UNIVERSITY OF CALIFORNIA SAN DIEGO

Structured Learning with Scale Mixture Priors

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

in

Electrical Engineering (Signal and Image Processing)

by

Igor Fedorov

Committee in charge:

Professor Bhaskar D. Rao, Chair
Professor Truong Q. Nguyen, Co-Chair
Professor Ery Arias-Castro
Professor Sebastian Obrzut
Professor Lawrence Saul
Professor Nuno Vasconcelos

2018

The dissertation of Igor Fedorov is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2018

DEDICATION

To my wife and family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Throughout my graduate studies, I benefited from the help and advice of many people. The following is my attempt to thank those that made my journey possible. First, I would like to thank my parents for giving me the opportunity to pursue my interests throughout my academic career. The chance to pursue a PhD is a privilege not available to many people and, at least in my case, represents the culmination of my parents' numerous sacrifices and years of hard work.

The process of working towards a PhD can be very grueling, from both an intellectual and emotional point of view. I was lucky enough to meet my wife, Katarina, at the beginning of my studies. Katarina's love and support propelled me through many of the hardships one inevitably faces in graduate school.

When I first came to UCSD, I was rather green in my ability to do meaningful scientific research. I was lucky enough be advised by Prof. Bhaskar Rao, who guided me towards becoming an independent researcher. I cannot overstate how much I learned from Prof. Rao. His patience, humility, and creativity set an example for me to strive to emulate. He taught me how to formulate research questions worth pursuing. Most importantly, he showed me the value of distilling complicated engineering problems into concise mathematical formulations, which can then be illuminated by a fundamental understanding of the mathematics involved.

I would also like to thank Prof. Truong Nguyen for welcoming me into his lab and supporting me throughout my PhD. In addition, I am very thankful for my committee members: Prof. Nuno Vasconcelos, Prof. Lawrence Saul, Prof. Ery Arias-Castro, and Prof. Sebastian Obrzut. Each of the committee members has given me valuable suggestions during the course of my work. I am especially grateful for Prof. Obrzut, who introduced me to the field of biomedical imaging.

I am very thankful for my fellow graduate students in the UCSD ECE department. I had the pleasure of collaborating with several students, including Ritwik Giri, Alican Nalci, and Maher Al-Shoukairi. I had many fruitful discussions with Yonatan Vaizman, Srinjoy Das, Narek

Rostomyan, and Narek Geghamyan.

Finally, I am thankful for the San Diego Chapter of the ARCS Foundation, Inc. for providing me with financial support and guidance during my graduate studies.

Chapter 2, in full, is a reprint of material published in the article Igor Fedorov, Ritwik Giri, Bhaskar D. Rao, and Truong Q. Nguyen, "Robust Bayesian Method for Simultaneous Block Sparse Signal Recovery with Applications to Face Recognition," IEEE International Conference on Image Processing, 2016. I was the primary author and B. D. Rao supervised the research.

Chapter 3, in full, is a reprint of material published in the article Igor Fedorov, Alican Nalci, Ritwik Giri, Bhaskar D. Rao, Truong Q. Nguyen, and Harinath Garudadri, "A Unified Framework for Sparse Non-Negative Least Squares using Multiplicative Updates and the Non-Negative Matrix Factorization Problem," Signal Processing, 2018. I was the primary author and B. D. Rao supervised the research.

Chapter 4, in full, is a reprint of material in Igor Fedorov and Bhaskar D. Rao, "Multimodal Sparse Bayesian Dictionary Learning," under review at the IEEE Transactions on Signal Processing. I was the primary author and B. D. Rao supervised the research.

VITA

2012	B. S. in Electrical Engineering, University of Illinois, Urbana-Champaign
2014	M. S in Electrical Engineering, University of Illinois, Urbana-Champaign
2018	Ph. D. in Electrical Engineering, University of California, San Diego

PUBLICATIONS

I. Fedorov, B.D. Rao, "Multimodal Sparse Bayesian Dictionary Learning," *under review at IEEE Transactions on Signal Processing*.

I. Fedorov, A. Nalci, R. Giri, B.D. Rao, T.Q. Nguyen, H. Garudadri, "A Unified Framework for Sparse Non-Negative Least Squares using Multiplicative Updates and the Non-Negative Matrix Factorization Problem," *Signal Processing*, 2018.

A. Nalci, **I. Fedorov**, M. Al-Shoukairi, T. T. Liu, B.D. Rao. "Rectified Gaussian Scale Mixtures and the Sparse Non-Negative Least Squares Problem," *IEEE Transactions on Signal Processing*, 2018.

I. Fedorov, B.D. Rao, T.Q. Nguyen, "Multimodal Sparse Bayesian Dictionary Learning Applied to Multimodal Data Classification," *IEEE Conference on Acoustic, Speech, and Signal Processing*, 2017.

I. Fedorov, S. Obrzut, B. Song, B.D. Rao, "SPECT Image Reconstruction under Imaging Time Constraints," *Asilomar Conference on Signals, Systems and Computers*, 2017.

I. Fedorov, B. Song, B.D. Rao, I. Levitan, S. Obrzut, "Total Variation Regularization in I-123 Ioflupane SPECT Reconstruction," *Journal of Nuclear Medicine*, 2017.

I. Fedorov, R. Giri, B.D. Rao, T.Q. Nguyen, "Robust Bayesian Method for Simultaneous Block Sparse Signal Recovery with Applications to Face Recognition." *IEEE International Conference on Image Processing (ICIP)*, 2016.

