from a simultaneous EEG + fMRI experiment.

3.1.2 Low-Rank + Sparse Decomposition

Let n be the number of EEG channels and m the number of samples taken in time. Denote the captured EEG signals with $\tilde{f}_i(k)$ for $i = 1, ..., n$ the channel index and $k = 1, ..., m$ the sample index. Assume all that remains to be separated is the BCG contribution from otherwise desirable neural activity. The signal is thus a sum $\tilde{f}_i(k) = g_i(k) + f_i(k)$ where g_i is the contribution due to the BCG artifact and f_i is the clean, uncorrupted EEG signal. Assuming that $g_i(k)$ is generated by p factors that contribute to $\tilde{f}_i(k)$ with p signatures, there exist coefficients a_{ij} indicating the contribution of factor j of the BCG artifact to sensor i. Then the total contribution for $g_i(k)$ is given by

$$g_i(k) = \sum_{j=1}^{p} a_{ij} f_j^A(k) \tag{3.1}$$

with $f_i^A(k)$ the contribution of factor j for sample k. These relationships can be stated concisely as follows. Let $\tilde{\mathbf{F}}$ and $\mathbf{F}$ be n-by-m matrices with $\tilde{\mathbf{F}}_{ik} = \tilde{f}_i(k)$ and $\mathbf{F}_{ik} = f_i(k)$. Let an n-by-p matrix $\mathbf{A}_{ij} = a_{ij}$ and a p-by-m matrix $\mathbf{F^A}_{jk} = f_j^A(k)$. Then the relationship between the observed signals, the BCG components, the BCG coefficients, and the clean signal is

$$\tilde{\mathbf{F}} = \mathbf{A}\mathbf{F}^{\mathbf{A}} + \mathbf{F}. \tag{3.2}$$

This motivates the use of a low-rank + sparse matrix decomposition. The LR+SD model takes some matrix $\mathbf{X}$ and decomposes it into a sum $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$ with $\mathbf{Y}$ a matrix of low-rank and $\mathbf{Z}$ has small 1-norm

$$||\mathbf{Z}||_1 = \sum_{ij} |\mathbf{Z}_{ij}|.$$

The decomposition is found via the solution to the problem

$$\arg\min_{\mathbf{Y},\mathbf{Z}} ||\mathbf{Y}||_* + \lambda ||\mathbf{Z}||_1 \quad \text{such that} \quad \mathbf{X} = \mathbf{Y} + \mathbf{Z} \tag{3.3}$$

where the nuclear norm $||\mathbf{Y}||_*$ is a convex relaxation of the matrix rank defined as the sum of singular values of $\mathbf{Y}$. The model 3.3 therefore decomposes the EEG signals into $\mathbf{Y}$ that has a sparse vector of singular values and $\mathbf{Z}$ that has sparse coefficients. The parameter λ can be selected to determine the trade-off between the two variables. This problem can be solved using a variety of techniques including singular value thresholding, augmented Lagrangian methods, or accelerated proximal gradient. In this work the inexact augmented Lagrangian method (inexact ALM) is used as it very fast and is simple to implement.

To solve 3.3 using an augmented Lagrangian approach the energy is augmented with terms to enforce the constraint with dual parameter matrix $\mathbf{\Lambda}$

$$L(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = ||\mathbf{Y}||_* + \lambda||\mathbf{Z}||_1 + \langle \mathbf{\Lambda}, \mathbf{X} - \mathbf{Y} - \mathbf{Z} \rangle + \frac{\mu}{2}||\mathbf{X} - \mathbf{Y} - \mathbf{Z}||_F^2.$$

The algorithm to solve 3.3 using the technique of the augmented Lagrangian method involves iteratively minimizing L with respect to $(\mathbf{Y}, \mathbf{Z})$ then performing a dual ascent of $\mathbf{\Lambda}$. The sub-problem solving for $(\mathbf{Y}, \mathbf{Z})$, however, does not need to be solved exactly. In practice the inexact iteration that simply improves $\mathbf{Y}$ and $\mathbf{Z}$ slightly at each iteration is used because it is much faster:

1: $\mathbf{\Lambda} \leftarrow \text{sgn}(\mathbf{X})/\max(||\text{sgn}(\mathbf{X})||_F, \lambda^{-1}||\text{sgn}(\mathbf{X})||_\infty)$.
2: $\mathbf{Z} \leftarrow \mathbf{0}$.
3: **while** not converged **do**
4: $\quad (\mathbf{U}, \mathbf{S}, \mathbf{V}) \leftarrow \text{svd}(\mathbf{X} - \mathbf{Z} + \frac{1}{\mu}\mathbf{\Lambda})$.
5: $\quad \mathbf{Y} \leftarrow \mathbf{U}\mathcal{S}_{\mu^{-1}}[\mathbf{S}]\mathbf{V}^\dagger$.
6: $\quad \mathbf{Z} \leftarrow \mathcal{S}_{\lambda\mu^{-1}}[\mathbf{X} - \mathbf{Y} + \frac{1}{\mu}\mathbf{\Lambda}]$.
7: $\quad \mathbf{Y} \leftarrow \mathbf{Y} + \mu(\mathbf{X} - \mathbf{Y} - \mathbf{Z})$.
8: $\quad \mu \leftarrow \rho\mu$.
9: **end while**.

The inexact ALM method is the standard augmented Lagrangian technique with only a single iteration of the sub-problem executed per dual ascent step.

3.1.3 Experimental Validation

Twenty individuals between the ages of 23 and 30 participated in this study, with written consent and UCLA IRB approval, in which concurrent EEG and fMRI was recorded through multiple stimuli at the Staglin IMHRO Center for Cognitive Neuroscience at UCLA. In the experiment subjects viewed 140 Gabor flashes with periods of 13.85 ± -2.8 seconds presented using an MR projector screen. This experiment is known to cause occipital event-related spectral perturbations (ERSPs) in the alpha spectral band (8-12 Hz). EEG was collected using a 256-channel GES 300 Geodesic Sensor Net by Electrical Geodesics, Inc. at a rate of 500 Hz. MRI clock signals were synchronized with the EEG data collection for magnetic gradient artifact removal. fMRI was concurrently collected using a 3-T Siemens Trio MRI Scanner with echo planar imaging gradient-echo sequence and an echo time of 25 ms, repetition time of 1s, 6mm slices, 2mm gap, flip angle of 90°, 3mm in-place resolution, and ascending acquisition. The magnetic gradient artifact was removed from the EEG signals through subtraction of an exponentially weighted moving average template.

One existing technique in practice for BCG artifact removal is the InfoMax ICA cleaning method included with Brain Analyzer v2.0.2 from Brain Products. This approach requires manual selection of cardiac signals that are used to identify independent components extracted using ICA that correlate with the cardiac signal. In addition to the concurrent EEG with fMRI acquisition the same experiment was performed outside of the MRI for control EEG data without BCG corruption. In figure 3.1 shows the alpha band activity over the patient's scalp averaged over events in the experiment for each technique: EEG+fMRI cleaned with ICA, EEG+fMRI cleaned with the proposed method, and finally the control. An ideal method, as in the control data, should result in a clear absence of alpha band activity for τ_2 which is immediately post-flash. This can also be observed in the averaged alpha band activity over all flashes on the right that demonstrates, ideally, a decrease in value after the flash time 0. Clearly the desired drop in activity is observed for both the control and proposed signals while there is little to no observable phenomenon for the ICA cleaned data.

Further evidence is found on examination in the full spectrum of the output for each method. In figure 3.2 the magnitude of the continuous wavelet transform (CWT) is averaged for a window of time surrounding each flash event. The result clearly demonstrates the removal of noise by the proposed method. In both uncleaned and ICA results there is significant activity across the spectrum for all time, but filtering using the proposed technique produces clear evience of alpha band activity with a decrease immediately following the time of the flash.

3.2 Hyperspectral Unmixing

3.2.1 Problem Overview

In the same way that standard photography images have red, green, and blue channels, hyperspectral images are captured with far more color channels numbering in the hundreds. Each channel represents the intensity of light at a specific frequency scattered off an object. Thus, while a human eye is restricted to threelight intensities, hyperspectral imaging devices are able to measure light intensity in very fine detail ranging from well below through infrared to well above ultraviolet. Each frequency of light observed is one *spectral band* captured by the imaging apparatus.

The human eye differentiates objects and materials using colors available, and analogously a hyperspectral image can be used to distinguish objects with high efficacy. For example, to the human eye grass is green and asphalt is black hence the problem of determining where in an image one or the other is present is easily solved. *Hyperspectral unmixing* [KM02] is a natural problem that arises from the analysis of hyperspectral images concerned with automatically discovering what materials are present in each pixel in general and for many spectral bands. Because hyperspectral images contain fine spectral information, materials that may appear the same to a human eye can frequently be distinguished with the additional spectral bands. For example, many types of vegetation appear to the human eye simply as green but hyperspectrally plants often have different disgnatures [BL06b]. In this work

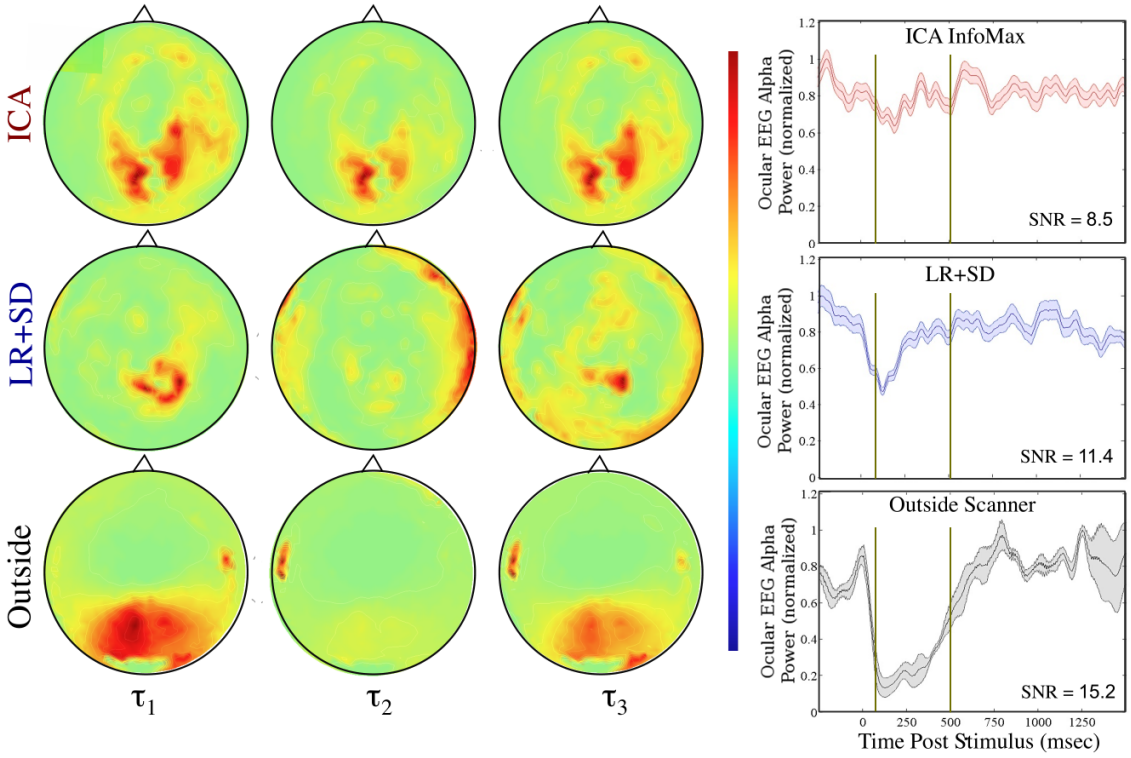

Figure 3.1: Alpha-band activity at three times pre-, mid-, and post-anomaly averaged over all epochs, for three experimental designs [GMD14]. The three designs are EEG + fMRI with the BCG artifact removed using ICA (top row), EEG + fMRI cleaned using the proposed technique (middle row), and the control EEG collected with no fMRI (bottom row). The time of activity displayed is 500msec pre-flash (τ_1), 50msec post-flash (τ_2), and 500msec post-flash (τ_3) for the activity maps on left, and a window of average alpha-band activity for ocular electrode 118 is on the right. SNR is the ratio of the signal extent from 0ms to 500ms to the standard deviation of alpha power from 0ms to 1000ms.

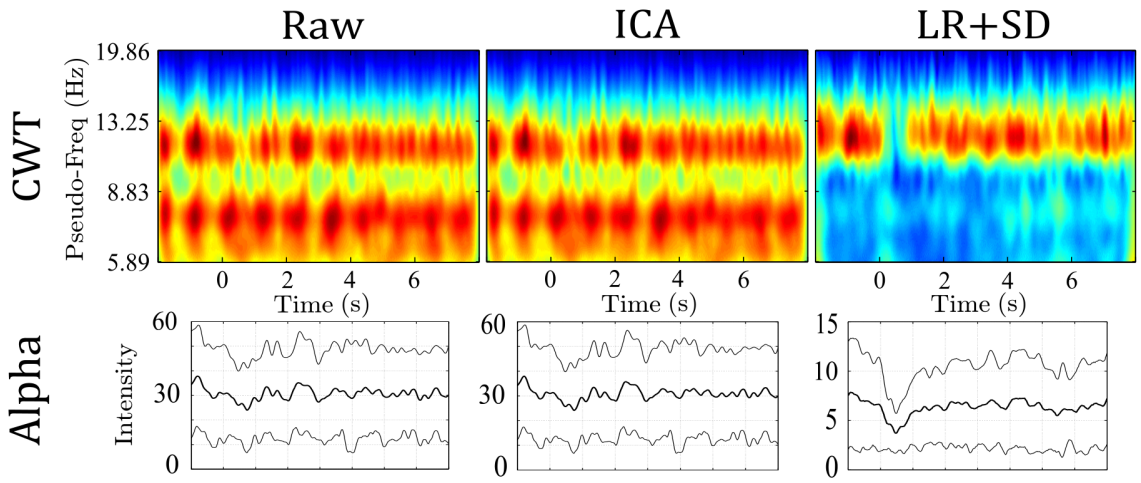

Figure 3.2: CWT frequency intensity/activity averaged over all events for one ocular electrode (118) during an experiment (top row) and alpha band specifically at 10Hz with standard deviations (bottom row) [GMD14]. Each column represents an experimental design, from left: no BCG artifact removal, BCG artifact removal with ICA, and finally BCG artifact removal with the proposed technique.

it is assumed that the number of materials, m, is known prior to unmixing however the endmembers, or "colors", for each material are yet to be extracted.

Complicating this unmixing task is the inevitable presence of multiple materials within each pixel. This results in a per-pixel mixture of spectral signatures. Because hyperspectral images are frequently captured by aircraft and satellites, each pixel often captures light emitted from within an area larger than one square meter. Multiple materials present in this area contribute to the resultant "color", or hyperspectral signature, of that pixel in the image. The unmixing problem therefore is a mathematical inverse problem with the goal of extracting from a hyperspectral image information about what materials are present, and in what abundance, within each pixel.

The information contained in a single hyperspectral image may be described by an s-by-n matrix $\mathbf{H}$ where s is a number of spectral bands as determined by the camera and n is the number of pixels in the hyperspectral image. The number of spectral bands measured may

range from tens to hundreds while n may be very large at 10^6 or more pixels. As a result, $\mathbf{H}$ is a very large matrix. The result of the unmixing problem is typically quantified by a collection of spectral signature vectors $\mathbf{w}_i \in \mathbb{R}^s$ for $i = 1, 2, ..., m$ where m is the number of materials. In addition, the unmixing method seeks abundance vectors $\mathbf{a}_j \in \mathbb{R}^n$ for each material j indicating the quantity of that material in each pixel of the original image. Let $\mathbf{M}$ be the matrix formed by placing $\mathbf{w}_i$ in columns and $\mathbf{A}$ be the matrix formed by placing $\mathbf{a}_i$ within rows. One possible formulation for the hyperspectral unmixing problem therefore can be stated

$$\mathbf{H} = \mathbf{MA} + \mathbf{N} \tag{3.4}$$

where $\mathbf{N}$ is noise resulting from the imaging apparatus. Within this framework there is some matrix $\mathbf{A}$ of abundances and another matrix $\mathbf{M}$ of spectral signatures that determine the final image data through a matrix product. This formulation is utilized by the method of linear spectral mixture analysis [ASJ86] through two stages. First, the material signatures, known as *endmembers*, are extracted from the data $\mathbf{H}$ to produce $\mathbf{M}$. With $\mathbf{M}$ now known, the matrix $\mathbf{A}$ is discovered in a second step through a fitting procedure. Two restrictions are typically placed on $\mathbf{A}$ for physical reasons. The abundance non-negativity constraint (ANC) permits only non-negative entries in this matrix while the abundance sum constraint (ASC) requires the columns to have unit sum. The union of these two conditions on $\mathbf{A}$ result in the fully-constrained least squares unmixing (FCLSU) problem for finding $\mathbf{A}$ [Cha03].

A multitude of prior techniques exist for endmember extraction approximately belonging two categories. The first category assumes that for each material there exists a pure pixel in the image containing only that material. This *pure pixel assumption* results in a variety of techniques [ND05b, Win99] that attempt to select from columns of $\mathbf{H}$ as spectral endmembers, though for images with low spatial resolution this assumption may not hold. An alternative category of techniques [LB08, Cra94, CCH09] do not depend on this pure pixel assumption. With such methods material endmembers form the extrema of a convex minimum-volume simplex containing the columns of $\mathbf{H}$.

$$\min_{\mathbf{A}} \frac{1}{2}||\mathbf{H} - \mathbf{MA}||_F^2. \tag{3.5}$$

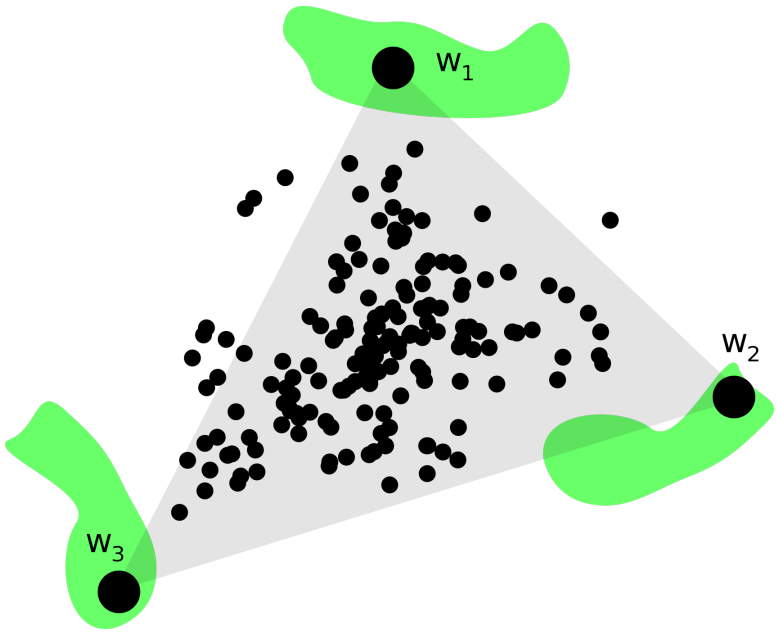

Figure 3.3: Material variability. The pixels in the hyperspectral image are a point cloud contained in a convex hull formed by material endmembers w_1-w_3. Under the assumption of material variability, the material representatives are not points but rather belong to some subspace that must be extrated as well.

Difficulties may arise in all cases, however, due to an issue known as *spectral variability* [ZH14, HDT14]. Spectral variability is the phenomenon in which the same material can have more than one spectral signature. A simple scenario in which this arises is due to illumination [ND05a] – a hill generally is more illuminated on one side than another as result of the position of the sun. The more illuminated side emits a spectral signature that is linearly scaled compared to that of the darker side. This, in addition to other variability sources such as grain size with gravels and material moisture, complicates the above process for hyperspectral unmixing. Endmember extraction techniques are originally only designed to find a single endmember for each material, a clear inadequacy if materials can present with multiple signatures. To adapt to this variability it is necessary to re-consider the model (3.4) and correctly characterize mathematically the desired result.

In the remainder of this section the work of [SZP12] is reviewed, an existing response

to the challenge of spectral variability. After that, the proposed contribution is presented as an augmentation of their work with the enforcement of *social sparsity* [KSD13]. Social sparsity a mathematical tool used to enforce sparsity among variables in which groupings of the variables are known to exist. Finally, the section is concluded with a presentation of computational results that demonstrate the efficacy of the proposed augmented hyperspectral unmixing model.

3.2.2 Existing Method

In [SZP12] the authors propose using any existing endmember extraction technique to extract endmembers from $\mathbf{H}$, repeating the process on random subsets of columns. This produces, for each run, some collection of endmembers extracted from a subset of image pixels. After accumulating the resulting endmembers from all runs into columns of a new matrix $\mathbf{M}$, the resultant matrix will contain more endmembers than there are materials. While some may be redundant, the selection of random pixel subsets is meant to produce representatives for each state of each material. At this point endmembers are clustered so that each cluster contains representatives, ideally, of different spectral signatures for a single material. Abundance maps are then estimated using this over-determined set of endmembers via the FCLSU fitting procedure and the final abundance maps for each material can be found by accumulating the abundances of each endmember within a cluster.

The VCA algorithm [ND05b] is used in this work to extract endmembers due simply to popularity, though in general VCA can be substituted with any other endmember extraction technique. VCA operates through projections of the pixel vectors onto the plane where, noting the pure pixel assumption, desirable endmembers necessarily occupy the role of extrema defining a triangle in the plane containing all other pixels. Viewing the data with various projections therefore identifies endmembers that form a simplex around columns of $\mathbf{H}$.

The entire method of [SZP12] is as follows. Let k be some number of times VCA is applied, a parameter to the method. Each application of VCA to random subsets of the

columns of $\mathbf{H}$ produces m endmembers so the final matrix $\mathbf{M}$ has km columns. Next, a clustering technique, for example spectral clustering [NJW02] with spectral angle similarity use in this work, is applied to group the extracted endmembers of $\mathbf{M}$ into m clusters. The use of the spectral angle as a measure of similarity in this work originates from the assumption that material variability arises due to illumination differences [ND05a]. Different assumptions pertaining to the anticipated material variability will imply using different clustering methods and similarities. This clustering produces a partition $G_1, G_2, ..., G_m$ of endmembers such that $\{\mathbf{M}_{:,j} \; \forall \; j \in G_i\}$ is the set of spectral endmembers for the the i^{th} material. The final step of the technique determines the abundance matrix with FCLSU, quantitatively stated as the problem

$$\min_{\mathbf{A}} ||\mathbf{H} - \mathbf{M}\mathbf{A}||_F^2 \tag{3.6}$$

subject to the FCLSU conditions enforcing $\mathbf{A}$ non-negative with unit-sum columns. Here $||\cdot||_F$ is the matrix Frobenius norm,

$$||\mathbf{Y}||_F = \sqrt{\sum_{ij} \mathbf{Y}_{ij}},$$

that naturally arises from the assumption of Gaussian data noise.

The over-determined endmember matrix $\mathbf{M}$ in connection with the nature of the Frobenius norm means that in general the mixture for each pixel will be dense. That is, each pixel is predicted to contain a small amount of many endmembers and, as a result, materials. A natural assumption to enforce is the selection of only a few endmembers [ZWF14, LSK12]. This is physically reasonable – each pixel may have a few materials but in general will not contain all materials, thus the matrix $\mathbf{A}$ should be sparse *in some sense*. Existing models [QJZ11] for $k = 1$ are extended to the $k > 1$ case in what follows via social norm penalties.

3.2.3 Proposed Method

The material clustering information can be incorporated into 3.6 by using social norms described in [KSD13]. Because G is known prior to solving 3.6, the assumption of each pixel being a mixture of a few materials can be enforced through penalty. This is demonstrated

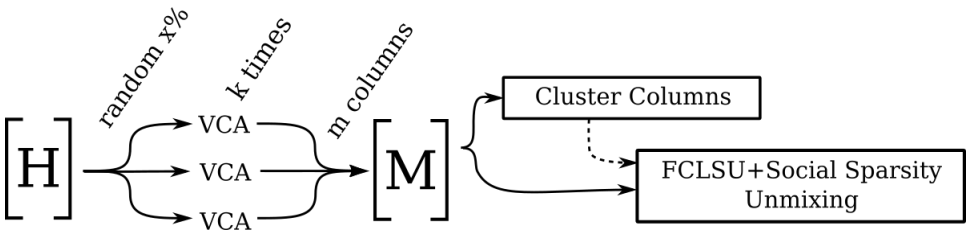

Figure 3.4: Workflow of the proposed technique. The additional step proposed involves using the clustering information within the unmixing stage via social sparsity, as denoted with the dotted line. Figure is ©2016 IEEE.

in figure 3.4 where the addition of a dotted arrow demonstrates the new information being used during unmixing. Let $\mathbf{x} \in \mathbb{R}^s$ be some vector and partition the set $\{1, 2, ..., km\}$ into m groups G_i for $i = 1, 2, ..., m$, as was done in the previous section using spectral clustering with the spectral angle measure. The vector ℓ_p norm of $\mathbf{x}$ is given by

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{s} |\mathbf{x}_i|^p \right)^{\frac{1}{p}}$$

with the cases of $p = 1$ and $p = 2$ very common, and limiting behaviours of $p \to 0$ and $p \to \infty$ approaching the number of non-zero entries of $\mathbf{x}$ and the maximum absolute value of all entries, respectively. The group pq-norm given the partition G is given by

$$||\mathbf{x}||_{G,p,q} = \left(\sum_{i=1}^{m} ||\mathbf{x}_{G_i}||_p^q \right)^{\frac{1}{q}}. \tag{3.7}$$

This can be generalized to matrices by summing the application of (3.7) to each column which, for a penalty parameter λ, results in the proposed model

$$\min_{\mathbf{A}} \frac{1}{2} ||\mathbf{H} - \mathbf{M}\mathbf{A}||_F^2 + \lambda \sum_{i=1}^{n} ||\mathbf{A}_{:,i}||_{G,p,q}. \tag{3.8}$$

The cases considered are the group lasso $(p, q) = (2, 1)$, elitist lasso $(p, q) = (1, 2)$, and a fractional case $(p, q) = (1, \frac{9}{10})$. Group lasso tends to select a few groups, in this case materials, and within groups it prefers a dense mixture of members. The elitist lasso selects a dense mixture over the groups and within each group selects a few representative "elites". The

final fractional lasso selects a few groups, similar to the group lasso, but it does so without preferring a dense mixture over groups. Because (3.8) is subject to unit-sum columns, the addition of these penalties can appear contradictory. For example, if each material has a single endmember the group lasso has no influence. This constraint also makes the traditional sparsity-enforcing lasso penalty nonsensical; a fraction of 9/10 can enforce sparsity even with constrained abundance at the expense of non-convexity.

One can solve (3.8) using the alternating direction method of multipliers, or ADMM [BPC11], that allows the constraints and complex penalty term to be split into distinct and easily calculable stages. To solve (3.8), first write the problem in a slightly different way. Consider the optimization problem

$$\min_{\mathbf{X},\mathbf{Y},\mathbf{Z}} \frac{1}{2}||\mathbf{H}-\mathbf{M}\mathbf{X}||_F^2 + \lambda||\mathbf{Y}||_{G,p,q} \tag{3.9}$$

subject to the constraints that $\mathbf{X} = \mathbf{Z}$, $\mathbf{X} = \mathbf{Y}$, $\mathbf{X}$ has unit-sum columns and $\mathbf{Z}$ is non-negative. This equivalent problem is a ready form for ADMM with multiplier variable matrices α and β that produce the augmented Lagrangian formulation

$$\begin{aligned}\min_{\mathbf{X},\mathbf{Y},\mathbf{Z}} \frac{1}{2}||\mathbf{H}-\mathbf{M}\mathbf{X}||_F^2 + \lambda||\mathbf{Y}||_{G,p,q} \\ +\langle\alpha, \mathbf{X}-\mathbf{Z}\rangle + \frac{\rho}{2}||\mathbf{X}-\mathbf{Z}||_F^2 \\ +\langle\beta, \mathbf{X}-\mathbf{Y}\rangle + \frac{\rho}{2}||\mathbf{X}-\mathbf{Y}||_F^2\end{aligned} \tag{3.10}$$

subject to the constraints of $\mathbf{X}$ with unit-sum columns and $\mathbf{Z}$ non-negative. Minimization with respect to $\mathbf{Z}$ is trivial as the problem is separable for each coordinate of the matrix, hence $\mathbf{Z}$ is updated by

$$\mathbf{Z} \leftarrow \left(\mathbf{X} + \frac{\alpha}{\rho}\right)_+$$

with $(\cdot)_+$ indicating the coordinate-wise positive part.

The update for $\mathbf{X}$ is slightly more difficult due to the unit-sum constraint, however this

only requires adding n multipliers μ_i

$$\begin{aligned}\min_{\mathbf{X},\mathbf{Y},\mathbf{Z}} \frac{1}{2}||\mathbf{H}-\mathbf{M}\mathbf{X}||_F^2+\lambda||\mathbf{Y}||_{G,p,q} \\ +\langle\alpha,\mathbf{X}-\mathbf{Z}\rangle+\frac{\rho}{2}||\mathbf{X}-\mathbf{Z}||_F^2 \\ +\langle\beta,\mathbf{X}-\mathbf{Y}\rangle+\frac{\rho}{2}||\mathbf{X}-\mathbf{Y}||_F^2 \\ +\sum_j \mu_j\left(\left(\sum_i \mathbf{X}_{i,j}\right)-1\right)\end{aligned} \tag{3.11}$$

which results in a linear system of equations given by

$$(\mathbf{M}^\dagger\mathbf{M}+2\rho I)\begin{pmatrix}\mathbf{X}\\ \mu\end{pmatrix}=\begin{pmatrix}-\mathbf{M}^\dagger\mathbf{H}+\alpha+\beta-\rho(\mathbf{Z}+\mathbf{Y})\\ 1\end{pmatrix} \tag{3.12}$$

where μ indicates the row vector of entries μ_i and 1 indicates a matrix of ones with the same shape as μ. This system is easily invertible and, conveniently, the system matrix is state-independent.

The last subproblem for $\mathbf{Y}$ requires the use of a group shrinkage operation $\mathcal{S}_{G,p,q}$, described in [KSD13] with approximate fractional *p-shrinkage* as used in [Cha09], denoted by

$$\mathbf{Y}=\mathcal{S}_{G,p,q}\left(\mathbf{X}+\frac{1}{\rho}\beta,\frac{\lambda}{\rho}\right). \tag{3.13}$$

The full iterative scheme, along with dual updates, is therefore

1: Initialize $\alpha\leftarrow 0$, $\beta\leftarrow 0$.
2: Randomly initialize $\mathbf{X}$.
3: $\mathbf{Y}\leftarrow\mathbf{X}$.
4: $\mathbf{Z}\leftarrow\mathbf{X}$.
5: **while** not converged **do**
6: $\quad\mathbf{Z}\leftarrow\left(\mathbf{X}+\frac{\alpha}{\rho}\right)_+$.
7: $\quad\mathbf{X}\leftarrow$ the solution of (3.12).
8: $\quad\mathbf{Y}\leftarrow\mathcal{S}_{G,p,q}\left(\mathbf{X}+\frac{1}{\rho}\beta,\frac{\lambda}{\rho}\right)$.
9: $\quad\alpha\leftarrow\alpha+\rho(\mathbf{X}-\mathbf{Z})$.

10: $\beta \leftarrow \beta + \rho(\mathbf{X} - \mathbf{Y})$.

11: **end while**.

For numerical demonstrations $\rho = 10$, requiring approximately a thousand iterations to converge sufficiently. The abundance estimation with social sparsity requires on the order of minutes to unmix data of size 100-by-100-by-56. The major challenge is the sum constraint that conflicts with the sparsifying penalties resulting at times in slow convergence.

3.2.4 Experimental Validation

The complete hyperspectral unmixing scheme proposed was applied to three datasets to evaluate performance. The first data set known as "cuprite" was generated using the AVIRIS cuprite data set via the method proposed in [HBG15] with added artificial spectral variability [DHV15]. That is, the abundance maps from real observations are used to generate through a stochastic process an artificial data set with variability present. The second dataset "islands" is entirely artificial also with endmember variability introduced. Finally, the "stadium" data set is a hyperspectral image taken of a football stadium from above with a variety of materials and illumination variability present. This data set has no ground truth and is used as a qualitative study. All three data sets are presented in figure 3.5with example abundance maps taken from the cuprite and islands data sets as well as an approximately RGB representation of the stadium data.

The mean pixel error

$$\mathrm{E}_{\mathrm{Model}} = \frac{1}{\#\ \mathrm{Pixels}} \sum_{\mathrm{Pixels}\ i} \sqrt{\frac{1}{\#\ \mathrm{Materials}} ||\mathbf{a}_i - \tilde{\mathbf{a}}_i||_2^2}$$

with a_i and $\tilde{a}_i$ the actual and approximate abundance maps for each pixel i can be used to quantify the agreement of predicted abundance maps using the proposed technique with known values. The error ratio $\mathrm{E}_{\mathrm{Model}}/\mathrm{E}_{\mathrm{FCLSU}}$ is used specifically to measure the relative change in performance for each type of sparsity enforcement where $\mathrm{E}_{\mathrm{FCLSU}}$ is the error of the unpenalized $\lambda = 0$ model. For these data sets $k = 5$ and VCA is applied to random subsets of 80% of the data using 100 iterations.

Optimizing over selection of λ for both synthetic data sets produces the mean pixel errors in table 3.7 as a percent of the $\lambda = 0$ model. Also shown is the "batchless" unmixing result when $k = 1$ and $\lambda = 1$, essentially not modelling material variability at all. Across the board some improvement is found with the group lasso generally performing the best. The fractional lasso also demonstrates a slight improvement. Unfortunately, the elitist lasso performs the worst on the cuprise data set. Furthermore, no improvement was demonstrated on the islands data set for positive values of λ. This is likely because the elitist lasso does not in fact enforce the desired type of sparsity between, rather than with, the material groups. Oddly enough, the most improvement for the islands data set was achieved with a negative value of λ and the elitist lasso. While this is interesting and indeed analytically there is reason to believe a negative elitist lasso may result in the desired sparsity properties the algorithm and model provide no guarantees in this case and therefore it is presented simply as a curious development.

The result of processing the stadium data set is presented in figure 3.6 for qualitative analysis of the behaviour of each norm. The group lasso enforces sparsity across the groups, thereby clearing up ambiguities that arise. For example, the parking area in the lower-left corner is made of asphalt however the FCLSU result generates a dense mixture with very small contributions from all materials. The group lasso forces fewer materials to contribute to each pixel which improves these areas. Similar results are observed with the fractional lasso though the effect is slightly different with significantly more low-abundance contributions from material groups being eliminated. The elitist lasso does not seem to improve the results.

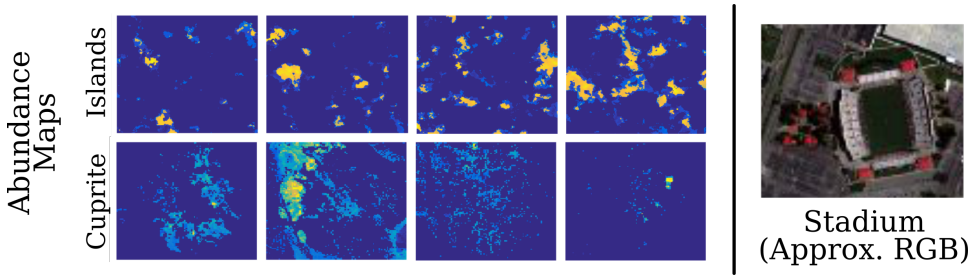

Figure 3.5: Data sets considered. The the islands synthetic data (top row) is a 100-by-100-by-56 hyperspectral cube and the cuprite synthetic data (bottom row) has a 100-by-100-by-47 data cube, therefore in both cases $\mathbf{H}$ has $10,000$ columns. Shown are the known exact abundance maps for four materials out of a respective total of 15 and 14 simulated materials. The third data set is a 105-by-128-by-144 hyperspectral cube of the since demolished Robertson stadium at the University of Houston. Shown is an approximate RGB representation. Figure is ©2016 IEEE.