ABSTRACT OF THE DISSERTATION

Structured Learning with Scale Mixture Priors

by

Igor Fedorov

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California San Diego, 2018

Professor Bhaskar D. Rao, Chair
Professor Truong Q. Nguyen, Co-Chair

Sparsity plays an essential role in a number of modern algorithms. This thesis examines how we can incorporate additional structural information within the sparsity profile and formulate a richer class of learning approaches. The focus is on Bayesian techniques for promoting sparsity and developing novel priors and inference schemes.

The thesis begins by showing how structured sparsity can be used to recover simultaneously block sparse signals in the presence of outliers. The approach is validated with empirical results on synthetic data experiments as well as the multiple measurement face recognition problem.

In the next portion of the thesis, the focus is on how structured sparsity can be used to extend approaches for dictionary learning. Dictionary learning refers to decomposing a data matrix into the product of a dictionary and coefficient matrix, subject to a sparsity constraint on the coefficient matrix.

Chapter 3 studies structure in the form of non-negativity constraints on the unknowns, which is referred to as the sparse non-negative least squares (S-NNLS) problem. It presents a unified framework for S-NNLS based on a novel prior on the sparse codes and provides an efficient multiplicative inference procedure. It then extends the framework to sparse non-negative matrix factorization (S-NMF) and proves that the proposed approach is guaranteed to converge to a set of stationary points for both the S-NNLS and a subclass of the S-NMF problems.

Finally, Chapter 4 addresses the problem of learning dictionaries for multimodal datasets. It presents the multimodal sparse Bayesian dictionary learning (MSBDL) algorithm. The MSBDL algorithm is able to leverage information from all available data modalities through a joint sparsity constraint on each modality's sparse codes without restricting the coefficients themselves to be equal. The proposed framework offers a considerable amount of flexibility to practitioners and addresses many of the shortcomings of existing multimodal dictionary learning approaches. Unlike existing approaches, MSBDL allows the dictionaries for each data modality to have different cardinality. In addition, MSBDL can be used in numerous scenarios, from small datasets to extensive datasets with large dimensionality. MSBDL can also be used in supervised settings.

Chapter 1

Introduction

1.1 Sparse Signal Recovery

Much of the motivation for the results in this thesis originates in the problem of sparse signal recovery (SSR). Signal recovery refers to the problem of recovering a coefficient vector x from an observation y generated according to the signal model

$$y = Dx, \tag{1.1}$$

where $D \in \mathbb{R}^{N \times M}$ represents the forward model from unknowns to observations. In certain applications, D is overdetermined, i.e. $N > M$, such that solving the least squares problem

$$\arg\min_x \|y - Dx\|_2^2 \tag{1.2}$$

can recover x with a reasonably high fidelity. Nevertheless, in a large number of cases D is underdetermined, i.e. $N < M$. For instance, in biomedical imaging, D represents the mapping embodied by the camera system. Each row of D represents the acquisition of one measurement of the patient. In order to limit the imaging time, one must reduce the number of measurements,

leading to an underdetermined D [1].

The fundamental challenge in recovering x in underdetermined problems is that there is an infinite number of solutions to (1.1). For any x which satisfies (1.1) and x_n in the null space of D, i.e. $Dx_n = 0$, $x + x_n$ also satisfies (1.1). Since the null space is guaranteed to be non-empty for underdetermined D, it is impossible to identify the true x which generated the given observations from the other possible candidates.

In order to resolve the identifiability issues inherent in (1.1), one possibility is to restrict the space of possible solutions. In many practical situations, including medical imaging [1], photography [2], and face recognition [3], it is reasonable to assume that x is sparse, meaning that x has a small number of non-zero elements. In this case, the SSR problem can be stated as

$$\arg\min_{y=Dx} \|x\|_0 \tag{1.3}$$

where $\|\cdot\|_0$ refers to the ℓ_0 pseudo-norm, which counts the number of non-zero entries. The benefit of the SSR formulation is that the solution to (1.3) is unique under some restrictions on D and x. The downside to (1.3) is that it is NP-hard [4]. As such, there is a wealth of research in approximating (1.3). Greedy methods, such as orthogonal matching pursuit [5], seek to find the support of x by beginning with an initially empty estimate of the support set and iteratively adding a single element to the support set using a locally optimal decision rule. Other methods seek to approximate the ℓ_0 norm in (1.3) by the ℓ_1 norm, which is both sparsity promoting and convex, thereby making the entire optimization problem convex [6]. Although the ℓ_1 norm has desirable optimization properties, there are many other non-convex norms which exhibit superior SSR performance [7, 8, 9]. Interestingly, it is possible to transform the resulting non-convex optimization problems to convex ones by bounding the non-convex norm by its linearization around a given estimate. The resulting algorithms often take the form of weighted ℓ_1 norm [8] or weighted ℓ_2 norm minimization problems.

1.2 Dictionary Learning

Whereas the dictionary D is assumed to be known in the context of SSR, it is well known that, if possible, learning D leads to superior performance in a range of tasks covering image denoising [10], classification [11, 12, 13, 14], and audio source separation [15], among many others. In fact, dictionary learning was first proposed as a model of the human visual system [16]. Broadly speaking, given a dataset $Y = \begin{bmatrix} y^1 & \cdots & y^L \end{bmatrix}$, the dictionary learning problem can take the form

$$\underset{D,\{\|x^i\|_0 \leq s\}_{i=1}^L}{\arg\min} \|Y - DX\|_2^2 \tag{1.4}$$

where $X = \begin{bmatrix} x^1 & \cdots & x^L \end{bmatrix}$. The optimization in (1.4) is often performed in a block-coordinate descent fashion, where D is updated while holding X fixed, followed by updating the sparse codes $\{x^i\}_{i=1}^L$ while holding D fixed, with the procedure iterated until a given termination criterion is met. In other words, the optimization strategy is summed up by iterating

$$\text{Update } D \text{ given } X \tag{1.5}$$

$$\text{Update } X \text{ given } D. \tag{1.6}$$

The motivation behind the block-coordinate descent scheme is to leverage mature existing sparse coding algorithms for (1.6). Interestingly, the original work by Olshausen et al. [16] used a simple gradient descent procedure for (1.5)-(1.6).