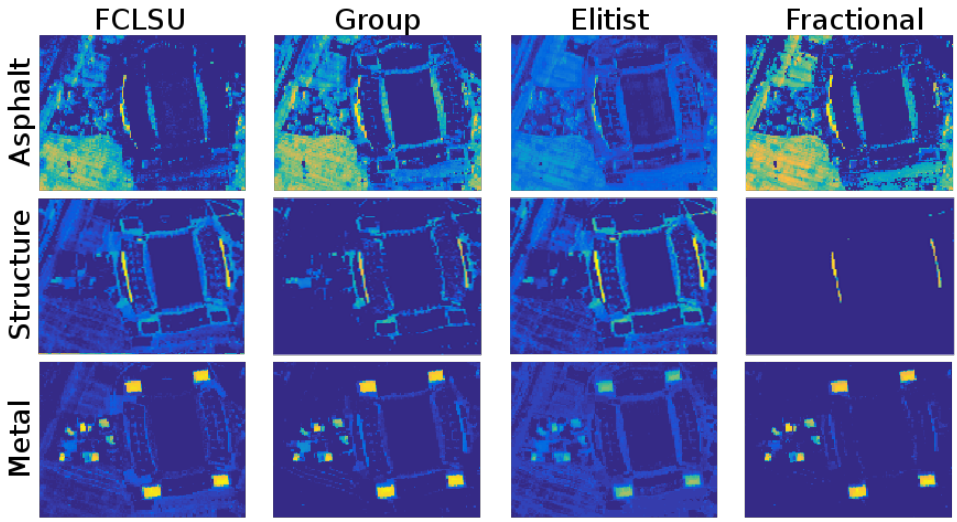

Figure 3.6: Stadium data set extracted abundance maps. Material abundance maps corresponding to (top row) asphalt, (middle row) a painted structure, and (bottom row) metal roofing. Shown is the result for each of unpenalized FCLSU, group lasso, elitist lasso, and fractional lasso. Figure is ©2016 IEEE.

Set	Group	Elitist	Fractional (0.9)	Batchless
Cuprite	88.2%	96.4%	93.2%	203.5%
Islands	94.9%	*94.4%	97.8%	565.0%

Figure 3.7: Mean pixel errors for each of three types of sparsity enforcement algorithms and FCLSU without bundles applied to two synthetic data sets. Shown is the error as a percent of the unpenalized bundle model. (*) performance achieved with a negative value of λ.

CHAPTER 4

Topic Models

4.1 Introduction

Topic models [Ble12] are used to find trends in a collection, or corpus, of documents. For the purposes of these models, a document is simply a set of word tokens taken from a dictionary and the collection of documents is therefore a collection of these sets. While such models were originally produced with the intention of understanding text documents, the mathematics is generally applicable to any data that can be described as a collection of sets with each set a collection of tokens from a dictionary. Denote by $w_1, ..., w_m$ the words of the dictionary a topic model will use, m being the total number of unique words appearing in all documents. Each document is, to the topic model, a collection of these words. This is known as the "bag-of-words" assumption as it discards the order of words in the document thereby reducing the document to simply a "bag" of tokens.

A topic model finds trends in a corpus by finding words that tend to co-occur in the same documents. An example of how topic modeling results may be presented is shown in table 4.1. Each column represents a topic, and for each four top words are displayed. The title for each topic is the interpretation given by the authors in [BNJ03]. The topic model extracted from the corpus, in this case new articles, the fact that "film" and "music" tend to appear in the same documents. Similarly, "school" and "education" also were found to be related in that documents containing one of these words tended to contain the other. In practice topic models can be applied to document collections numbering in the millions and can extract hundreds of topics each with dozens of meaningful top words.

Arts	Budgets	Children	Education
news	million	children	school
film	tax	women	students
show	program	people	schools
music	budget	child	education

Table 4.1: Top words learned by a topic model for four topics [BNJ03] extracted from a collection of news articles. The column labels are descriptions chosen by the authors.

The first topic model to come to widespread popularity is latent Dirichlet allocation (LDA) [BNJ03]. LDA is a Bayesian generative model for words appearing in documents. That is, LDA supposes that a set of yet-unknown probability distributions along with a prescribed sampling process is responsible for producing each word in each document. Through various techniques a user can find these distributions and from them generate information such as the to words in table 4.1. Another technique arising from an entirely different collection of research is non-negative matrix factorization (NMF) [Paa97, LS99]. NMF is used to write a non-negative matrix as the product of two non-negative matrices of much smaller size. In finding these two matrices that best reproduce the original matrix, NMF is also able to recover information such as the top words shown in table 4.1. While both techniques seem to solve the topic modeling problem with varying degrees of efficacy, they appear to be completely unrelated in the literature. NMF is presented as an energy model for matrix decomposition while LDA is a Bayesian statistical model for the generation of words. It is the purpose of this work to understand these techniques as well as their inference methods from a common perspective and with a common language of optimization and energy models.

Topic models begin by transforming the data into a large non-negative *word count* matrix. Let $w_1, ..., w_m$ denote the dictionary of words appearing in the corpus where m is the number of unique dictionary words. With something known as the *bag-of-words* assumption, each document is viewed as a bag of these tokens and therefore the entire corpus can be transcribed into a large matrix, the *word count matrix*, with entry (i, j) equal to the number

of appearances of word i from the dictionary in the j^{th} document. As a result, this matrix is non-negative and is m-by-n where n is the number of documents in the corpus. Both LDA and NMF have a parameter t that must be set beforehand to specify the number of topics to learn from the corpus. While the result and mechanism of LDA and NMF are certainly not identical, both ultimately produce t functions that map the dictionary $\{w_i\}$ to non-negative values. The top words are simply the words for each topic that are mapped to the largest values as these are most relevant to that topic. Thus to be more explicit, shown in table 4.1 are the four words that map to the largest value for each of four topics learned using LDA.

In the next section we present the general connections between probability formulations such as LDA and energy formulations such as NMF through the negative log-likelihood transform. We then summarize in detail the two model formulations separately and in their own languages. After this, we summarize the model-level connections between LDA and NMF as understood in the literature. Unfortunately, topic models are non-convex problems and therefore different inference techniques applied to the same problem can produce different results. We therefore investigate three inference techniques - two for LDA and one for NMF - from the perspective of energy minimization to understand their similarities and differences without the barriers of language and notation. Finally, we conclude this chapter with a variety of numerical experiments that highlight findings from the analytic study, reveal the general effectiveness of topic modeling techniques, and demonstrate the utility of topic models when trying to understand a complex data set comprised of tweets taken from Madrid city over a year. To assist with this study of Madrid through tweets, we present various metrics that capture important properties of the learned topics with respect to both the location and time of tweets.

4.2 Probability and Energy Frameworks

Two dominant frameworks for handling general data problems, and in our case topic models, are Bayesian statistical models and energy models. Bayesian models, such as LDA, are

based on the analytic construction of a posterior probability followed by the application of numerical techniques to study the probable values, and uncertainty, of model parameters. This approach is stated most simply using Bayes' law [BC16],

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{4.1}$$

where θ represents model parameters to be found and X is the data observed. Because the data X is observed and fixed, $P(X)$ is simply a constant and can be ignored. The crucial elements of a Bayesian model are analytically describing a *likelihood* $P(X|\theta)$ and *prior probability* $P(\theta)$ that are sensible and tractable. Numerical algorithms are then applied to $P(\theta|X)$ to understand which models fit the data. For example, one way to numerically study the posterior distribution (4.1) is by finding the most probable model known as the maximum a posteriori (MAP) estimate

$$\theta_{\text{MAP}} = \arg\max_{\theta} P(\theta|X) \tag{4.2}$$

Skilled Bayesian statisticians are able to construct intricate models that can be easily studied numerically and fit observations well. The prime example of a Bayesian topic model is LDA itself which is discussed in further detail in section 4.4. For LDA, the likelihood $P(X|\theta)$ captures similarity of the word count matrix X to a generative probabilistic model with parameters given by θ. The underlying model θ consists of t discrete distributions over the dictionary words as well as t discrete distributions over documents. The prior probability $P(\theta)$ is used to place assumptions on these distributions. In the case of LDA, for example, the prior captures the probability that these distributions were sampled from a Dirichlet distribution which is explained in detail further on. The MAP estimate (4.2) can be used with LDA [VP14, AWS09] though because such methods do not retain uncertainty information MAP estimation is not as common as other techniques discussed later.

The second framework for data problems uses no statistical notation or probability models and is responsible for NMF which is discussed in section 4.3. Energy models are techniques for cleaning, decomposing, and studying data through the design of an energy functional

consisting of a sum over terms. These terms are used to enforce similarity of the data to the model as well as seek desirable geometric properties of parameters. Again taking θ to be some collection of parameters to be learned, an energy model is simply a function $E(\theta)$. The problem is solved by finding the values θ that minimize this energy function. In contrast with the Bayesian framework, mathematicians studying energy models are concerned only with guaranteeing a minimum exists and is unique, as well as with finding the minimum quickly through optimization techniques. That is, uncertainty and probabilistic questions are rarely considered. In the context of topic modeling with NMF, θ is a parameter space consisting of two matrices [XLG03]. The energy E for NMF has one term that measures deviation of $\mathbf{X}$ from the product of these matrices, and therefore the solution is the matrix product that most closely approximates $\mathbf{X}$.

These two frameworks exist in almost entirely different literature with almost entirely different researchers speaking in entirely different languages. However, as is summarized in section 4.5, the underlying structure of probabilistic and energy models can be very similar in cases such as LDA and NMF. Specifically, in some cases $P(\theta|X)$ and $E(\theta)$ can be seen as interchangeable representations of the same problem where the MAP estimate is precisely the energy minimization problem. The remainder of this section outlines the general analytic connection between operations on a model probability and operations on an energy function.

Foremost an energy can in some cases be transcribed into a probability distribution. Suppose that

$$\int e^{-E(\theta)} d\theta \tag{4.3}$$

exists and equals c. Then the energy E is the *negative log-likelihood* of a probability distribution P

$$E(\theta) = -\log\left(cP(\theta)\right). \tag{4.4}$$

When $c = 1$ the energy will be called *probabilistic*. If E depends on more arguments ζ, but

$$\int e^{-E(\theta,\zeta)} d\theta = 1 \tag{4.5}$$

for all ζ then E will be called *probabilistic in* θ.

If $E(\theta, \zeta)$ is probabilistic then the marginalization of E with respect to θ is

$$E_{\hat{\theta}}(\zeta) = -\log\left(\int e^{-E(\theta,\zeta)} d\theta\right). \tag{4.6}$$

When marginalization can be achieved analytically it provides a useful way to view minimization of E on a smaller parameter space, though minimization of the marginalized energy is not in general equivalent to minimization of the full energy. Another expression for this process is *marginalizing out* θ.

Conditioning of a probability distribution on a specific subset of variables can be stated in the energy framework using marginalization. To make an energy $E(\theta, \zeta)$ probabilistic in a subset of parameters ζ one need only shift by the marginal energy

$$E_{\bar{\theta}}(\theta, \zeta) = E(\theta, \zeta) - E_{\hat{\zeta}}(\theta). \tag{4.7}$$

That is, for every fixed value of θ it is trivial to show that this energy is probabilistically normalized with respect to ζ.

The final operation that connects the general energy framework to the probabilistic framework we will use is that of sampling from a distribution. Given a probability distribution $P(\theta)$ a trivial operation to perform is sampling θ from this distribution. If $E(\theta)$ is probabilistic then the notation

$$\theta \sim E \tag{4.8}$$

will denote sampling θ from the probability distribution $e^{-E(\theta)}$. When E depends on other arguments than θ, the notation

$$\theta \sim_\theta E(\theta, \zeta) \tag{4.9}$$

will be used to represent sampling θ in place of the more verbose

$$\theta \sim_\theta E_{\bar{\zeta}}(\theta, \zeta). \tag{4.10}$$

The ability to carry out this operation on pieces of the energy is frequently important for probability inference algorithms to construct an algorithm which migrates to the minimum of E, or equivalently to locations where the equivalent probability distribution is large. After

the next two sections summarizing NMF and LDA, section 4.5 outlines the relationship between LDA and NMF using the negative log-likelihood and marginalization.

4.3 Non-Negative Matrix Factorization

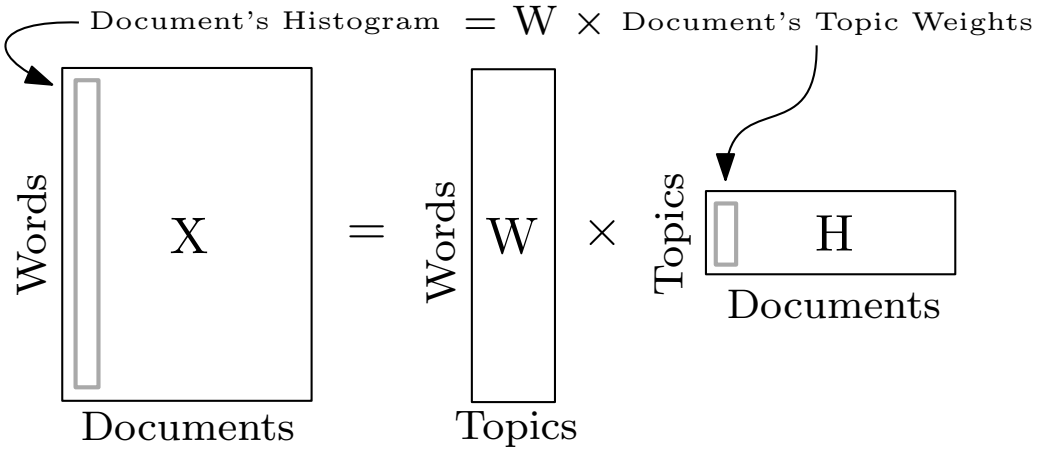

Figure 4.1: Non-negative matrix factorization diagram. Each document's histogram is modeled as a linear combination of the topic vectors which form the columns of **W**. In NMF all entries of these matrices are non-negative, each document's histogram is modeled by a strictly additive combination of these columns.

The concept of NMF was introduced in [Paa97] as positive matrix factorization then later became widespread with the publication of [LS99] that includes an easily implemented algorithm based on alternating multiplicative updates [LS01]. In essence, NMF is a technique similar to principal component analysis or singular value decomposition, but with an entry-wise non-negativity condition on the computed matrices. This results in something known as a *parts-based decomposition* with data factors that have large positive values for collections of features that tend to co-occur in the data.

We describe the NMF model, with optional penalty terms, as follows. Denote by n the number of documents in the corpus and m the size of the dictionary containing all words that appear in all documents. The document collection is parsed into an m-by-n data matrix $\mathbf{X}$

such that $\mathbf{X}_{i,j}$ is the number of times the i^{th} word appears in the j^{th} document. The columns of $\mathbf{X}$ are thus document word-histograms that discard word order. NMF approximates this histogram matrix using a product of two non-negative matrices $\mathbf{X} \approx \mathbf{WH}$ with an inner product dimension $t \ll \min(m, n)$ where t is the number of topics. The number of topics is simply a parameter to the model in the same way that an unsupervised clustering algorithm may require the number of clusters as a parameter.

The extracted information produced by NMF is the matrices $\mathbf{W}$ and $\mathbf{H}$. These are found using an optimization framework by solving the problem

$$\arg \min_{\mathbf{W},\mathbf{H} \succeq \mathbf{0}} D(\mathbf{X}, \mathbf{WH}) + r_{\mathbf{W}}(\mathbf{W}) + r_{\mathbf{H}}(\mathbf{H}) \tag{4.11}$$

with D a measure of divergence, or disagreement, between the data and the model. The optimization problem takes places over all possible matrices $\mathbf{W}$ and $\mathbf{H}$ with non-negative entries. Two common choices for D are the squared Frobenius matrix norm

$$D_F(\mathbf{X}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{Y}||_F^2 = \sum_{i,j} (\mathbf{X}_{i,j} - \mathbf{Y}_{i,j})^2 \tag{4.12}$$

or generalized Kullback-Leibler divergence

$$D_{KL}(\mathbf{X}, \mathbf{Y}) = \sum_{i,j} \mathbf{X}_{i,j} \log \left(\frac{\mathbf{X}_{i,j}}{\mathbf{Y}_{i,j}} \right) - \mathbf{X}_{i,j} + \mathbf{Y}_{i,j}. \tag{4.13}$$

The selection of norm depends on assumptions about the data matrix $\mathbf{X}$. In the literature for NMF (4.13) is sometimes used due to the simple multiplicative algorithm it enables as presented by [LS99], though later literature for NMF with fast optimization techniques typically use divergence (4.12) due to the simple analytic form of the gradient. The regularity terms, or penalties, $r_{\mathbf{W}}$ and $r_{\mathbf{H}}$ in (4.11) can capture additional assumptions placed on the problem such as sparsity of the matrices or columns norm constraints.

Given the definition of $\mathbf{X}$ chosen with documents occupying columns, the optimal matrix $\mathbf{W}$ has columns that are each a distribution of weight over words in the dictionary. The largest values within each column are the most significant words contributing to each topic. The top words for each topic, for example as shown in table 4.1 , are simply the words

with the largest value for each column of $\mathbf{W}$. Each column of $\mathbf{H}$ contains t weights used to combine the topic columns of $\mathbf{W}$ to approximate one document's histogram vector. For example, a column of $\mathbf{H}$ pertaining to a document about agriculture will likely have large a large coefficient associated with an agriculture column of $\mathbf{W}$. A topic will only be captured by the model if it is frequent enough throughout the corpus relative to the magnitude of t. Selecting smaller values for t will tend to find a few general trends in the data while larger t produces more detailed topics. Too large of a value for t will begin to over-fit the data and no longer produce meaningful topics. Unfortunately, the choice of this parameter for a general data set depends on a number of factors. In general, a heuristic such as one topic per 1000 documents may perform well, but the complexity of the data, the length of the documents, and the finer details of constructing $\mathbf{X}$ all play a role in the quality of results.

Model variations on NMF have been designed for a variety of more general data analysis techniques. In addition to simple penalties such as sparsity or norm constraints manifesting through $r_{\mathbf{W}}$ and $r_{\mathbf{H}}$, researchers have investigated regularizing the columns of these matrices using graph regularity terms [CHH11]. Generalizations of NMF include replacing the data matrix $\mathbf{X}$ with a higher-dimensional non-negative tensor [SH05] resulting in a model built from the sum of t rank-1 tensors rather than a matrix product. NMF can be used in interesting ways to form new data science techniques, for example by applying NMF to a graph similarity matrix to cluster data [KDP12]. The work of [WZ13] may be consulted for a general survey.

A variety of algorithms have been proposed for solving this problem in addition to the popular multiplicative method of [LS99]. One such method involves alternating unconstrained minimizing of equation (4.11) while projecting back to the non-negative orthant. Unfortunately, this only approximately minimizes (4.11). Alternating non-negative least squares is a superior approach [KP07] that achieves a local minimizer and can be made computationally fast [KP11]. Other techniques utilize projected gradient descent [Lin07] or higher-order approximate Newton steps [KSD07], projected Newton methods [GZ12], splitting with the alternating direction method of multipliers [SF14], primal-dual

optimization [YB14], or block-coordinate optimization with Taylor approximations to the divergence [LLP12]. The separability assumption [DS03, AGK12] is applicable in many settings. This assumes that the rows of $\mathbf{X}$ contain the desired rows of the solution matrix $\mathbf{H}$. When this assumption holds, a multitude of very effective techniques can be used [GV14, KSK12, EMO12, RRT12]. Separability is a reasonable assumption in the case of topic modeling [GV14] where the problem becomes a search for "token words" that are associated uniquely with a topic. NMF is closely related to many other models that share a familiar latent matrix factorization structure [SG08] for example SVD, PCA, and spectral clustering [DHS05].

This work is focused on the application of NMF specifically to text documents for topic modeling. Using NMF for text document classification was first done in [XLG03] where it was shown to outperform many methods at the time. To accomplish this, [XLG03] argues that the input matrix to the problem should be pre-processed using a method known as term-frequency-inverse-document-frequency (TF-IDF) that scales down rows of $\mathbf{X}$ corresponding to common words while increasing the scale, and hence importance, of rows with unique words [SB88]. To this day, the re-normalization of the data matrix using TF-IDF is a common practice in particular when using the Frobenius divergence due to the tendency of (4.12) to place importance on extreme values. In results shown in section 4.9, NMF results are presented using the Frobenius divergence (4.12) both with and without TF-IDF re-weighting where it becomes clear that at times it is a necessary preprocessing step. The mathematics of TF-IDF are quite simple. First, let c_i be the number of documents that contain the word i. This quantity is used to indicate the importance of a word with infrequent words considered topically important and common words unimportant. While there are a variety of ways to state TF-IDF using different re-weighting schemes, a simple and effective scheme is

$$\tilde{\mathbf{X}}_{i,j} = \mathbf{X}_{i,j} \log\left(\frac{n}{c_i}\right) \tag{4.14}$$

where $\mathbf{X}$ is the re-weighted data.

4.4 Latent Dirichlet Allocation

LDA [BNJ03] is a Bayesian model that builds off a prior model known as probabilistic latent semantic indexing (pLSI) [Hof99]. pLSI posited that the co-occurrences of words in documents can be modeled as a sum over contributions from some fixed number of latent distributions. That is,

$$P(w, d) = \sum_{z=1}^{t} P(z|d)P(w|z) \tag{4.15}$$

is the probability of observing a word w in a document d given t latent topics. The index z is a sum over contributions from each topic. $P(z|d)$ is the probability of sampling from topic z for document d, and $P(w|z)$ is the probability that the word of type w is generated by topic z. The probability (4.15) is therefore the probability of finding a word of type w in the document d given a process whereby first a topic is sampled for the document, then a word is sampled for that topic. This relationship between data and latent distributions is used to infer the distributions $P(z|d)$ and $P(w|z)$ given the data consisting of observed instances of words in documents. pLSI unfortunately tends to over-fit data and arrive at sub-optimal solutions [BNJ03]. This is suggested by observing that the probability degenerates when zeros are introduced into the learned distributions. LDA adds priors, or penalties, to the distributions $P(z|d)$ and $P(w|z)$ that ensure zeros do not appear in the distributions. This slight change results in significantly better performance of the model in practice for many different algorithms.

LDA is a generative model for words appearing in documents. The model attempts to assign to each word appearance in each document the topic that generated that word. Let k be the number total number of words in all documents, henceforth known as *instances*. Each instance has corresponding values $w_i \in \{1, 2, ..., m\}$ denoting which word in the length-m dictionary the instance i is of, as well as $d_i \in \{1, 2, ..., n\}$ that denotes which of the n documents instance i occurred in. The values of w_i and d_i are given by the data set initially. LDA seeks to find values for new parameters $x_i \in \{1, 2, ..., t\}$ that indicate which extracted topic instance i was generated by. For example, the word "plant" appearing in a document

about gardening may have a value x_i equal to the index of a learned gardening topic, but "plant" appearing in a document about power plant designs may have a value x_i selecting a topic about power plants.

In addition to these x_i, the LDA model also seeks to find two matrices $\mathbf{W}$ and $\mathbf{H}$ which are non-negative, have columns with a unit-sum constraint, and which are respectively of dimensions m-by-t and t-by-n. The columns of these matrices represent probability distributions over, respectively, the dictionary for each topic and the topics for each document. Using the notation $\propto$ to represent proportionality up to a constant not dependent on the quantities to be found, the posterior model probability for LDA can be stated

$$P_{\text{LDA}}(\mathbf{W}, \mathbf{H}, x_1, x_2, ...) \propto P_D(\mathbf{W}, \alpha) P_D(\mathbf{H}, \beta) \prod_{i=1}^{k} \mathbf{W}_{w_i, x_i} \mathbf{H}_{x_i, d_i} \tag{4.16}$$

where the product takes place over all instances. The product expression on the right-hand side of (4.16) captures the probability, given the model, of observing the instances in the data. Specifically, $\mathbf{W}$ captures $P(w|z)$ in pLSI while $\mathbf{H}$ captures $P(z|d)$. The terms applying P_D to the two matrices are the aforementioned penalties that prevent zeros from appearing in the solution. The expression for these terms, with $\alpha, \beta > 0$ taken to be fixed quantities known as *hyperparameters*, is the probability that columns of the matrices are sampled from a Dirichlet distribution,

$$P_D(\mathbf{M}, \mu) = C(\mu) \prod_{i,j} \mathbf{M}_{i,j}^{\mu - 1} \tag{4.17}$$

with $C(\mu)$ a normalization constant depending only on μ. One can see from (4.17) that if any entry of the argument matrix $\mathbf{M}$ is zero and if $\mu > 1$, any entry being equal to zero results in $P_D(\mathbf{M}, \mu) = 0$. Therefore the only solutions corresponding to non-zero probability in this case must be fully dense. This penalty term serves as a means to prevent either the distributions in the columns of $\mathbf{W}$ or $\mathbf{H}$ having entries that are too small. The Dirichlet distribution directly addresses the degeneracy issue faced by pLSI. The statement of LDA (4.16) suggestively describes the model distributions inferred by LDA as columns of two matrices $\mathbf{W}$ and $\mathbf{H}$ with non-negative unit-sum columns. This notation is intentionally chosen as these matrices capture information identifiable with the information captured by

the two like-named matrices learned by NMF in section 4.3. In the literature for LDA, algorithms to find solutions typically explore the posterior distribution (4.16) to understand probable values for $\mathbf{W}$, $\mathbf{H}$, and the instance-topic assignments x_i. This is in contrast with NMF where only a single solution for the decomposition matrices is found. Additionally, NMF does not deal with anything analogous to the x_i found by LDA.

LDA, similar to NMF, forms the foundation of a number of methods in the statistics literature. For example, to address the topic number selection issue [GT04] develop the hierarchical Dirichlet process to select the number of topics using a topic tree structure. Improving basic assumptions, manifesting through the priors like P_D, can result in more meaningful topic information. The fundamental Dirichlet prior $P_D(\mathbf{H})$ in equation (4.16) was shown by [WMM09] to be too restrictive, and that relaxing it with an asymmetric Dirichlet prior can improve topic model performance. Similarly, the correlated topic model of [BL07] replaces the Dirichlet prior with something entirely different, the logit-normal distribution, that addresses a tendency for documents to have a single contributing topic. The "focused topic model" of [WWH09] is a variation on the hierarchical LDA model with a modified penalty on $\mathbf{H}$ that enforces sparsity and hence more of a focus on fewer topics. More recently, [ZHD12] generalized priors in ways which connect pLSI, LDA, and the focused topic model into a versatile and general count-data factor model. In the case of documents which have additional information such as a date or author along with the text, joint models built on the LDA framework include supervised LDA [MB08] which models this "metadata" as a generative result of the learned topics and the subsequent work of [MM12] in which Dirichlet-multinomial regression is used, differing by a conditioning on the metadata for the generative topic process. Both of these methods are related to other models which incorporate additional information such as the topics-over-time model [BL06a] and the author-topic model [RGS04].

Literature for LDA and related models largely utilize algorithms such as variational Bayes (VB) [BNJ03], expectation-maximization (EM) [DLR77, VP14, ZLC16], and Markov chain Monte Carlo (MCMC) methods such as Metropolis-Hastings [MRR53] or a specific case which we consider known as Gibbs sampling (GS) [CG92]. Variational Bayes [JGJ99], the

approach proposed in the original LDA work of [BNJ03], assumes a functional form for (4.16) and minimizes the difference of this approximation to the true distribution. Expectation-maximization [DLR77] is a way to find a MAP estimate for the model after marginalizing out x_i from (4.16), a process explained further in sections 4.5 and 4.6.1. The Gibbs sampler [CG92] explores (4.16) by forming a sequence of parameters which, after a sufficiently long time, approximately are sampled from (4.16) and therefore capture both good solutions as well as uncertainly in the solution. This algorithm is studied in section 4.6.2 in detail. More recent and efficient algorithms for LDA are stochastic variational inference [HBW13] which uses stochastic optimization to find approximations to (4.16) for large data sets, and work such as that of [YGH15] seeking to make the MCMC approach to LDA fast using parallel computation. In the next sections we outline the model-level connections between the probability for LDA (4.16) and the energy for NMF (4.11) as it is known in the literature.

4.5 Analytic Comparisons

The LDA and NMF models are known to be similar latent factor models. The pLSI model that forms the basis for LDA has, up to column normalization, a negative log-likelihood equivalent to NMF using the KL-divergence (4.13) [DLP08]. LDA builds on pLSI by adding Dirichlet priors [BNJ03] to the matrices, hence it stands to reason that NMF with KL-divergence is closely related to LDA. In this section the relationship is summarized using the energy framework.

Both models start with a corpus of m dictionary elements, n documents, and attempt to find some number t of latent topics. We denote by $\mathbf{X}$ the data histogram matrix with entry (i, j) equal to the number of times word i appears in document j. As was the case for LDA, let w_i, d_i, and x_i denote the integer word, document, and topics for each instance of a word in a document with i ranging $1, ..., k$ for k the total number of instances. While $\mathbf{X}$, w_i, and d_i are computed directly using the data the values of x_i are to be found by the model.

Recall the LDA probability (4.16)

$$P_{\text{LDA}}(\mathbf{W}, \mathbf{H}, x_1, x_2, ...) = P_D(\mathbf{W}, \alpha) P_D(\mathbf{H}, \beta) \prod_i \mathbf{W}_{w_i, x_i} \mathbf{H}_{x_i, d_i}. \tag{4.18}$$

Taking the negative logarithm of (4.18) produces up to additive constant LDA with non-probabilistic penalty terms [VP14]

$$E_{\text{LDA}}(\mathbf{W}, \mathbf{H}, x_1, x_2, ...) \propto -\sum_i \log(\mathbf{W}_{w_i, x_i} \mathbf{H}_{x_i, d_i}) + (1-\alpha)\log(\mathbf{W}) + (1-\beta)\log(\mathbf{H}) \tag{4.19}$$

with the logarithm of a matrix $\log(\mathbf{W})$ in (4.19) denoting the sum of the logarithm of all entries

$$\log(\mathbf{W}) = \sum_{i,j} \log \mathbf{W}_{i,j}.$$

Marginalizing (4.19) over instance-topic assignments x_i produces an energy identifiable with NMF using KL divergence

$$E_{\text{LDA}, \hat{x_i}}(\mathbf{W}, \mathbf{H}) = D_{KL}(\mathbf{X}, \mathbf{WH}) + (1-\alpha)\log(\mathbf{W}) + (1-\beta)\log(\mathbf{H}). \tag{4.20}$$

The marginalization operation transforms a dependence on individual word labels to a dependence only on the word count matrix $\mathbf{X}$. Though NMF can be formulated with the single-word assignment parameters x_i as well [FC09] it is almost exclusively stated in the post-marginalization form. The only remaining distinction between the above marginalized LDA and NMF energies, other than particular choice of divergence and penalty terms, is the column-normalization of the matrices $\mathbf{W}$ and $\mathbf{H}$. Incorporating this constraint into NMF concludes the model-level similarities. The underlying mechanisms of LDA as a matrix factorization technique have not gone unnoticed. LDA can also be understood as a Bayesian model for Poisson matrix factorization due to the KL-divergence corresponding to a Poisson noise assumption [ZHD12]. Discovering relationships such as these naturally lead to the extension of progress with NMF to the LDA domain, for example by extending non-negative tensor factorization to Bayesian Poisson tensor factorization [SPB15]. NMF with Frobenius divergence can also be referred to as Gaussian matrix factorization as the norm follows from the negative logarithm of a Gaussian noise assumption.

Many algorithms for LDA such as the popular collapsed Gibbs sampler depend on the per-word assignment parameters x_i. Indeed, the collapsed Gibbs sampler uses this representation exclusively with the parameters $\mathbf{W}$ and $\mathbf{H}$ being marginalized over. For these reasons, relating the models as summarized above is not sufficient – the algorithms applied remain largely disjoint, and the non-convex and indeed NP-hard nature of the problem [AGM12] means different algorithms may impact the quality of solutions. In the interest of further understanding x_i and how various algorithms compare, the remainder of this section presents a broad perspective from which different inference techniques for both NMF and LDA can be characterized.

In the general setting of a topic model there are k observations $y_i \in Y$, each to which the model assigns a label $x_i \in X$. Here Y and X represent, respectively, the type of data which may be observed and the labels which are to be assigned to the observations. With the topic model Y is the set of integer pairs

$$Y = \{(w, d) : w \in [1, ..., m], d \in [1, ..., n]\} \tag{4.21}$$

indicating word type and document index for instances, while X is the set of integer topic assignments $\{1, ..., t\}$. Additionally, a collection of model selection parameters $\theta \in \Theta$ are sought that describe the data-label relationship. For topic models these parameters are the matrices, or distributions, $\mathbf{W}$ and $\mathbf{H}$. A model is a relationship between the data and the parameters to be learned. This is described through some posterior probability distribution

$$P(\theta, x_1, ..., x_N) \tag{4.22}$$

that describes the likelihood of model and assignment parameters. In this work the aim is limited to finding the MAP estimate in the probabilistic framework, or equivalently the energy minimum in the energy framework. This form (4.22) is too general for topic modeling, however, as it may depend on the ordering of observations x_i. Assuming the observations are exchangeable e.g. the bag-of-words assumption, the probability (4.22) can be written only as depending on the number of observations of each type Y. This is captured by the empirical data histogram. The empirical histogram is simply a function $p(x, y) \to \mathbb{R}$ for $x \in X$ and

$y \in Y$ where $p(x, y)$ is proportional to the number of times an observation of type y is given assignment x. For topic models, this corresponds to the number of times a certain word in a certain document is assigned to a topic. Define the empirical data histogram

$$p(x, y) = \frac{1}{N} \sum_i 1_{x=x_i} 1_{y=y_i}. \tag{4.23}$$

This distribution can be factored into a product

$$p(x, y) = f(x, y)p(y)$$

where the probability of each observation is

$$p(y) = \sum_X p(x, y)$$

and therefore for y that have been observed at least once

$$f(x, y) = p(x, y)/p(y). \tag{4.24}$$

$f(x, y)$ is, for observed y, a distribution over class assignments x. For unobserved y the value of $f(x, y)$ has no impact on the model. Let $\mathcal{F}$ denote all functions $g(x, y)$ such that for each $y \in Y$ the function $g(\cdot, y)$ is a probability distribution on X. Hence $f(x, y)$ is, for any choice of assignments x_i, a member of $\mathcal{F}$. Denote by $\hat{\mathcal{F}} \subset \mathcal{F}$ the subset of such functions which correspond to $f(x, y)$ in (4.24) for some choice of assignments x_i. That is, $\hat{\mathcal{F}}$ is the empirical set of choices for f that can result from the data through the empirical histogram (4.23). For example, if there is only a single observation of type y the distribution $f(x, y)$ must be an indicator on a lone assignment x and cannot be a general distribution. In conclusion an exchangeable model probability, and therefore model energy, can be written in terms of the choice of assignment function $f \in \hat{\mathcal{F}}$ and model parameters $\theta \in \Theta$ alone

$$E(\theta, f) = -\log(P(\theta, f)).$$

This statement of the general topic model enables the study of inference techniques for both NMF and LDA in the energy framework subsequently in section 4.6.