The difference among the large array of available dictionary learning algorithms is in the method of sparsity promotion. For instance, the K-SVD algorithm seeks to solve (1.4) directly, using a greedy sparse coding scheme for the coefficient update stage. On the other hand, the approach by Mairal et al. [17] replaces the ℓ_0 norm by its convex ℓ_1 norm relaxation. The resulting optimization problem then has the desirable property that both the sparse coding and

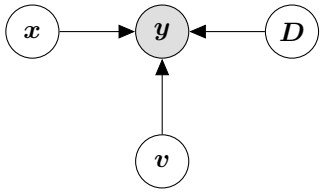


Figure 1.1: Graphical model for dictionary learning.

dictionary update sub-problems are convex. Other approaches replace the ℓ_0 norm in (1.4) with the ℓ_p norm, $0 \leq p < 1$ [18].

1.3 Bayesian Techniques

In the Bayesian framework, all of the unknowns are represented by random variables and the task is to estimate the most likely values of those random variables given the observations by maximizing an appropriate cost function. Let $\boldsymbol{y}, \boldsymbol{x}$, and $\boldsymbol{D}$ represent the observation vector, the coefficient vector, and the dictionary, respectively. The assumption is that the relationship among these random variables is given by

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} + \boldsymbol{v} \tag{1.7}$$

where $\boldsymbol{v}$ is a random variable representing noise, i.e. $\boldsymbol{v} \sim \mathsf{N}\left(v; 0, \sigma^2 \mathsf{I}\right)$. A visualization of the dictionary learning problem is given by the graphical model in Fig. 1.1. There are no direct connections between $\boldsymbol{x}$ and $\boldsymbol{D}$, reflecting the assumption that $\boldsymbol{x}$ and $\boldsymbol{D}$ are a-priori independent. Unless otherwise stated, we also assume a non-informative prior on $\boldsymbol{D}$. The key remaining component is to specify the prior distribution on $\boldsymbol{x}$. As before, the goal is promote sparse coefficient vectors. In the following, we review sparsity promoting distributions and describe Bayesian inference techniques.

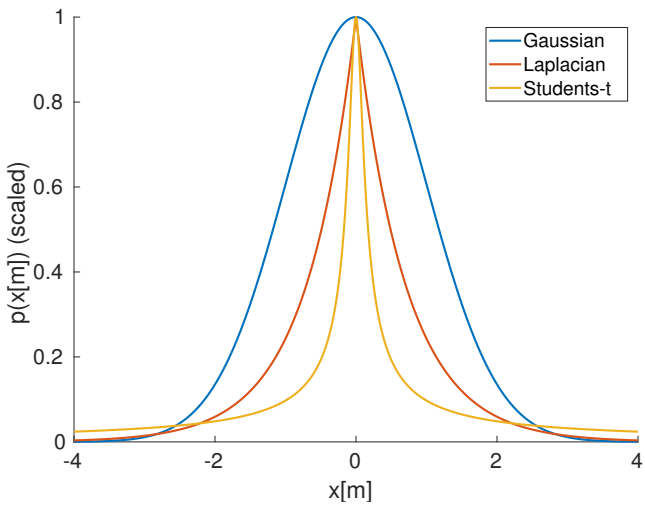


Figure 1.2: Visualization of the pdf of super-Gaussian distributions. All pdf's have been scaled such that $p(0) = 1$ for visualization purposes.

1.3.1 Sparsity Promoting Distributions

For simplicity and ease of exposition, we assume that $p(x)$ is separable:

$$p(x) = \prod_{m=1}^{M} p(x[m]) \tag{1.8}$$

where $x[m]$ denotes the m'th entry of x. In order to promote sparsity, it is important to choose a super-Gaussian $p(x[m])$. Loosely speaking, a given distribution[1] is super-Gaussian if it is more peaky around the origin than a Gaussian distribution [19]. Fig. 1.2 shows the difference between a Gaussian probability density function (pdf) and two super-Gaussian pdf's, including the Laplacian and Student's-t. Note that the pdf's in Fig. 1.2 have been scaled for visualization purposes. A more formal characterization of super-Gaussianity states that a scalar random variable $\boldsymbol{x}$ is super-Gaussian if the kurtosis of $\boldsymbol{x}$ is greater than that of a Gaussian with the same variance

[1] In the context of assessing super-Gaussianity, we assume that the given distribution is symmetric about the origin.

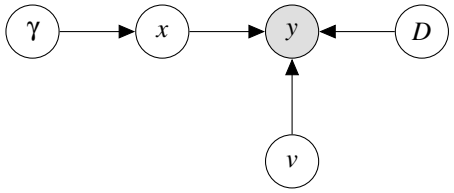


Figure 1.3: Hierarchical graphical model for dictionary learning.

as $\boldsymbol{x}$, where kurtosis is defined as

$$kurtosis = \frac{E_{\boldsymbol{x}}\left[(x-\mu_x)^4\right]}{\sigma_x^4} \tag{1.9}$$

and μ_x and σ_x denote the mean and standard deviation of $\boldsymbol{x}$, respectively [19]. For reference, the kurtosis of the standard Laplacian distribution is 6[2].