4.6 Inference Techniques

4.6.1 Expectation-Maximization

Expectation-maximization (EM) [DLR77] is a technique for finding a MAP estimate for a marginal probability. The parameter space is separated into one group that is marginalized over and another that the probability is maximized with respect to. In the case of $E(\theta, f)$ for $\theta \in \Theta$ and $f \in \hat{\mathcal{F}}$ both the iteration

$$\theta \leftarrow \arg\min_{\phi} \int_{\mathcal{F}} E(\phi, f) e^{-E_{\bar{\theta}}(\theta, f)} df \qquad \text{(EM-}\theta\text{)} \tag{4.25}$$

and

$$f \leftarrow \arg\min_{g} \int_{\Theta} E(\theta, g) e^{-E_{\bar{f}}(\theta, f)} d\theta \qquad \text{(EM-f)} \tag{4.26}$$

correspond to EM methods. The first maximizes the probability, hence minimizes the energy, with respect to θ while marginalizing over f. The second iteration does exactly the reverse. Taking the EM-θ iteration (4.26) and re-arranging terms clarifies this behaviour slightly

$$\theta \leftarrow \arg\min_{\phi} \int (E_{\hat{f}}(\phi) + E_{\bar{\theta}}(\phi, f)) e^{-E_{\bar{f}}(\theta, f)} df \tag{4.27}$$

$$\Rightarrow \quad \theta \leftarrow \arg\min_{\phi} E_{\hat{f}}(\phi) + \int \left(E_{\bar{f}}(\phi, f) - E_{\bar{\theta}}(\theta, f) \right) e^{-E_{\bar{\theta}}(\theta, f)} df \tag{4.28}$$

$$\Rightarrow \quad \theta \leftarrow \arg\min_{\phi} E_{\hat{f}}(\phi) + I_{KL}(\theta, \phi). \tag{4.29}$$

This is known as the a Kullback-proximal point iteration [CH00] that converges to a local minimum of the marginal energy $E_{\hat{f}}$ [XJ96, Wu83, HF95]. In particular, the quantity I_{KL} becomes insignificant with more iterations. The EM-f algorithm similarly converges to a local minimum of $E_{\hat{\theta}}$. To further simplify these iterations, assume in either (4.25) or (4.26) that the exponential in the integral is perfectly concentrated about a single point. This occurs when the energy has a clear, single minimal point in the integrated direction. In both cases the iteration reduces to

$$f \leftarrow \arg\min_{g} E(\theta, g)$$
$$\theta \leftarrow \arg\min_{\phi} E(\phi, f)$$

which corresponds to alternating coordinate descent of the energy [Ama95].

4.6.2 Collapsed Gibbs Sampling

The Gibbs sampler (GS) [CG92] iteratively cycles through parameters on an individual basis. Similar to coordinate descent, the algorithm iterates over all individual assignments sampling repeatedly

$$x_i \sim_{x_i} E(x_1, ..., x_N) \tag{4.30}$$

for $i = 1, 2, ..., N$ until stability around the minimum of E is reached.

The Gibbs sampler differs from EM in that it does not converge to a single minimizing value of E corresponding to a MAP estimate, but rather in the limit of many iterations becomes a random walk near minima of E. This for example enables analysis of uncertainly in the solution and the study of potentially multiple modes. The stochastic nature can also be advantageous when E is challenging and the perturbations due to sampling are able to navigate out of local optima. In the context of f this manifests as a probabilistic update to f where the contribution of a single variable x_i is modified probabilistically based on the probability distribution P with negative log-likelihood E. To make this concrete, the update for f in a single iteration first selects a new assignment for an observation i with the probability of x_i becoming a new value $x_i \in X$ given by

$$P(x_i = x') \propto P\left(f(x, y) + \frac{(1_{x=x'} - 1_{x=x_i})1_{y=y_i}}{Np(y_i)}\right). \tag{4.31}$$

With the new assignment selected, f is updated as such

$$f(x, y) \leftarrow f(x, y) + \frac{(1_{x=x'} - 1_{x=x_i})1_{y=y_i}}{Np(y_i)}. \tag{4.32}$$

These steps are iterated for all observations i until the algorithm stabilizes around high probability states for f. In order to understand the geometric relationship of the Gibbs sampler to other algorithms used by topic models, the remainder of this section will investigate the continuum limits of the Gibbs sampler with single-datum assignment extended to re-sampling of multiple observations simultaneously known as the *block* Gibbs sampler. Consider re-sampling k_i observations of type $z_i \in Y$ for $i = 1, ..., M$, for some M, at once.

Let $\epsilon_i = k_i/(Np(z_i))$ be the fraction of data re-assigned for each type of observation under consideration z_i. In this case the steps become as follows

$$q \sim \mathcal{F}_{k_i,z_i} \quad \text{subject to} \quad f \geq \epsilon_i q \tag{4.33}$$

$$P(g \in \mathcal{F}_{k_i,z_i}) \propto P\left(f + \epsilon_i(g-q)\right) \tag{4.34}$$

$$f \leftarrow f + \epsilon_i(g-q) \tag{4.35}$$

where

$$\mathcal{F}_{k_i,z_i} = \left\{h(x,y) \text{ s.t. } h(\cdot,z_i) \in D_{k_i} \ \ \forall i = 1,...,M\right\},$$
$$D_n = \left\{\frac{1}{n}\sum_{j=1}^{n} 1_{v_j} \text{ s.t. } v_1,...,v_n \in X\right\}.$$

The set $\mathcal{F}_{k_i,z_i}$ is a set of functions similar to those of $\hat{\mathcal{F}}$ but restricted to represent the possible contribution of the selected observations. D_n is the set of distributions on X that represent empirical histograms of n quanta. What has changed is the parameters that are sampled in one step. In (4.33) a subset of observations for the desired types Y are selected at random by finding a function q that is similar to f but which represents the contributions to f of only the subset of observations being sampled. A new contribution for these observations to f is found with probability given by (4.34) using the topic model probability, and finally this is used to update f.

Consider now the continuum limit $\epsilon_i \to \epsilon/p(z_i) > 0$ and $N \to \infty$ with $p(y)$ is held constant. $D_n(X)$ becomes general distributions on X while $\mathcal{F}_{k_i,z_i}$ becomes $\mathcal{F}$. In this limit as well, because the fraction of each observation type is uniform and constant, a fraction of f is updated at each step. Hence the block Gibbs sampler under this limit becomes first a matter of selecting an appropriate re-assignment for the ϵ fraction of the data

$$P(g \in \mathcal{F}) \propto P\left((1-\epsilon)f + \epsilon g\right) \tag{4.36}$$

then updating the assignments via $f \leftarrow (1-\epsilon)f + \epsilon g$. These two steps represent a continuum limit of the block Gibbs sampler when re-assigning a fraction ϵ of the data uniformly and simultaneously. The state f stochastically migrates through the solution space at a rate

given by the re-assignment fraction ϵ. To finally connect this with the other algorithms under consideration that seek only MAP estimates, suppose the sampler takes only the most probable path in (4.36) thus making the process a deterministic update

$$f \leftarrow f + \epsilon(\mathcal{M}_\epsilon(f) - f) \tag{4.37}$$

where

$$\mathcal{M}_\epsilon(f) = \arg\max_{g \in \mathcal{F}} P((1-\epsilon)f + \epsilon g).$$

In the limit of small ϵ (4.37) is approximately local gradient ascent of P, each iteration being a step to the optimal solution in a neighbourhood. In the case $\epsilon = 1$ solving the problem is moved completely into the re-assignment selection step

$$f \leftarrow \arg\max_{g \in \mathcal{F}} P(g).$$

Thus far the analysis for the Gibbs sampler has ignored θ that captures $\mathbf{W}$ and $\mathbf{H}$ in the topic modeling problem. For LDA these parameters ave simply been marginalized over to bring the model probability completely to a dependence on x_i and, thus, f. The final algorithm for a maximum likelihood interpretation of the continuum collapsed Gibbs sampler is given by

$$\begin{aligned} f &\leftarrow (1-\epsilon)f + \epsilon \left(\arg\min_{g \in \mathcal{F}} E_{\hat{\theta}}((1-\epsilon)f + \epsilon g) \right) \\ &\Rightarrow f \leftarrow \arg\min_{g \in \mathcal{F}} \{E_{\hat{\theta}}(g) + I_\epsilon(g, f)\} \end{aligned} \tag{4.38}$$

with a suggestive use of the distance

$$I_\epsilon(g, f) = \begin{cases} 0 & \text{if } \frac{1}{\epsilon}(g - (1-\epsilon)f) \in \mathcal{F} \\ \infty & \text{otherwise.} \end{cases}$$

Coordinate stepping becomes gradient descent due to the nature of the topic modeling problem, the exchangeability assumption, and because large N makes individual observation assignments result in infinitesimal changes to f. Recall for the EM-f algorithm in (4.26) the iteration

$$f \leftarrow \arg\min_{g \in \mathcal{F}} \{E_{\hat{\theta}}(g) + I_{KL}(g, f)\}$$

for I_{KL} a proximal penalty described in [CH00]. The form (4.38) of the Gibbs sampler can be interpreted as a proximal mapping analogous to the proximal iteration of the EM algorithm. We have therefore shown that when applying the collapsed Gibbs sampler to data adhering to the above limits the result is a proximal descent similar to the proximal descent of EM-f, and in a continuum data limit this is approximately a gradient descent of the marginal energy. In this sense the EM-f algorithm can be seen as a non-stochastic variation of random sampling with a much larger proximal step but, as a consequence, a much higher cost in per-step calculation. We expect therefore that in a continuum data limit the Gibbs sampler is more likely to become stuck in a local minimum that EM iteration may avoid. Additionally, we speculate that away from this limit the stochastic properties of Gibbs sampler will allow a more effective exploration of the solution space than EM.

Taking the energy (4.19) and after performing standard but lengthy calculations, the marginal energy can be explicitly written

$$E_{\hat{\theta}}(x_i) \propto \sum_k \lg(Z_k + m\alpha) - \sum_{k,j} \lg(N_{k,j} + \beta) - \sum_{i,k} \lg(M_{i,k} + \alpha) \tag{4.39}$$

with Z_k the number of words assigned to the k^{th} topic, $N_{k,j}$ the number of words in document j assigned to topic k, and $M_{i,k}$ the number of words of dictionary type i assigned to topic k. The function lg is the natural logarithm of the gamma function. N and M are discrete matrices which can be thought of similarly to $\mathbf{H}$ and $\mathbf{W}$ respectively, though they are stored implicitly through word assignments x_i. The first term of (4.39) favors classes of equal size, while the second and third terms seek count matrices N and M having larger values where large values already exist. Although the original factorization matrices $\mathbf{W}$ and $\mathbf{H}$ are not explicitly represented, this marginal energy still desires selecting x_i that are as consistent as possible with representation as two factor matrices. We also note that the non-zero parameters α and β ensure that (4.39) remains finite. The continuum limit is therefore a limit in which the integer-valued matrices M and N behave effectively as real-values matrices, and in this limit the stochastic nature of the Gibbs sampler approaches gradient descent. These count matrices can be averaged for a number of iterations to approximate

the distributions captured by $\mathbf{W}$ and $\mathbf{H}$ for comparison with other techniques.

4.6.3 Alternating Minimization

Algorithms in the NMF literature typically seek simply to minimize $E_{\hat{f}}$ and therefore operate in the space $\theta = (\mathbf{W}, \mathbf{H})$. Alternating minimization (AM) is one such algorithm that accomplishes this through alternating optimization with respect to $\mathbf{W}$ and $\mathbf{H}$. This consists of performing the two steps in repetition

$$\mathbf{W} \leftarrow \arg\min_{\mathbf{W} \geq 0} E_{\hat{f}}(\mathbf{W}, \mathbf{H})$$

$$\mathbf{H} \leftarrow \arg\min_{\mathbf{H} \geq 0} E_{\hat{f}}(\mathbf{W}, \mathbf{H})$$

until convergence. This is particularly useful in the case of Frobenius divergence as this reduces to alternating non-negative least squares for which there exist fast algorithms [KP11]. Similarly to EM, AM minimizes a marginal form of the energy. In contrast, however, it explicitly operates on $\mathbf{W}$ and $\mathbf{H}$ separately while EM seeks a new state $(\mathbf{W}, \mathbf{H})$ that is in KL-proximity to the previous state.

4.7 Tensor Comparison

In this section we review the three algorithms studied as energy minimization problems in a tensor space. There are m words in the dictionary, n documents, and the topic models seek to reduce the observations to t latent topics. The space Y can be identified with the set of all pairs of integers (i, j) for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$, each pair indicating the presence of a particular word i in a particular document j. The topic labels $x_i \in X = \{1, 2, ..., t\}$ are given by integer assignments to be discovered assigning each instance to a topic. Let $\mathcal{T}$ be the set of all m-by-n-by-t tensors $\mathbf{Z}_{i,j,k}$ with respective dimensions indexed by i, j, and k. Define

$$\mathcal{F} = \left\{ \mathbf{Z} \in \mathcal{T} : \mathbf{Z}_{i,j,k} \geq 0, \ \sum_k \mathbf{Z}_{i,j,k} = 1 \ \ \forall \ \ i, j \right\}$$

and

$$\Theta = \left\{(\mathbf{W}, \mathbf{H}) : \mathbf{W} \in \mathbb{R}_+^{m \times t}, \;\; \mathbf{H} \in \mathbb{R}_+^{t \times n}\right\}$$

as the realizations for topic modeling of previously discussed sets. $\mathcal{F}$ represents assignments of the observed data to topics and is the representation space for the collapsed Gibbs sampler of section 4.6.2. There is a natural map from Θ to $\mathcal{T}$ given by

$$F((\mathbf{W}, \mathbf{H}) \in \Theta)_{i,j,k} = \mathbf{W}_{i,k}\mathbf{H}_{k,j} \in \mathcal{T}$$

that produces $F(\Theta) \in \mathcal{T}$, the natural representation space for algorithms such as the alternating minimization of NMF of section 4.3. These two sets capture the geometry of latent factorization problem as well as provide insights into the computational challenges. $F(\Theta)$ is perfectly modelled low-rank data, while $\mathcal{F}$ represents assignments of observations to underlying mixture components. Choice of divergence provides an analytic measurement for distance between elements in each set. For $A \in \mathcal{F}$ and $B \in F(\Theta)$ the two divergences (4.12) and (4.13) considered in section 4.3 are given by

$$D_F(A, B) = \sum_{i,j} \left(\sum_k X_{i,j} A_{i,j,k} - B_{i,j,k} \right)^2 \tag{4.40}$$

$$D_{KL}(A, B) = -\sum_{i,j,k} X_{i,j} A_{i,j,k} \log(B_{i,j,k}). \tag{4.41}$$

This latent factorization problem can thus be viewed as a problem in the tensor space: given a measure of distance such as (4.40) and (4.41), an ideal solution is a pair of points in each of $\mathcal{F}$ and $F(\Theta)$ that are close in the divergence sense.

The EM algorithm alternates projection between these two sets with the variants EM-θ and EM-f holding explicit representations in $F(\Theta)$ or $\mathcal{F}$, respectively, while performing a proximal descent of the marginalized energy approaching the opposite set. This is schematized in figure 4.2 where the dashed EM path is shown along with two marginalized variants in grey. The difference between the two marginalized EM algorithms is which set is explicitly represented versus implicitly represented through the weighted averaging of equations (4.26) and (4.25). It is important to note that $F(\Theta)$ is non-convex and significantly smaller with

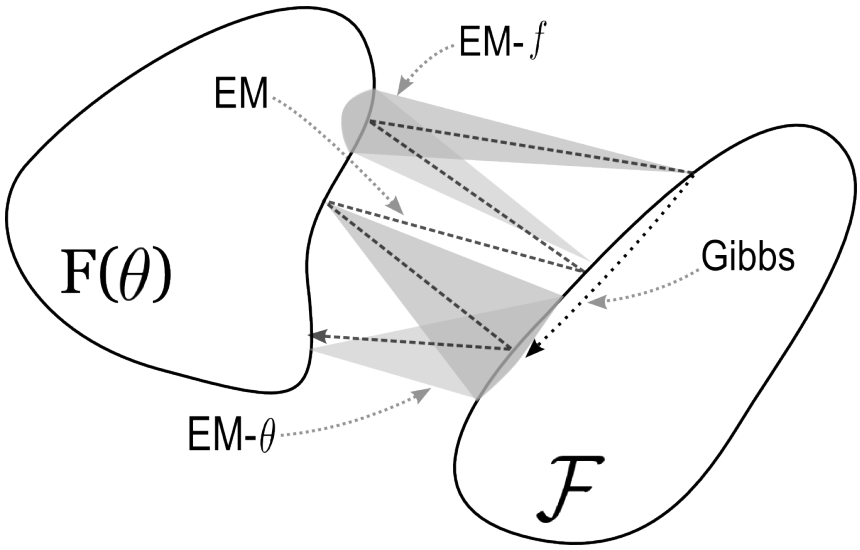

Figure 4.2: Visualization for the EM algorithm variants and the collapsed Gibbs sampler (in the small ϵ limit). The grey cones represent approximately those which points are contributing the the minimization of distance in the next iteration for the EM-f and EM-θ algorithms.

dimension $mt + nt$ than the convex counterpart $\mathcal{F}$ of dimension mnt, hence representations constrained to this space will be naturally more restricted than representations in $\mathcal{F}$. Indeed, the difficult nature of the topic modeling problem arises from the shape of $F(\Theta)$.