In most instances, a super-Gaussian density can be represented as a Gaussian scale mixture (GSM), i.e.

$$p(x) = \int_0^\infty \mathsf{N}(x;0,\gamma)\, p(\gamma)\, d\gamma \tag{1.10}$$

where the choice of prior on $\boldsymbol{\gamma}$ determines the eventual form of $p(x)$ [20][3]. For example, if $\boldsymbol{\gamma}[\boldsymbol{m}] \sim p^{Exp}\left(\gamma[m];\frac{\tau^2}{2}\right)$, then $\boldsymbol{x}[\boldsymbol{m}] \sim p^{Exp}(x[m];\tau)$ where $p^{Exp}(\cdot)$ refers to the Exponential or, equivalently, Laplacian pdf's. Likewise, if $\boldsymbol{\gamma}[\boldsymbol{m}] \sim p^{\mathsf{IGa}}\left(\gamma[m];\frac{\tau}{2},\frac{\tau}{2}\right)$, then $\boldsymbol{x}[\boldsymbol{m}] \sim p^{\mathsf{ST}}\left(x^i[m];\tau\right)$, where $p^{IGa}(\cdot)$ and $p^{ST}(\cdot)$ refer to the Inverse-Gamma and Student's-t pdf's, respectively.

In order to incorporate the hierarchical representation of $p(x)$, we modify the graphical model in Fig. 1.1 to the one in Fig. 1.3. Although the decomposition of $p(x)$ may appear to be adding more complexity to the model than needed, in the following we will show that the GSM prior enables two types of inference, with each type offering its own set of trade-offs.

[2]Interestingly, the kurtosis is undefined for the Student's-t distribution with degrees of freedom less than 4, which is the case in Fig. 1.2.

[3]Theorem 3 in [20] gives the exact conditions which a given distribution must satisfy in order to admit a GSM representation. All of the super-Gaussian densities used in this thesis admit such a representation.

1.3.2 Inference for Sparsity Promoting Priors: Type I (MAP)

Given the graphical model in Fig. 1.3, one sensible option for estimating $\boldsymbol{x}$ is to form the maximum a-posteriori (MAP) estimator:

$$\arg\max_{X,D} p\left(X,D|Y\right). \tag{1.11}$$

As before, a block-coordinate descent scheme similar to (1.5)-(1.6) can be adopted, where the dictionary update stage consists of solving

$$\begin{aligned}\arg\max_{D} p\left(D|Y,X\right) &= \arg\min_{D} -\log p\left(D|Y,X\right)\\ &= \arg\min_{D} p\left(Y|D,X\right)\\ &= \arg\min_{D} \|Y-DX\|_F^2 .\end{aligned} \tag{1.12}$$

The coefficient update stage is given by

$$\begin{aligned}\arg\max_{X} p\left(X|Y,D\right) &= \arg\min_{X} -\log p\left(X|Y,D\right)\\ &= \arg\min_{X} -\log p\left(Y|D,X\right) - \log p\left(X\right)\\ &= \arg\min_{X} \sum_{i=1}^{L} \left\|y^i - Dx^i\right\|_2^2 - 2\sigma^2 \log p\left(x^i\right)\end{aligned} \tag{1.13}$$

where it is assumed that x^i and x^j are a-priori independent for $i \neq j$.

One important observation is that (1.13) looks exactly like an SSR problem where the regularizer is determined by $p\left(x\right)$. The equivalence between regularized SSR problems derived in the deterministic paradigm and MAP estimates in the Bayesian paradigm is well documented, most notably in [7]. For example, consider the choice of regularizer in [8], which leads to the

following SSR optimization problem

$$\arg\min_x \|y - Dx\|_2^2 + \lambda \sum_{m=1}^{M} \log\left(|x[m]| + \tau\right) \tag{1.14}$$

where λ and τ are user-specified parameters. One of the challenges with solving (1.14) is that the first term is convex whereas the second term is concave. The approach taken in [8] is to bound the concave term by a linear function of $x[m]$ at a given estimate of the coefficients, x^t, leading to the convex weighted ℓ_1-norm regularized problem

$$\arg\min_x \|y - Dx\|_2^2 + \lambda \left\| \frac{x}{|x^t| + \tau} \right\|_1 \tag{1.15}$$

where we define the division of two vectors as elementwise division. The advantage of (1.15) over (1.14) is that (1.15) is convex and can be solved by one of many mature convex optimization packages (i.e. [21]). By iterating between solving (1.15) and re-computing the bound of the concave term, the objective in (1.14) is iteratively minimized. In other words, the objective in (1.14) is non-increasing under the sequence of updates $\{x^t\}_{t=1}^{\infty}$. Because the original problem is non-convex, it is not possible to guarantee convergence to a global minimum of (1.14). Moreover, it is not even clear whether the sequence $\{x^t\}_{t=1}^{\infty}$ itself has a limit point. The only claim that is made in [8] is that the objective function itself converges to a local minimum. The contribution made in [7] is to show that (1.14) is equivalent to the MAP estimate of $\boldsymbol{x}$ under a Generalized Double Pareto prior and that the iterative re-weighting strategy in [8] is equivalent to applying the expectation-maximization (EM) algorithm [22] to (1.13) with a clever choice of nuisance variable. The fundamental insight is to treat γ in Fig. 1.3 as the nuisance variable, leading to the EM update rule

$$x^{t+1} = \arg\min_x \|y - Dx\|_2^2 + \lambda \sum_{m=1}^{m} x[m] E_{\gamma[m]|\boldsymbol{x}[m]} \left[\gamma[m]^{-1}\right]. \tag{1.16}$$

In fact, we can compute $E_{\gamma[m]|x[m}\left[\gamma[m]^{-1}\right]$ without even forming $p(\gamma[m]|x[m])$ using a trick from [23, 24]:

$$E_{\gamma[m]|x[m]}\left[\gamma[m]^{-1}\right] = \frac{\partial \log p(x[m])}{\partial x[m]}. \tag{1.17}$$

Note that the expression in (1.17) applies specifically to the Generalized Double Pareto prior, but similar expressions can be derived for other sparsity promoting priors [7, 25]. The connection of re-weighted methods to Bayesian methods in [7] lays the foundation for Chapter 3, where we study the non-negatively constrained dictionary learning problem (i.e. non-negative matrix factorization) and leverage the convergence properties of the EM algorithm [26] to prove that our approaches converge to the set of stationary points of the underlying objective function.

1.3.3 Inference for Sparsity Promoting Priors: Type II (Evidence Maximization)

In the following, we begin by describing the evidence-maximization approach to SSR and dictionary learning before giving an overview of how this strategy leads to sparse solutions [27, 28].

In the context of SSR, the evidence-maximization framework proceeds by first forming a maximum-likelihood estimate of γ:

$$\gamma_{ML} = \arg\max_{\gamma} p(y|\gamma). \tag{1.18}$$

This estimate is then used to approximate the true posterior $p(x|y)$ by $p(x|y;\gamma_{ML})$. For the choice of a Gaussian data-likelihood, it turns out that the posterior $p(x|y,\gamma)$ is Gaussian. As a result, (1.18) lends itself to EM optimization with closed form update rules [27].

If we extend the SSR evidence maximization recipe to dictionary learning, the goal

becomes to find [29]

$$\gamma_{ML}, D_{ML} = \arg\max_{\gamma, D} p\left(y|\gamma, D\right). \tag{1.19}$$

As before, (1.19) can be optimized through the use of the EM algorithm [30, 31]. One unexpected result of the EM procedure detailed in [30, 31][4] is that both γ and D are naturally updated simultaneously. In other words, there is no need for block-coordinate descent to solve (1.19), in contrast with many existing dictionary learning approaches which directly solve for the MAP (or, equivalently, regularized least squares) estimate [10, 17, 32].

In practice, it is observed that evidence-maximization inference is superior to MAP inference for both the SSR and dictionary learning problems [27, 31, 7, 33]. At the same time, it is not self-evident why evidence-maximization should produce sparse estimates nor why the sparse estimates that are produced should be better than those obtained by directly maximizing the posterior $p\left(x|y\right)$ (or $p\left(X, D|Y\right)$ for dictionary learning). Partial answers to these questions in the context of SSR can be found in [27, 7, 34, 35], which we summarize next. Empirical evidence for the superiority of evidence maximization in the context of dictionary learning can be found in [30, 31][5], but there remains a theoretical gap in understanding why direct MAP estimation is inferior in this context.

The general analysis in [34, 35] gives some intuition about why evidence-maximization works so well, at least in the context of SSR. The central idea is that evidence-maximization naturally embodies the Occam's razor principle [36], which states that simple models are preferable to complex ones. The question of whether or not the Occam's razor principle is true universally is a philosophical one, but it seems that it is very fitting to the SSR problem. Indeed, one could argue that the Occam's razor principle is just a high-level statement of the SSR problem: Look for the coefficient vector with smallest number of non-zero entries that still aligns with the observations.

[4] The EM algorithm for dictionary learning is described in full detail in Chapter 4.

[5] Chapter 4 of this thesis.

Two additional arguments for evidence-maximization can be found in [27, 7]:

- Evidence-maximization is more robust than MAP inference [7].
- Evidence-maximization offers a tractable approximation to MAP inference when direct MAP inference is intractable (i.e. in the case where MAP inference requires solving an NP-hard problem) [27].

The claim that evidence-maximization is more robust than MAP estimation is supported by analysis which shows that γ_{ML} can be viewed as the maximizer of the posterior mass of $\boldsymbol{x}$ over the set of non-zero indices of γ_{ML} [7]. If the posterior has multiple modes, then a MAP estimator may give a poor estimate if the highest posterior peak occurs in an area of small posterior mass. On the other hand, evidence maximization will only look for areas of high mass. Of course, the analysis in [7] rests on the assumption that the set of non-zero indices of γ_{ML} and $\boldsymbol{x}$ are the same, which is difficult to guarantee in practice. Theorem 1 in [27] guarantees that the non-zero indices of γ_{ML} coincide with the non-zero indices of x if x is the sparsest solution to $y = Dx$, assuming that there is no noise in the observations. The challenge is that the objective in (1.18) is non-convex, so it is not possible to guarantee that γ_{ML} would be found in practice, as opposed to one of the stationary points of the objective. Moreover, any practical system will inherently pollute the observations y with noise, violating the assumptions of Theorem 1 in [27].

Lastly, it is informative to consider the argument presented in [27], which proves that (1.18) represents maximizing the evidence of a variational approximation to the actual signal prior, which may not lend itself to direct MAP estimation. Due to the geometry of the true prior and the properties of evidence-maximization, the estimates generated by this optimized variational approximation tend to be sparse.