Alternating minimization for NMF is constrained to $F(\Theta)$ and navigates toward $\mathcal{F}$ by exploiting convexity of $F((\mathbf{W}, \mathbf{H}))$ when either $\mathbf{W}$ or $\mathbf{H}$ is held constant. The Gibbs sampler is an infinitesimal form of EM-f with small proximal descent steps, shown in figure 4.2 as the path through $\mathcal{F}$. The state of the Gibbs sampler is represented by assignments of single words to topics and amounts to a tensor in $\mathcal{F}$ restricted to having integer values. The marginal energy (4.39) that the Gibbs sampler minimizes causes the solution to navigate toward $F(\Theta)$ using the stochastic process (4.30) studied in section 4.6.2. With sparse data the stochastic nature of the sampler can be advantageous, while more dense data results in a limiting behavior of (4.37) producing a technique almost identical to gradient descent.

4.8 Topic Characterization

In some circumstances, text documents against which the topic model is applied may have additional information. For example, each document may have a corresponding location in space or time of authorship. Using such information in the context of topic models is interesting for two reasons. First, correlations between topics and additional information not included in the topic model provides a new and interesting way to understand the topics and data themselves. Secondly, high correlation also indicates performance of the topic model and validates assumptions when a practitioner is interpreting topic meaning.

Suppose there exists some value $z_j \in [0,1]^r$ for the j^{th}. The first metric proposed is the *mean squared distance* (MSD). Let A be the set of indices corresponding to documents belonging to a specific topic where topic assignments are found by taking the maximum entry in each column of $\mathbf{H}$. The MSD for the topics is given by precisely what the name implies,

$$\text{MSD} = \frac{1}{|A|^2} \sum_{i \in A} \sum_{j \in A} ||z_i - z_j||_2^2, \tag{4.42}$$

and is the expectation of the squared distance between two documents in the topic. Small values of the MSD imply that all documents in a topic are appearing approximately near a single point, while large values imply there is no such point.

The second metric proposed is the L^p norm (LP). The L^p norm of function on a domain Ω is defined by

$$||f||_p = \left(\int_\Omega |f(x)|^p dx \right)^{\frac{1}{p}}.$$

This is commonly utilized with $p = 2$ resulting in the Euclidean norm. Different values of p capture different information about the function f. For example, in the limit $p \to \infty$ it can be shown that, under some additional assumptions, this norm approaches the maximum value of f on Ω. In contrast, as $p \to 0$ this norm approximates the area on which f is non-zero [Rud91]. The proposed metric is computed by fixing p beforehand and applying the L^p norm to the histogram of z_j. This quantifies approximately how concentrated each topic is into a few areas of space versus spread out over a large area. In contrast to the MSD

metric, the LP metric will be small if z_j are concentrated at a few points but these points are far apart. Therefore these two metrics capture similar but certainly not identical properties of the topic's z_j distribution.

4.9 Numerical Comparisons

4.9.1 Synthetic Examples

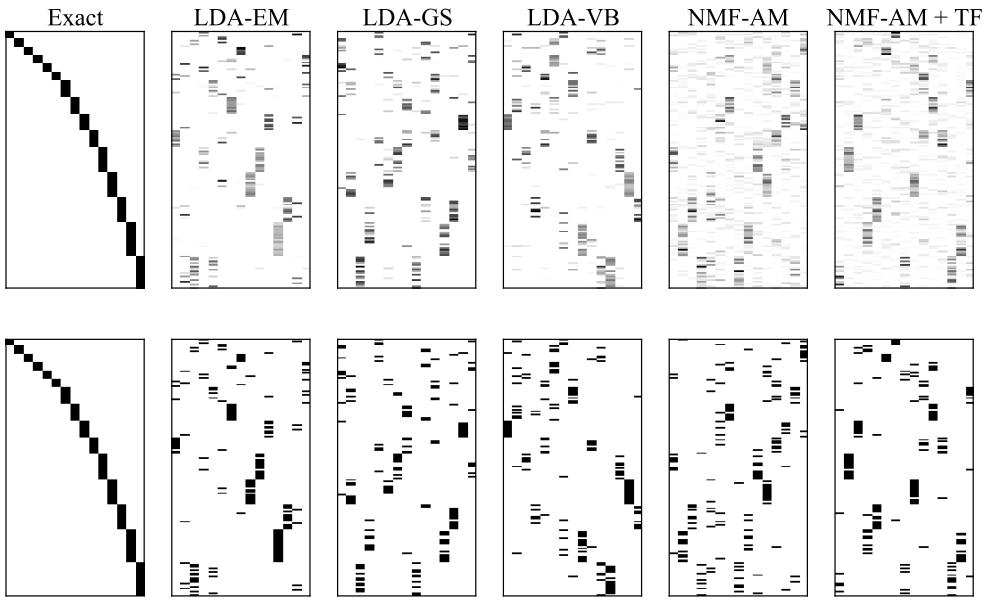

Figure 4.3: Topic distribution matrix $\mathbf{W}$, learned and exact, for "sparse" synthetic data with entries rounded to integers in $[0, 5]$ and 85% sparsity. The top row is the original data (darker means higher value with white equal to zero) and the bottom row demonstrates the row-wise maximal element indicating the accuracy of word assignments to topics. The word-topic assignment purity for the methods, from left, are 75%, 68%, 74%, 64%, and 79%.

The first demonstrations are based on synthetically generated matrices. In figures 4.3 and 4.4 we apply all algorithms to a small synthetic matrix formed using entirely disjoint topics, shown as the matrix on left, and a matrix $\mathbf{H}$. Two cases are considered. For the

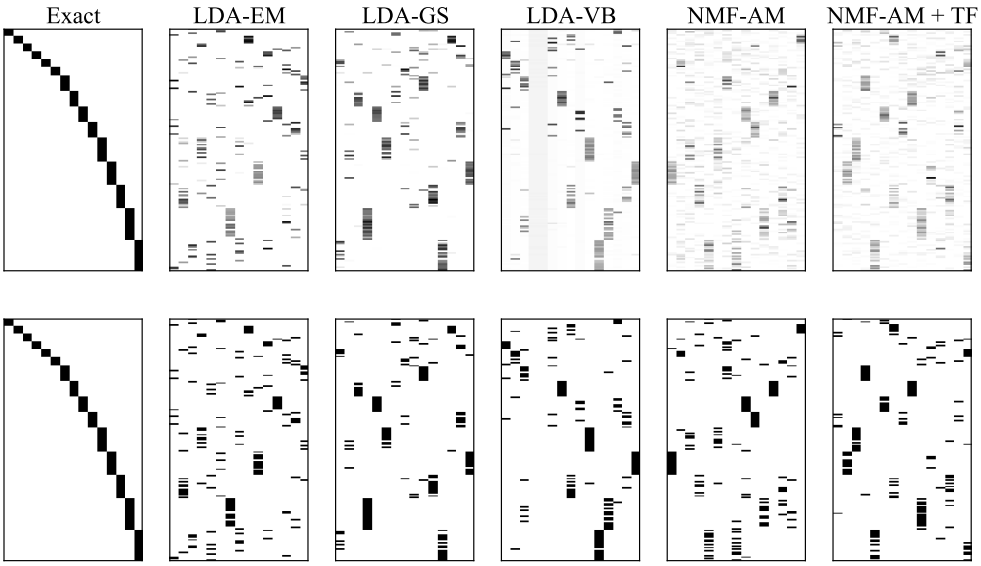

Figure 4.4: Topic distribution matrix $\mathbf{W}$, learned and exact, for "dense" synthetic data with entries rounded to $[0, 24]$ and 50% sparsity. The top row is the original data (darker means higher value with white equal to zero) and the bottom row demonstrates the row-wise maximal element indicating the accuracy of word assignments to topics. The word-topic assignment purity for the methods, from left, are 58%, 82%, 76%, 74%, and 75%.

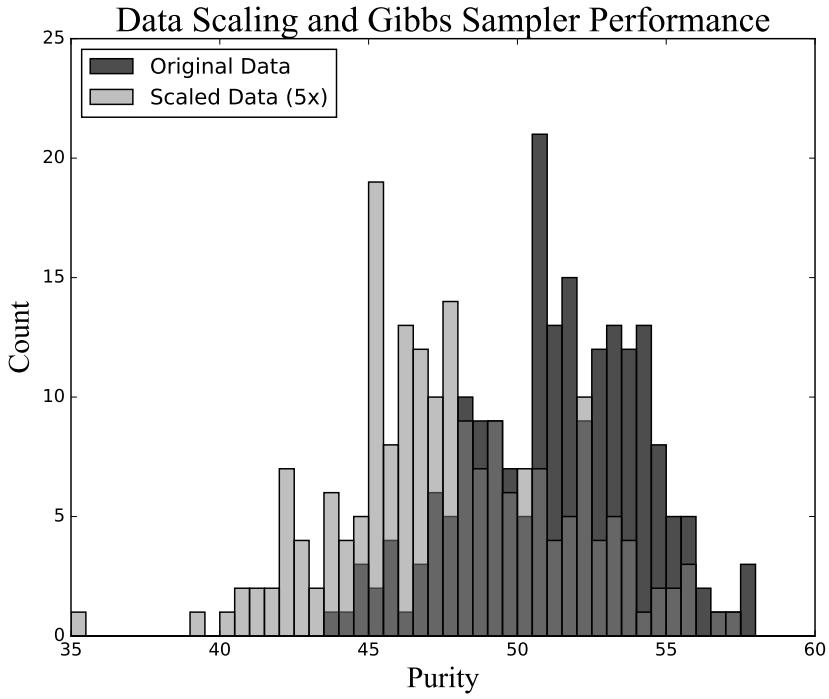

Figure 4.5: Percent of documents correctly classified using purity score for a subset of 10 classes taken from the 20 Newsgroups data set. The dark bars indicate the performance of the Gibbs sampler over 300 runs. The lighter histogram represents the performance of the Gibbs sampler after the data matrix is scaled by a factor of 5.

"dense" simulated data the matrix $\mathbf{H}$ is fully dense and the product $\mathbf{WH}$ is rounded to $[0, 24]$ with 50% sparsity randomly introduced to give the final data matrix. The "sparse" simulated data follows a similar process with $\mathbf{H}$ now 50% sparse, and the product scaled to $[0, 5]$ with 85% sparsity. All LDA-based algorithms are run with hyper-parameters $\alpha = 0.2$ and $\beta = 0.5$ with respective offsets as mentioned previously.

The first synthetic data results, or the "dense" example, are shown in figure 4.3. Here the Gibbs sampler is behind other methods as for this type of data the sampling process, described in section 4.6.2, becomes approximately gradient descent. The stochasticity of the sampling does not prove advantageous.

For the "sparse" data matrix, shown in figure 4.4, the Gibbs sampler outperforms other

LDA-GS	LDA-EM	LDA-VB	NMF-AM	NMF-AM, TF-IDF
62% (57%)	66% (57%)	50% (45%)	13% (12%)	52% (43%)

Table 4.2: Classification accuracy via purity score for each algorithm when applied to the 20 Newsgroups corpus. Shown is the best score over ten runs, with the average score in parenthesis.

techniques. The discrete count nature of the observations produces a setting in which continuum methods such as EM, VB, and AM can become stuck in sub-optimal minima while the sampler, being discrete in nature and stochastic, naturally handles the problem and easily finds good solutions.

This behavioural change of the Gibbs sampler arises with real data as well. We applied both LDA-GS and LDA-EM to a reduced version of the 20 Newsgroups corpus, using only ten of the classes to reduce the computational time, and calculated the classification accuracy for each of 300 runs. In the upper-left portion of figure 4.5, we see that the Gibbs sampler generally attains a higher document classification accuracy than EM frequently achieving nearly 60% accuracy compared to roughly 50% for EM. We then performed the same experiment three more times, each duplicating the data either once, twice, or thrice. The result is a decrease in performance of the Gibbs sampler, while the EM, VB, and NMF-based algorithms are unaffected by such a modification up to hyper-parameter adjustment.

4.9.2 20 Newsgroups

Here we apply all methods to the 20 Newsgroups data set extracting 20 topics with $\alpha = 0.1$ and $\beta = 0.5$. After the removal of stop words and words appearing less than twice the 11,314 documents are represented by histograms over a dictionary of 32,095 terms. The ideal topic model should be able to separate the newsgroups with clustering evidence in the matrix $\mathbf{H}$. We show this matrix in figure 4.6 for all methods, where each column indicates a distribution of one document over topics. The 20 Newsgroups clusters appear in sequence so that the

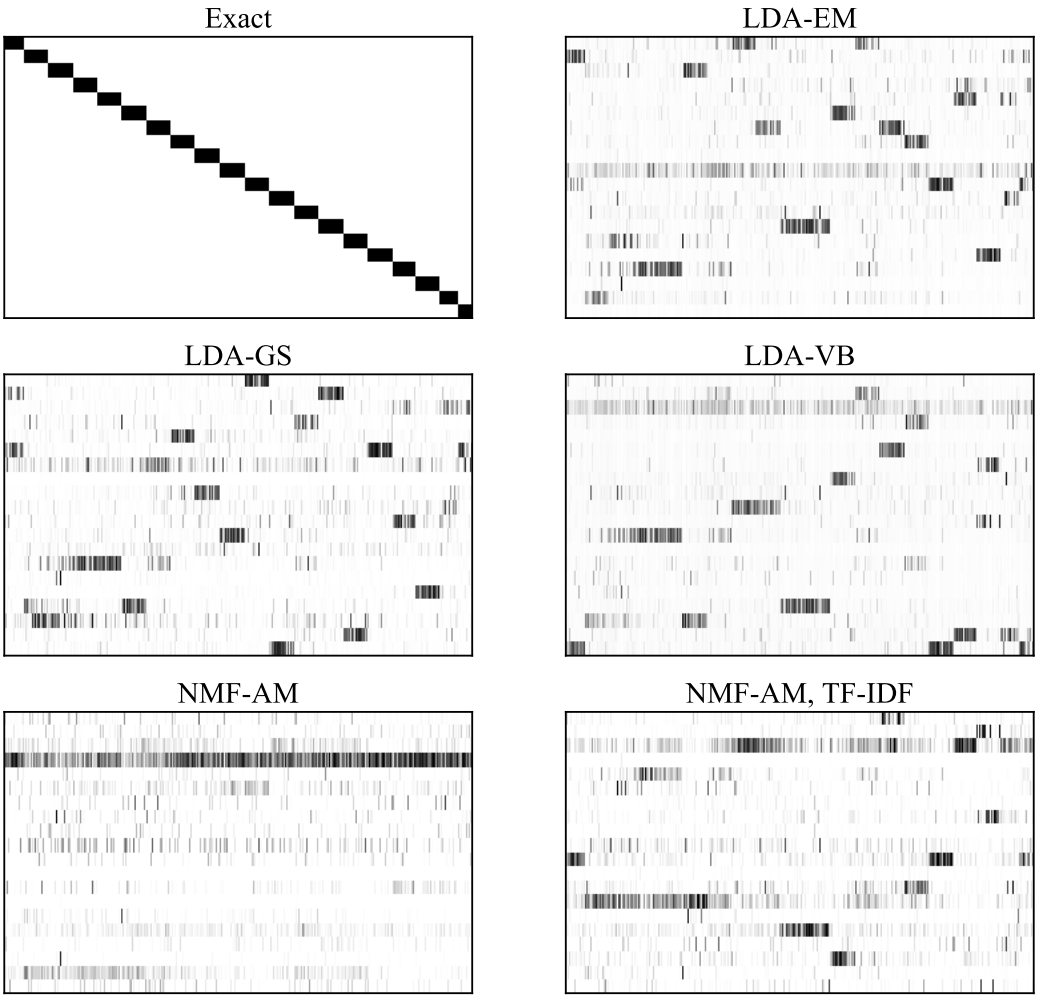

Figure 4.6: Learned matrices **H** for each algorithm when applied to the 20 Newsgroups corpus. Darker color indicated a higher value, with white equal to zero.

LDA-GS	LDA-EM	LDA-VB	NMF-AM	NMF-AM, TF-IDF
space	space	space	will	nasa
nasa	nasa	nasa	people	space
gov	gov	gov	subject	gov
will	organization	will	time	henry
access	subject	earth	god	jpl
earth	lines	launch	lines	launch
launch	writes	center	organization	orbit
center	article	data	writes	alaska
moon	will	orbit	article	moon
digex	access	1993	good	toronto

Table 4.3: Top words taken from topics with highest weight on the word "space" when each method is applied to the 20 Newsgroups corpus.

grouping of similar columns indicates clustering performance. The classification accuracy in terms of purity is in table 4.2. Both LDA-EM and LDA-GS perform comparably. LDA-VB manages to isolate a handful of topics but remains behind the other LDA algorithms overall. Finally, FRO-AM struggles to isolate the documents efficiently into clusters but the re-weighting of TF-IDF improves the data by penalizing more common words and emphasizing infrequent terms. This improvement, however, is not enough to separate the clusters as well as the LDA methods. The LDA model is capable of detecting the newsgroup clusters with performance that is comparable to prior art on unsupervised clustering of this data [BHL14], though notably there is considerable variation between the three LDA methods.

In table 4.3 we show an example topic for each method pertaining to space and NASA, where topics are selected by choosing the one with the highest probability of producing the word "space". All methods other than FRO-AM are able to discover well-defined topics. Because the data is not normalized, the FRO-AM model is dominated by common terms and longer documents resulting in very poor topics. The TF-IDF re-weighting corrects for this, but the LDA model is still able to model the data without TF-IDF very well.

4.9.3 Tweets of Madrid Evaluation

In this final study we qualitatively study the results of a topic model applied to a Twitter corpus and present the utility of simple metrics for quantifying spatial and temporal properties of topics. We extracted 300 topics using LDA-EM from a corpus consisting of tweets in the city of Madrid throughout the year 2011 that were created with geo-locations provided. As a result, the 1.4 million tweets we process have both a time and a place with which they are associated. This allows us to compute, for each topic, the spatial quantities MSD_s and LP_s for which the location in the city is used as well as the temporal quantities MSD_t and LP_t computed using the time within the year. For LP_s a 100-by-100 grid over the city is used to compute the histogram and $p = 0.8$. For LP_t a 100-bin discrete histogram over the year is used and $p = 0.1$.