1.4 Thesis Outline and Contributions

As is the trend in Bayesian learning literature, there is a constant battle between the complexity of the prior and the tractability of the inference scheme. This thesis seeks to extend the class of signal priors and Bayesian inference schemes available for researchers in both the SSR and dictionary learning communities. The goal throughout this work is to devise models which better fit a given generative process in the real-world and then study how Type I and Type II inference can be performed for these novel priors.

Chapter 2 covers SSR in the context of multiple observations which are subjected to both stationary and non-stationary noise. Our main contribution is a novel hierarchical signal model that incorporates three key properties:

- the signal is sparse and the sparsity profile does not change with time,
- the observations are corrupted by stationary Gaussian noise, and
- the observations are corrupted by non-stationary sparse noise, whose sparsity profile changes with time.

It turns out that this model is well-aligned with the conditions encountered in recognition problems in videos and we show that our approach achieves considerable improvement in classification accuracy in such scenarios.

Chapter 3 studies the problem of sparse non-negative least squares and sparse non-negative matrix factorization, which correspond to the SSR and dictionary learning problems under the constraint that the unknowns are non-negative. Our contributions are summarized by:

- We provide a general class of signal priors which admit a hierarchical representation. This class of priors encompasses a number of existing priors used in the literature as well as other, novel priors.

- We describe a unified MAP inference framework for the proposed class of signal priors. Specifically, inference is embodied by a set of multiplicative update rules, which is a low-complexity, widely used optimization technique in the non-negative matrix factorization literature.

- We prove that the proposed inference framework admits convergence guarantees.

Finally, Chapter 4 extends the evidence-maximization framework for dictionary learning to multimodal dictionary learning. Multimodal dictionary learning refers to learning representations for multiple datasets simultaneously, subject to the constraint that the atoms of the learned representations are related to each other in some way. Our contributions are:

- We provide several novel signal priors which extend the class of associations that can be learned by existing multimodal dictionary learning algorithms. The priors we propose allow learning dictionaries whose cardinality is a function of modality, which is distinctly unique to our work.

- We show that inference can be done at low memory and computational cost. In other words, we make steps towards making our approach scalable to large datasets, which Bayesian approaches often have trouble with.

- We provide an automatic hyperparameter tuning strategy, which obviates the need for performing a search over a hyperparameter space which grows exponentially with the number of modalities.

- We incorporate supervised learning into the proposed framework.

- We conduct a theoretical analysis of the proposed framework.

Chapter 2

Robust Bayesian Method for Simultaneous Block Sparse Signal Recovery with Applications to Face Recognition

2.1 Introduction

Sparse Signal recovery (SSR) refers to algorithms which seek sparse solutions to underdetermined systems of equations [4], which occur naturally when one seeks a representation of a given signal under an overcomplete dictionary. Overcomplete dictionaries have gained popularity in a wide range of applications because they are much more flexible than their undercomplete counterparts and lead to unique solutions under certain constraints, when sparsity has been enforced [37]. Constraining the solution of underdetermined problems to be sparse represents prior knowledge about the solution and makes finding it tractable. In certain applications, structured sparsity, such as block sparsity, has been enforced on the desired coefficient vector, i.e., a small number of blocks of the solution are non-zero [38].

SSR has become a very active research area in recent times because of its wide range

of engineering applications. For example, in several popular computer vision problems, such as face recognition [3], motion segmentation [39], and activity recognition [40], signals lie in low-dimensional subspaces of a high dimensional ambient space. An important class of methods to deal with this depends on exploiting the notion of sparsity. Following this path, Sparse Representation based Classification (SRC) [3] was proposed and produced state of the art results in a face recognition (FR) task.

In many applications, we often encounter outliers in measurements, which leads traditional SSR algorithms to fail and necessitates the development of an outlier robust SSR algorithm. The need for outlier resistant SSR algorithms motivates our present work, in which we develop a robust SSR algorithm and extend it to recover simultaneously block sparse signals. To show the efficacy of our approach, we focus on FR, which refers to identifying a subject's face given a labeled database of faces. The pioneering work of Wright et al. [3] on SRC showed that a face classifier can be devised by using the downsampled images from the training database as a dictionary and considering the sparse representation of a given image under that dictionary as the "identity" of the person. In this scenario, it is intuitive to assume that the dictionary is broken up into blocks corresponding to each specific person and constraining the encoding of the image to be block-sparse leads to performance gains [41].

One significant challenge in the FR problem is dealing with occlusions. Occlusions are outliers within the SRC model because the model assumes that the dictionary spans the space of all possible observations. A popular way to incorporate robustness to outliers into the SSR model is to assume that the outliers themselves have a sparse representation [3], which has been shown to yield improved resilience to various forms of face occlusion and corruption [3, 42, 43, 41, 44]. In certain cases, such as when the entire face is occluded or lighting conditions are extremely poor, FR within the SRC framework can yield unsatisfactory results because it is difficult to solve the single measurement vector (SMV) SSR problem. When possible, it is advantageous to acquire multiple measurements of the same source and instead solve the multiple measurement vector

(MMV) problem. The MMV problem assumes that the support of the non-zero coefficients that encode each measurement does not change, while the actual values of the coefficients can vary. It is well known that the MMV problem yields much better recovery results than the SMV problem [45].

In this work, we extend the SRC framework to the MMV case and consider performing FR when multiple images of the same subject, corrupted by non-stationary occlusions, are presented to the classifier. Our work is motivated, in part, by the person re-identification problem [46, 47, 48]. Srikrishna et al. [48] addressed the re-identification problem by applying SSR to each individual image of the subject and aggregating the results to form a global classifier. As such, [48] did not address the MMV nature of the problem. The main motivation behind our work is to enforce the prior knowledge that the input images correspond to the same person within the SSR process, while still maintaining resilience to time-varying occlusions.

Our SSR framework builds upon the hierarchical Bayesian framework discussed in [49, 50, 51], known as Sparse Bayesian Learning (SBL). This choice is motivated by the superior recovery results obtained for the standard SSR problem [50, 7] and the Bayesian framework is convenient for extensions to problems with structure [52]. In this work, we extend the SBL framework to the MMV block-sparse case and explicitly model time-varying occlusions, referring to our method as robust SBL (Ro-SBL).

2.1.1 Contributions

- We introduce a novel hierarchical Bayesian Robust SSR algorithm, Ro-SBL, for solving the MMV block-sparse problem with time-varying outliers. This work has connections to [44], where a Robust Block Sparse Bayesian Learning (BSBL) method was proposed. In contrast with our work, BSBL only considered the SMV problem and did not harness the ability of the SBL framework to capture non-stationary outliers.

- We validate our proposed method with synthetic data results and also apply our method to a robust simultaneous FR task. Unlike [48], our proposed approach exploits the prior knowledge that the input images correspond to the same person within the SSR process.

2.2 Ro-SBL for Simultaneous Block Sparse Recovery

The signal model for simultaneous block sparse recovery is given by

$$\boldsymbol{Y} = D\boldsymbol{X} + \boldsymbol{E} + \boldsymbol{V} \tag{2.1}$$

where, $\boldsymbol{Y} \in \mathbb{R}^{N\times L}$ is the matrix of L measurements, $D \in \mathbb{R}^{N\times M}$ is the dictionary, $\boldsymbol{V} \in \mathbb{R}^{N\times L}$ is the independent and identically distributed Gaussian noise term with mean zero and variance σ^2, $\boldsymbol{X} \in \mathbb{R}^{M\times L}$ is the encoding of the measurements under D, and $\boldsymbol{E} \in \mathbb{R}^{N\times L}$ is the matrix containing the outliers in the measurements.

The key assumption in the MMV problem is that, if a given column of D is activated (i.e. its corresponding coefficient in X is non-zero) for one of the measurements, then it will be activated for all of the measurements [45]. This means that the same set of basis vectors have been used to generate all of the measurements, which is reflected in the encoding matrix X in the form of joint sparsity, i.e. $\{x^i\}_{i=1}^{L}$ share the same support, where x^i is the i'th column of X [45]. Within a Bayesian framework, the joint sparsity assumption translates to placing a prior on the *rows* of X. In the context of our work, we build upon the extension of the SBL framework to the MMV problem in [50] and adopt a hierarchical prior, namely a Gaussian Scale Mixture (GSM), over the rows of $\boldsymbol{X}$:

$$p\left(X[m,:]|\gamma[m]\right) = \mathsf{N}\left((X[m,:])^T;0,\gamma[m]\mathsf{I}\right) \tag{2.2}$$

where, $X[m,:]$ denotes the m'th row of X and $\gamma[m]$ is the unknown variance hyperparameter. In addition, we also consider block sparsity in each x^i, where the block structure is shared

among all of the encoding vectors. Assuming that the support of x^i is separated into disjoint sets $\mathscr{T}^k, 1 \leq k \leq K$, which are known a-priori and shared across all i, $p(\boldsymbol{X})$ is amended to reflect that each of the rows in a given group $\mathscr{T}^k$ share the same $\boldsymbol{\gamma}[\boldsymbol{k}]$ [52]:

$$p(X) = \prod_{k=1}^{K} \prod_{m \in \mathscr{T}^k} p(X[m,:]|\gamma[k]) \tag{2.3}$$

Although the joint block sparsity constraint is a valid one for the encoding matrix $\boldsymbol{X}$, it does not hold for the outlier matrix $\boldsymbol{E}$ since the outliers could be non-stationary, i.e., time varying. Therefore, we will treat each $\boldsymbol{e^i}$ independently and not constrain the outliers to share the same support across all measurements. As such, we adopt a sparsity enforcing GSM prior on $\boldsymbol{e^i}$, which induces the following prior on $\boldsymbol{E}$:

$$p(E|\Delta) = \prod_{n=1,i=1}^{N,L} p\left(e^i[n]|\delta^i[n]\right) = \prod_{n=1,i=1}^{N,L} \mathsf{N}\left(e^i[n]; 0, \delta^i[n]\right) \tag{2.4}$$

where $\Delta = \begin{bmatrix} \delta^i & \cdots & \Delta^L \end{bmatrix}$. This set of assumptions is unique to this work and is motivated by the FR task.

2.2.1 Incorporating Robustness to Outliers

For an SMV problem, [44][53] showed that sparse (under the standard basis) outliers can be incorporated into the well known SBL framework by introducing a simple modification to the dictionary D. In the present work, we extend this idea to the MMV case, which results in the following modification to the signal model in (2.1):

$$\boldsymbol{Y} = \underbrace{\begin{bmatrix} D & \mathsf{I} \end{bmatrix}}_{\tilde{\boldsymbol{A}}} \underbrace{\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{E} \end{bmatrix}}_{\tilde{\boldsymbol{X}}} + \boldsymbol{V}. \tag{2.5}$$

Note that (2.5) and (2.1) are equivalent, but, as will be shown next, the signal model in (2.5) lends itself much more nicely to a closed form inference procedure.