In table 4.4 the most probable words for six example topics are displayed with names designated by the authors. These represent the variety of topics discovered in the data using the topic model. For example, the first column "FITUR" is the result of an international turism event that was held in the city. The "15-M" column corresponds to an anti-austerity protest that arose within Madrid, first gathering on the 15^{th} of March. "Airport" corresponds to tweets generated around the airport pertaining to travel. Finally, the last three displayed top words are the result of topics resulting from language differences in the city. Becuase the topic model discovers topics based on the co-occurrence of words in documents, here tweets, different languages tend to have disjoint topics. In this case, these three languages are rare among the predominantly Spanish corpus and therefore become confined to their own topics. The topics therefore capture a wide variety of activity taking place in the city.

In figure 4.7 we show for all topics the distributions of the various metrics herein proposed for understanding the distributions in space and time of the topics. Foremost, the two metrics though correlated do not capture identical information. This results from cases where topic activity is concentrated at a handful of locations with these locations spread apart. Such a situation produces a small LP metric value with a large MSD value. This can be see

by comparing the examples of temporal and spatial distributions in figures 4.8 and 4.9 respectively. The spatial distributions are shown here using Google's maps API.

The examples reveal the variety of distributions that arise on examination of the topics as well as exemplars for which the metric quantities are most clearly demonstrated. Uniform activity in time corresponds to topics such as "Airport" with activity that is associated with no particular time. In space an analogous property is found in the "Foursquare" topic that corresponds to user activity in the Foursquare social application across all locations of the city. Both temporally and spatially there are highly concentrated topics – in time there are events such as New Year's celebrations and in space there is the airport topic. Finally, there are a handful of cases where the metrics are not in agreement. This arises in cases such as the "Ticket Sales" topic in time and the "Three Wise Men" topic in space. For the former example activity is concentrated at a few points in the year with large periods of no activity. For the latter example activity is concentrated in space but at two locations. Hence these topics serve as examples where the study of both metrics is useful to understand the tweet distributions.

Quite evidently these metrics are capturing useful properties of the distributions. A quick examination of the scatter plots in figure 4.7 reveal interesting topics for further examination. In addition, the fact that the distributions of tweets within topics produce a variety of metric values serves as an additional validation for the performance of the topic model since the topic model has no knowledge of the additional tweet information.

4.10 Discussion

Topic modeling is a widely applicable approach to dimension reduction for non-negative high-dimensional and sparse data. Both non-negative matrix factorization and latent Dirichlet allocation approaches to the problem are impressive models with very disparate origin stories. As we have demonstrated, the underlying mechanism for why these models function the way they do is very similar. The common factor – non-negative matrix product representa-

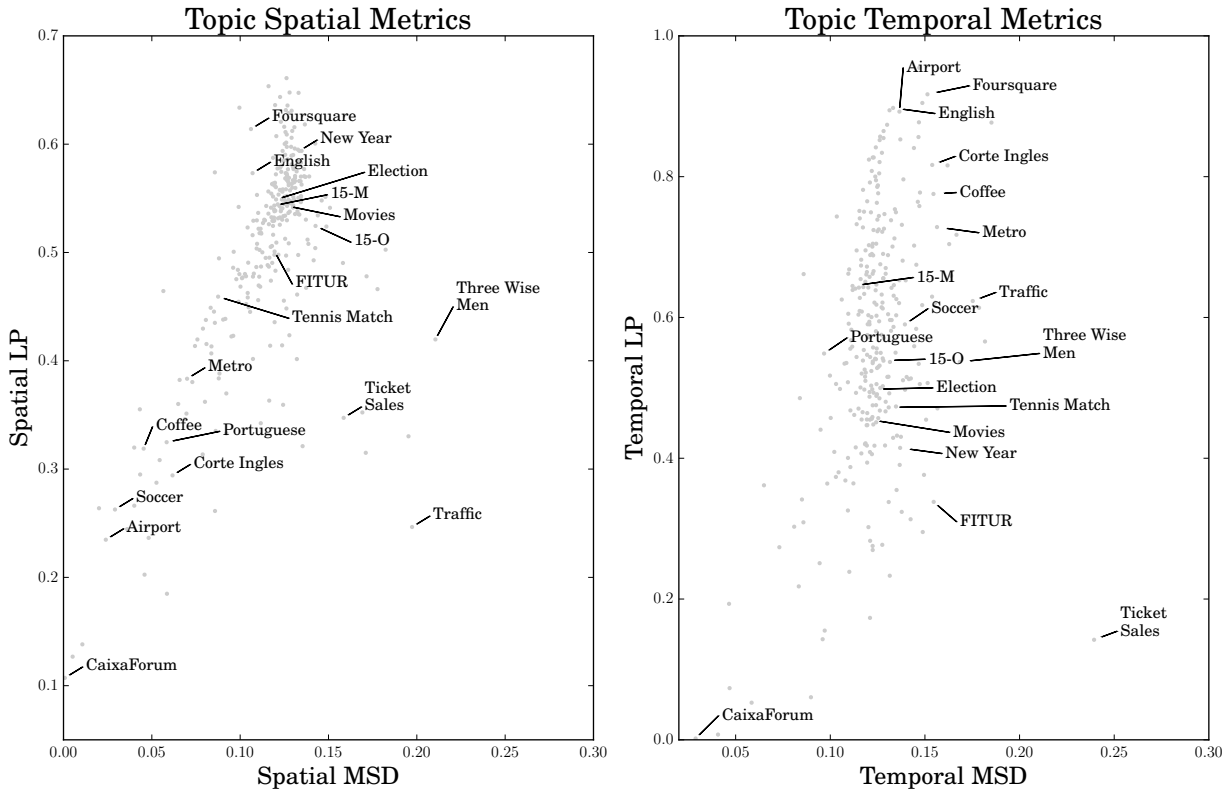

Figure 4.7: Metrics for all topics. Shown are the values of the spatial (left) and temporal (right) metrics proposed to study the topics learned from the corpus.

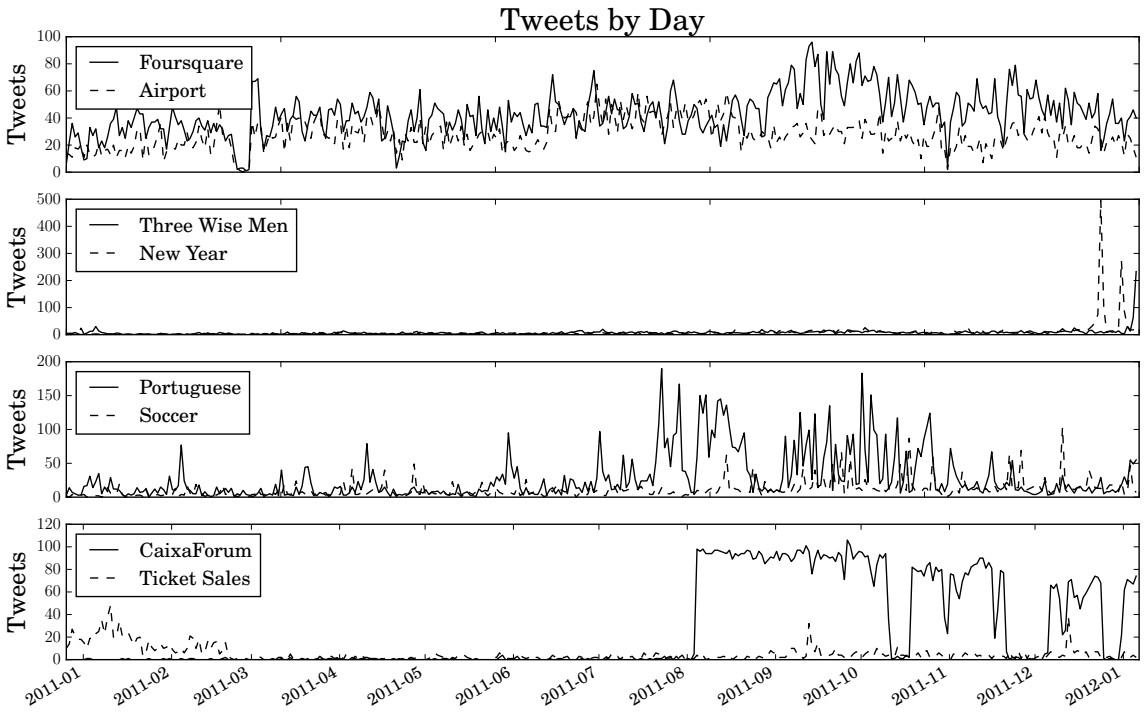

Figure 4.8: Types of temporal histograms. Different topics are characterized by different metric values that indicate the type of temporal activity. Shown are a few examples of background topics (top row), singular events in the year (second row), event topics with many activity spikes in the year (third row), and outliers arising from automated tweeting (bottom row). The metric values help to understand these distributions.

Table 4.4: Top words. Each topic is described by a probability distribution over words in the vocabulary. Shown here are the most probable words as learned by latent Dirichlet allocation when applied to geotagged tweets from the city of Madrid in 2011. The title for each topic is the author's interpretation.

FITUR (249)	15-M (74)	Airport (248)	English (26)	French (30)	Portuguese (91)
prensa	del	4	in	l	amor
internacional	sol	barajas	and	él	sueño
evento	sgae	mad	for	des	é
orgullo	campeonato	aeropuerto	day	et	eu
turismo	apertura	terminal	thanks	une	em
rueda	miau	t4	more	davidperez	não
francia	suchil	t1	us	à	um
aniversario	ancha	iberia	have	pas	pra
revista	selva	t2	nice	est	pro
fitur	carnes	airport	life	je	mais
marcatv	samurai	gate	last	ganitas	nose

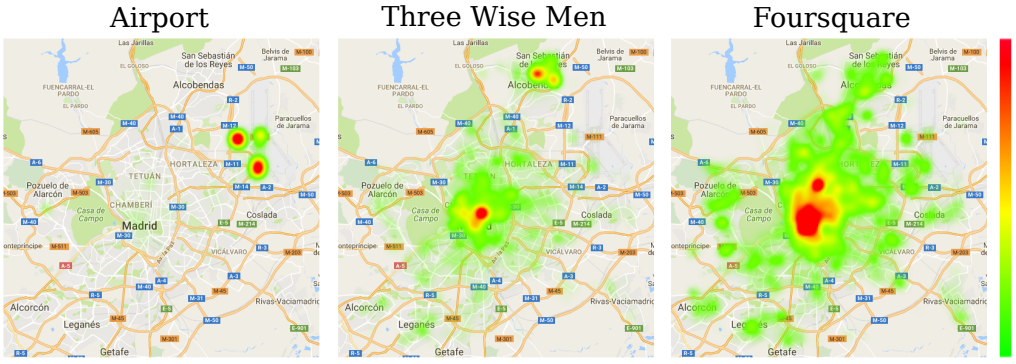

Figure 4.9: Example histograms in space via Google's mapping API. These three histograms demonstrate the characteristics captured by the metrics of figure 4.7: airport activity (small LP_s, small MSD_s), the Three Wise Men festival (small LP_s, large MSD_s), and check-ins to the Foursquare service (large LP_s, large MSD_s).

tions of data – results in a low-dimensional representation with each dimension remaining meaningful for interpretation, similar to a soft data clustering problem. The two methods, though appearing very different in statement, are analytically similar. We summarize this similarity by concluding that NMF with the Kullback-Liebler divergence is LDA with a Dirichlet penalty, up to normalization. This is unhelpful, unfortunately, when attempting to understand the inference techniques used for each model.

We therefore both summarize the popular inference techniques used in practice and contribute a consistent interpretation of these techniques in the same variational notation. Specifically, we demonstrate that in a variational setting the stochastic Gibbs sampler is, in the limit of many observations, simply a gradient method for the topic modeling problem. This re-framing of the Gibbs sampler also allows us to draw analogies with the expectation-maximization which is known to be a proximal-point technique and the alternating-minimization method which operates in an entirely different way but in the same high-dimensional tensor setting.

We demonstrate the behavioural expectations of the Gibbs sampler via numerical experimentation as well as the benefits of stochasticity in practical experiments. The benefits of matrix normalization in the case of NMF with the Frobenius divergence is also demonstrated numerically within a larger experiment demonstrating the clustering efficacy of the various topic modeling approaches outlined. Finally, in a practical application we demonstrate that topic models can be used with social media to produce a summary view of a city automatically. Additionally, because many hundreds of topics may result from a topic model on a large corpus, we proposed and studied a variety of metrics that quantify directly the spatial and temporal properties of topics. Using such tools, a practitioner is able to quickly sort through the topics and evaluate their distributional properties when additional information is available for each document in the corpus.

REFERENCES

[AGK12] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. "Computing a nonnegative matrix factorization–provably." In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 145–162. ACM, 2012.

[AGM12] Sanjeev Arora, Rong Ge, and Ankur Moitra. "Learning topic models–going beyond SVD." In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pp. 1–10. IEEE, 2012.

[AK06] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*, volume 147 of *Applied Mathematical Sciences*. Springer Science + Business Media, LLC, 2nd edition, 2006.

[Ama95] Shun-Ichi Amari. "Information geometry of the EM and em algorithms for neural networks." *Neural networks*, **8**(9):1379–1408, 1995.

[APK98] Philip J Allen, Giovanni Polizzi, Karsten Krakow, David R Fish, and Louis Lemieux. "Identification of EEG events in the MR scanner: the problem of pulse artifact and a method for its subtraction." *Neuroimage*, **8**(3):229–239, 1998.

[ASJ86] John B Adams, Milton O Smith, and Paul E Johnson. "Spectral mixture modeling: A new analysis of rock and soil types at the Viking Lander 1 site." *Journal of Geophysical Research: Solid Earth*, **91**(B8):8098–8112, 1986.

[AWS09] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. "On smoothing and inference for topic models." In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27–34. AUAI Press, 2009.

[BC16] William M Bolstad and James M Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.

[BHL14] Xavier Bresson, Huiyi Hu, Thomas Laurent, Arthur Szlam, and James von Brecht. "An incremental reseeding strategy for clustering." *arXiv preprint arXiv:1406.3837*, 2014.

[BL06a] David M Blei and John D Lafferty. "Dynamic topic models." In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM, 2006.

[BL06b] Peter Bunting and Richard Lucas. "The delineation of tree crowns in Australian mixed species forests using hyperspectral Compact Airborne Spectrographic Imager (CASI) data." *Remote Sensing of Environment*, **101**(2):230–248, 2006.

[BL07] David M Blei and John D Lafferty. "A correlated topic model of science." *The Annals of Applied Statistics*, pp. 17–35, 2007.

[Ble12] David M Blei. "Probabilistic topic models." *Communications of the ACM*, **55**(4):77–84, 2012.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet allocation." *the Journal of machine Learning research*, **3**:993–1022, 2003.

[BPC11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. "Distributed optimization and statistical learning via the alternating direction method of multipliers." *Foundations and Trends® in Machine Learning*, **3**(1):1–122, 2011.

[BQG86] G. Binnig, C. F. Quate, and Ch. Gerber. "Atomic Force Microscope." *Phys. Rev. Lett.*, **56**:930–933, Mar 1986.

[BS10] Christoph Braunsmann and Tilman E Schaffer. "High-speed atomic force microscopy for large scan sizes using small cantilevers." *Nanotechnology*, **21**(22):225705, 2010.

[BSC00] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. "Image Inpainting." In *Proceedings of the 27th annual conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pp. 417–424, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[BVS03] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. "Simultaneous structure and texture image inpainting." In *Proceedings of Computer Vision and Pattern Recognition Conference*, volume 2, pp. 707–712. IEEE Computer Society, June 2003.

[CCH09] Tsung-Han Chan, Chong-Yung Chi, Yu-Min Huang, and Wing-Kin Ma. "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing." *IEEE Transactions on Signal Processing*, **57**(11):4418–4432, 2009.

[CG92] George Casella and Edward I George. "Explaining the Gibbs sampler." *The American Statistician*, **46**(3):167–174, 1992.

[CH00] Stéphane Chrétien and Alfred O Hero III. "Kullback proximal algorithms for maximum-likelihood estimation." *Information Theory, IEEE Transactions on*, **46**(5):1800–1810, 2000.

[Cha03] Chein-I Chang. *Hyperspectral imaging: techniques for spectral detection and classification*, volume 1. Springer Science & Business Media, 2003.

[Cha09] Rick Chartrand. "Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data." In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pp. 262–265. IEEE, 2009.

[CHH11] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. "Graph regularized nonnegative matrix factorization for data representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**(8):1548–1560, 2011.

[CMS98] V. Caselles, J.M. Morel, and C. Sbert. "An axiomatic approach to image interpolation." *Transactions on Image Processing*, **7**(3):376–386, March 1998.

[Cra94] Maurice D Craig. "Minimum-volume transforms for remotely sensed data." *IEEE Transactions on Geoscience and Remote Sensing*, **32**(3):542–552, 1994.

[CSD01] D. Croft, G. Shed, and S. Devasia. "Creep, hysteresis, and vibration compensation for piezoactuators: Atomic force microscopy applications." *Journal of Dynamic Systems, Measurement, and Control*, **123**(1):35–43, 2001.

[CWT11] A. Chen, T. Wittman, A.G. Tartakovsky, and A.L. Bertozzi. "Efficient Boundary Tracking Through Sampling." *Applied Mathematics Research eXpress*, **2011**(2):182–214, 2011.

[DBM13] Sven Dähne, Felix Bießmann, Frank C Meinecke, Jan Mehnert, Siamac Fazli, and Klaus-Robert Müller. "Integration of multivariate data streams with bandpower signals." *IEEE Transactions on Multimedia*, **15**(5):1001–1013, 2013.

[DHS05] Chris HQ Ding, Xiaofeng He, and Horst D Simon. "On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering." In *SDM*, volume 5, pp. 606–610. SIAM, 2005.

[DHV15] Lucas Drumetz, Simon Henrot, Miguel Angel Veganzones, Jocelyn Chanussot, and Christian Jutten. "Blind hyperspectral unmixing using an Extended Linear Mixing Model to address spectral variability." In *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2015)*, 2015.

[DLF09] Jean Daunizeau, Helmut Laufs, and Karl J Friston. "EEG–fMRI information fusion: biophysics and data analysis." In *EEG-fMRI*, pp. 511–526. Springer, 2009.

[DLP08] Chris Ding, Tao Li, and Wei Peng. "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing." *Computational Statistics & Data Analysis*, **52**(8):3913–3927, 2008.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[DS03] David Donoho and Victoria Stodden. "When does non-negative matrix factorization give a correct decomposition into parts?" In *Advances in neural information processing systems*, p. None, 2003.

[DSS07] Stefan Debener, Alexander Strobel, Bettina Sorger, Judith Peters, Cornelia Kranczioch, Andreas K Engel, and Rainer Goebel. "Improved quality of auditory event-related potentials recorded simultaneously with 3-T fMRI: removal of the ballistocardiogram artefact." *Neuroimage*, **34**(2):587–597, 2007.

[EMO12] Ernie Esser, Michael Möller, Stanley Osher, Guillermo Sapiro, and Jack Xin. "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space." *Image Processing, IEEE Transactions on*, **21**(7):3239–3252, 2012.

[FC09] Cédric Févotte and A Taylan Cemgil. "Nonnegative matrix factorizations as probabilistic inference in composite models." In *Signal Processing Conference, 2009 17th European*, pp. 1913–1917. IEEE, 2009.

[GMD14] Jérôme Gilles, Travis Meyer, and Pamela K. Douglas. "Leveraging Sparsity: A Low-Rank + Sparse Decomposition (LR+SD) Method for Automatic EEG Artifact Removal." In *Proceedings of the Second International Workshop on Sparsity Techniques in Medical Imaging (STMI)*, pp. 80–88, Boston, USA, 2014.

[GMM09] Leo Gross, Fabian Mohn, Nikolaj Moll, Peter Liljeroth, and Gerhard Meyer. "The Chemical Structure of a Molecule Resolved by Atomic Force Microscopy." *Science*, **325**(5944):1110–1114, 2009.

[GO09] T. Goldstein and S. Osher. "The Split Bregman Method for L1-Regularized Problems." *SIAM Journal on Imaging Sciences*, **2**(2):323–343, 2009.

[GSE02] Robin I Goldman, John M Stern, Jerome Engel Jr, and Mark S Cohen. "Simultaneous EEG and fMRI of the alpha rhythm." *Neuroreport*, **13**(18):2487, 2002.

[GT04] DMBTL Griffiths and MIJJB Tenenbaum. "Hierarchical topic models and the nested Chinese restaurant process." *Advances in neural information processing systems*, **16**:17, 2004.

[GV14] Nicolas Gillis and Stephen Vavasis. "Fast and robust recursive algorithms for separable nonnegative matrix factorization." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **36**(4):698–714, 2014.

[GZ12] Pinghua Gong and Changshui Zhang. "Efficient nonnegative matrix factorization via projected Newton method." *Pattern Recognition*, **45**(9):3557–3565, 2012.

[HBG15] Zhipeng Hao, Mark Berman, Yi Guo, Glenn Stone, and Iain Johnstone. "Semi-realistic simulations of natural hyperspectral scenes." In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, pp. 1004–1007. IEEE, 2015.

[HBW13] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. "Stochastic variational inference." *The Journal of Machine Learning Research*, **14**(1):1303–1347, 2013.

[HDT14] Abderrahim Halimi, Nicolas Dobigeon, and Jean-Yves Tourneret. "Unsupervised unmixing of hyperspectral images accounting for endmember variability." *arXiv preprint arXiv:1406.5071*, 2014.

[HF95] Alfred O Hero and Jeffrey A Fessler. "Convergence in norm for alternating expectation-maximization (EM) type algorithms." *Statistica Sinica*, **5**(1):41–54, 1995.

[HMH05] A. D. L. Humphris, M. J. Miles, and J. K. Hobbs. "A mechanical microscope: High-speed atomic force microscopy." *Applied Physics Letters*, **86**(3):–, 2005.

[Hof99] Thomas Hofmann. "Probabilistic latent semantic analysis." In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

[HSF06] P K Hansma, G Schitter, G E Fantner, and C Prater. "High-Speed Atomic Force Microscopy." *Science*, **314**(5799):601–602, October 2006.

[Hun10] Shao-Kang Hung. "Spiral Scanning Method for Atomic Force Microscopy." *J. Nanosci. Nanotech.*, **10**(7):45114516, Jul 2010.

[JGJ99] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An introduction to variational methods for graphical models." *Machine learning*, **37**(2):183–233, 1999.

[JMH00] Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J Mckeown, Vicente Iragui, and Terrence J Sejnowski. "Removing electroencephalographic artifacts by blind source separation." *Psychophysiology*, **37**(02):163–178, 2000.

[KDP12] Da Kuang, Chris Ding, and Haesun Park. "Symmetric nonnegative matrix factorization for graph clustering." In *Proceedings of the 2012 SIAM international conference on data mining*, pp. 106–117. SIAM, 2012.

[KFC04] Johannes H Kindt, Georg E Fantner, Jackie A Cutroni, and Paul K Hansma. "Rigid design of fast scanning probe microscopes using finite element analysis." *Ultramicroscopy*, **100**(34):259 – 265, 2004. Proceedings of the Fifth International Conference on Scanning Probe Micrscopy, Sensors and Nanostructures.

[KLH11] Priyanka Kohli, Jeff Lyons, Andrew D. L. Humphris, Benjamin D. Bunday, Abraham Arceo, Akira Hamaguchi, Dilip Patel, and David Bakker. "High-speed atmospheric imaging of semiconductor wafers using rapid probe microscopy." *Proc. SPIE*, **7971**:797119–797119–9, 2011.

[KM02] Nirmal Keshava and John F Mustard. "Spectral unmixing." *IEEE signal processing magazine*, **19**(1):44–57, 2002.

[KP07] Hyunsoo Kim and Haesun Park. "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis." *Bioinformatics*, **23**(12):1495–1502, 2007.

[KP11] Jingu Kim and Haesun Park. "Fast nonnegative matrix factorization: An active-set-like method and comparisons." *SIAM Journal on Scientific Computing*, **33**(6):3261–3281, 2011.

[KSD07] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon. "Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem." In *SDM*, volume 7, pp. 343–354. SIAM, 2007.

[KSD13] Matthieu Kowalski, Kai Siedenburg, and Monika Dorfler. "Social sparsity! neighborhood systems enrich structured shrinkage operators." *Signal Processing, IEEE Transactions on*, **61**(10):2498–2511, 2013.

[KSK12] Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur. "Fast conical hull algorithms for near-separable non-negative matrix factorization." *arXiv preprint arXiv:1210.1190*, 2012.

[KYI10] Noriyuki Kodera, Daisuke Yamamoto, Ryoki Ishikawa, and Toshio Ando. "Video imaging of walking myosin V by high-speed atomic force microscopy." *Nature*, **468**(7320):72–76, April 2010.

[LB08] Jun Li and José M Bioucas-Dias. "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data." In *IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pp. III–250. IEEE, 2008.

[LCM10] Zhouchen Lin, Minming Chen, and Yi Ma. "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices." *arXiv preprint arXiv:1009.5055*, 2010.

[Lin07] Chuan-bi Lin. "Projected gradient methods for nonnegative matrix factorization." *Neural computation*, **19**(10):2756–2779, 2007.

[LLP12] Liangda Li, Guy Lebanon, and Haesun Park. "Fast Bregman divergence nmf using Taylor expansion and coordinate descent." In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 307–315. ACM, 2012.

[LS99] Daniel D Lee and H Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature*, **401**(6755):788–791, 1999.

[LS01] Daniel D Lee and H Sebastian Seung. "Algorithms for non-negative matrix factorization." In *Advances in neural information processing systems*, pp. 556–562, 2001.

[LSK12] Chengbo Li, Ting Sun, Kevin F Kelly, and Yin Zhang. "A compressive sensing and unmixing scheme for hyperspectral data processing." *Image Processing, IEEE Transactions on*, **21**(3):1200–1210, 2012.

[MAF07] Richard AJ Masterton, David F Abbott, Steven W Fleming, and Graeme D Jackson. "Measurement and reduction of motion and ballistocardiogram artefacts from simultaneous EEG and fMRI recordings." *Neuroimage*, **37**(1):202–211, 2007.

[MB08] Jon D Mcauliffe and David M Blei. "Supervised topic models." In *Advances in neural information processing systems*, pp. 121–128, 2008.

[MM09] I A Mahmood and S O Reza Moheimani. "Fast spiral-scan atomic force microscopy." *Nanotechnology*, **20**(36):365503, 2009.

[MM10] I.A. Mahmood and S.O.R. Moheimani. "Spiral-scan Atomic Force Microscopy: A constant linear velocity approach." In *Nanotechnology (IEEE-NANO), 2010 10th IEEE Conference on*, pp. 115–120, Aug 2010.

[MM12] David Mimno and Andrew McCallum. "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression." *arXiv preprint arXiv:1206.3278*, 2012.

[MRR53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. "Equation of state calculations by fast computing machines." *The journal of chemical physics*, **21**(6):1087–1092, 1953.

[MZB14] Travis R. Meyer, Dominik Ziegler, Christoph Brune, Alex Chen, Rodrigo Farnham, Nen Huynh, Jen-Mei Chang, Andrea L. Bertozzi, and Paul D. Ashby. "Height drift correction in non-raster atomic force microscopy." *Ultramicroscopy*, **137**:48 – 54, 2014.

[NBI05] RK Niazy, CF Beckmann, GD Iannetti, JM Brady, and SM Smith. "Removal of FMRI environment artifacts from EEG data using optimal basis sets." *Neuroimage*, **28**(3):720–737, 2005.

[ND05a] José MP Nascimento and José M Bioucas Dias. "Does independent component analysis play a role in unmixing hyperspectral data?" *Geoscience and Remote Sensing, IEEE Transactions on*, **43**(1):175–187, 2005.

[ND05b] José MP Nascimento and José M Bioucas Dias. "Vertex component analysis: A fast algorithm to unmix hyperspectral data." *Geoscience and Remote Sensing, IEEE Transactions on*, **43**(4):898–910, 2005.

[NJW02] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. "On spectral clustering: Analysis and an algorithm." *Advances in neural information processing systems*, **2**:849–856, 2002.

[Paa97] Pentti Paatero. "Least squares formulation of robust non-negative factor analysis." *Chemometrics and intelligent laboratory systems*, **37**(1):23–35, 1997.

[PBU07] L M Picco, L Bozec, A Ulcinas, D J Engledew, M Antognozzi, M A Horton, and M J Miles. "Breaking the speed limit with atomic force microscopy." *Nanotechnology*, **18**(4):044030, 2007.

[QJZ11] Yuntao Qian, Sen Jia, Jun Zhou, and Antonio Robles-Kelly. "Hyperspectral unmixing via sparsity-constrained nonnegative matrix factorization." *Geoscience and Remote Sensing, IEEE Transactions on*, **49**(11):4282–4297, 2011.

[RGS04] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. "The author-topic model for authors and documents." In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494. AUAI Press, 2004.

[RPP14] M.S. Rana, H.R. Pota, and I.R. Petersen. "Spiral Scanning With Improved Control for Faster Imaging of AFM." *Nanotechnology, IEEE Transactions on*, **13**(3):541–550, May 2014.

[RRT12] Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. "Factoring nonnegative matrices with linear programs." In *Advances in Neural Information Processing Systems*, pp. 1214–1222, 2012.

[Rud91] Walter Rudin. *Functional Analysis*. McGraw-Hill Science/Engineering/Math, 1991.

[SB88] Gerard Salton and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." *Information processing & management*, **24**(5):513–523, 1988.

[SF14] Dennis L Sun and Cedric Fevotte. "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence." In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6201–6205. IEEE, 2014.

[SG08] Ajit P Singh and Geoffrey J Gordon. "A unified view of matrix factorization models." In *Machine Learning and Knowledge Discovery in Databases*, pp. 358–373. Springer, 2008.

[SH05] Amnon Shashua and Tamir Hazan. "Non-negative tensor factorization with applications to statistics and computer vision." In *Proceedings of the 22nd international conference on Machine learning*, pp. 792–799. ACM, 2005.

[SPB15] Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. "Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1045–1054. ACM, 2015.

[STH08] Georg Schitter, Philipp J. Thurner, and Paul K. Hansma. "Design and input-shaping control of a novel scanner for high-speed atomic force microscopy." *Mechatronics*, **18**(5-6):282 – 288, 2008. Special Section on Optimized System Performances Through Balanced Control StrategiesThe 4th {IFAC} Symposium on Mechatronic Systems - Mechatronics 2006.

[SZP12] Ben Somers, Maciel Zortea, Antonio Plaza, and Gregory P Asner. "Automated extraction of image-based endmember bundles for improved spectral unmixing." *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, **5**(2):396–408, 2012.

[THR12] T. Tuma, W. Haeberle, H. Rothuizen, J. Lygeros, A. Pantazi, and A. Sebastian. "A dual-stage nanopositioning approach to high-speed scanning probe microscopy." In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 5079–5084, Dec 2012.

[TVR11] Rachel Thornton, Serge Vulliemoz, Roman Rodionov, David W Carmichael, Umair J Chaudhary, Beate Diehl, Helmut Laufs, Christian Vollmar, Andrew W McEvoy, Matthew C Walker, et al. "Epileptic networks in focal cortical dysplasia revealed using electroencephalography–functional magnetic resonance imaging." *Annals of neurology*, **70**(5):822–837, 2011.

[VP14] Konstantin Vorontsov and Anna Potapenko. "Additive regularization of topic models." *Machine Learning*, pp. 1–21, 2014.

[VSS09] Pedro A Valdes-Sosa, Jose Miguel Sanchez-Bornot, Roberto Carlos Sotero, Yasser Iturria-Medina, Yasser Aleman-Gomez, Jorge Bosch-Bayard, Felix Carbonell, and Tohru Ozaki. "Model driven EEG/fMRI fusion of brain oscillations." *Human brain mapping*, **30**(9):2701–2721, 2009.

[Wie01] M Wieczorowski. "Spiral sampling as a fast way of data acquisition in surface topography." *International Journal of Machine Tools and Manufacture*, **41**(1314):2017 – 2022, 2001.

[Win99] Michael E Winter. "N-FINDR: an algorithm for fast autonomous spectral endmember determination in hyperspectral data." In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pp. 266–275. International Society for Optics and Photonics, 1999.

[WMM09] Hanna M Wallach, David M Mimno, and Andrew McCallum. "Rethinking LDA: Why priors matter." In *Advances in neural information processing systems*, pp. 1973–1981, 2009.

[Wu83] CF Jeff Wu. "On the convergence properties of the EM algorithm." *The Annals of statistics*, pp. 95–103, 1983.

[WWH09] Sinead Williamson, Chong Wang, Katherine Heller, and David Blei. "Focused topic models." In *NIPS workshop on Applications for Topic Models: Text and Beyond, Whistler, Canada*, 2009.

[WZ13] Yu-Xiong Wang and Yu-Jin Zhang. "Nonnegative matrix factorization: A comprehensive review." *Knowledge and Data Engineering, IEEE Transactions on*, **25**(6):1336–1353, 2013.

[XJ96] Lei Xu and Michael I Jordan. "On convergence properties of the EM algorithm for Gaussian mixtures." *Neural computation*, **8**(1):129–151, 1996.

[XLG03] Wei Xu, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 267–273. ACM, 2003.

[YB14] Felipe Yanez and Francis Bach. "Primal-Dual Algorithms for Non-negative Matrix Factorization with the Kullback-Leibler Divergence." *arXiv preprint arXiv:1412.1788*, 2014.

[YGH15] Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. "Lightlda: Big topic models on modest computer clusters." In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1351–1361. ACM, 2015.

[ZH14] Alina Zare and KC Ho. "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing." *Signal Processing Magazine, IEEE*, **31**(1):95–104, 2014.

[ZHD12] Mingyuan Zhou, Lauren Hannah, David B Dunson, and Lawrence Carin. "Beta-Negative Binomial Process and Poisson Factor Analysis." In *International Conference on Artificial Intelligence and Statistics*, pp. 1462–1471, 2012.

[ZHJ12] Yuan Zou, John Hart, and Roozbeh Jafari. "Automatic EEG artifact removal based on ica and hierarchical clustering." In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 649–652. IEEE, 2012.

[ZLC16] Jia Zeng, Zhi-Qiang Liu, and Xiao-Qin Cao. "Fast Online EM for Big Topic Modeling." *IEEE Transactions on Knowledge and Data Engineering*, **28**(3):675–688, 2016.

[ZMA16] Dominik Ziegler, Travis R Meyer, Andreas Amrein, Andrea Bertozzi, and Paul D Ashby. "Ideal Scan Path for High-Speed Atomic Force Microscopy." *IEEE/ASME Transactions on Mechatronics*, 2016.

[ZMF13] Dominik Ziegler, Travis R Meyer, Rodrigo Farnham, Christoph Brune, Andrea L Bertozzi, and Paul D Ashby. "Improved accuracy and speed in scanningprobe microscopy by image reconstruction from non-gridded positionsensor data." *Nanotechnology*, **24**(33):335703, 2013.

[ZSC10] Yusheng Zhou, Guangyi Shang, Wei Cai, and Jun-en Yao. "Cantilevered bimorph-based scanner for high speed atomic force microscopy with large scanning range." *Review of Scientific Instruments*, **81**(5):–, 2010.

[ZWF14] Feiyun Zhu, Ying Wang, Bin Fan, Shiming Xiang, Geofeng Meng, and Chunhong Pan. "Spectral unmixing via data-guided sparsity." *Image Processing, IEEE Transactions on*, **23**(12):5412–5427, 2014.