2.2.2 Ro-SBL Inference Procedure

The goal of the inference procedure is to estimate the $\theta = \{\gamma, \Delta\}$. As in [49][50], we adopt an Expectation Maximization (EM) procedure where we treat $\tilde{\boldsymbol{X}}$ as the hidden data. In the E-step, we seek the expectation of the complete data $\left\{\boldsymbol{Y}, \tilde{\boldsymbol{X}}, \gamma, \boldsymbol{\Delta}\right\}$ log likelihood under the posterior $p\left(\tilde{X}|Y, \theta^t, \sigma^2\right)$, where θ^t denotes the estimate of θ at iteration t. Because $\left\{\tilde{x}^i\right\}_{i=1}^L$ are conditionally independent given Y, γ, and Δ, the E-step reduces to

$$Q(\theta, \theta^t) = \sum_{i=1}^{L} E_{\tilde{x}^i|y^i,\theta^t}\left[\log p\left(y^i, \tilde{x}^i, \theta\right)\right]. \tag{2.6}$$

The posterior needed to compute (2.6) is given by

$$p\left(\tilde{x}^i|y^i, \theta\right) = \mathsf{N}\left(\tilde{x}^i; \mu^i, \Sigma^i\right) \tag{2.7}$$

$$\Sigma^i = \Phi^i - \Phi^i \tilde{D}^T \left(\sigma^2 \mathsf{I} + \tilde{D}\Phi^i \tilde{D}^T\right)^{-1} \tilde{D}\Phi^i \tag{2.8}$$

$$\mu^i = \sigma^{-2}\Sigma^i \tilde{D}^T y^i \tag{2.9}$$

where Φ^i is a diagonal matrix with

$$\Phi^i[m,m] = \begin{cases} \gamma^i[k] & \text{if } m \in \mathscr{T}^k \\ \delta^i[m-M] & \text{else.} \end{cases} \tag{2.10}$$

Unlike [50], where the covariance of the posterior is shared for all i, the covariance is a function of i here because each $\tilde{x}^i$ consists of x^i, whose support does not vary with i, and e^i, whose support does vary.

In the M-step, $Q(\theta,\theta^t)$ is maximized with respect to θ, leading to the update rules:

$$\gamma[k] = \sum_{i=1}^{L} \sum_{m\in\mathscr{T}^k} \frac{\Sigma^i[m,m] + \left(\mu^i[m]\right)^2}{|\mathscr{T}^k|L} \tag{2.11}$$

$$\delta^i[n] = \Sigma^i[n+M,n+M] + \left(\mu^i[n+M]\right)^2, \;\; 1 \le n \le N \tag{2.12}$$

$$\sigma^2 = \sum_{i=1}^{L} \frac{\left\|y^i\right\|^2 - 2\left(y^i\right)^T \tilde{D}\mu^i + \mathsf{tr}\left(\tilde{D}^T\tilde{D}\left(\Sigma^i + \mu^i\left(\mu^i\right)^T\right)\right)}{Ln} \tag{2.13}$$

where $\mathsf{tr}(\cdot)$ refers to the trace operator.

Upon convergence, the EM algorithm produces an estimate of θ, denoted by $\hat{\theta}$. Given $\hat{\theta}$, X and E can be estimated using the maximum a-posteriori (MAP) estimator:

$$\begin{bmatrix} \hat{x}^i \\ \hat{e}^i \end{bmatrix} = \arg\max_{\tilde{x}^i} p\left(\tilde{x}^i | y^i, \hat{\theta}\right). \tag{2.14}$$

Since the posterior is Gaussian, the mean and mode are identical, such that the optimizer of (2.14) is given by (2.9).

2.3 Results

2.3.1 Synthetic Data Results

To validate the proposed method, we conducted SSR experiments on synthetic data. To generate the synthetic data, we begin by randomly selecting s sets from $\{\mathscr{T}^k\}_{k=1}^K$ and generate x^i such that the non-zero elements are indexed by one of the selected sets. We use equally sized blocks of length 8. The non-zero elements of x^i are drawn from the $\mathsf{N}(0,1)$ distribution. We generate $D \in \mathbb{R}^{80\times 160}$ by drawing its elements from the $\mathsf{N}(0,1)$ distribution and normalizing the columns to have unit ℓ_2 norm. Finally, we use the robust modeling strategy and replace D by

$\tilde{D} = \begin{bmatrix} D & \mathsf{I} \end{bmatrix}$. In order to simulate a noisy SSR scenario, we generate v^i by drawing its elements from the $\mathsf{N}(0,1)$ distribution and e^i by drawing its elements form the Student-t distribution with one degree of freedom. Finally, we generate observations y^i according to (2.1) after scaling v^i and e^i to achieve a specified Signal-to-Gaussian noise ratio (SGNR) and Signal-to-Outlier noise ratio (SONR).

Let $\hat{X}$ denote the approximation to X generated by the SSR algorithm. We measure the quality of the recovery using the relative ℓ_2 error:

$$\frac{\left\|X - \hat{X}\right\|_F^2}{\|X\|_F^2} \tag{2.15}$$

We performed the synthetic data experiment 500 times and report the average performance results.

We compare the performance of the proposed method to several standard SSR algorithms. As a baseline, we use the ℓ_1 SSR approach and the block sparse extension of the ℓ_1 approach, the $\ell_2 - \ell_1$ block SSR algorithm (also known as Group LASSO [54]), which seeks

$$\arg\min_x \|y - Dx\|_2^2 + \lambda \sum_{k=1}^{K} \left\| x[\mathscr{T}^k] \right\|_2. \tag{2.16}$$

Note that (2.16) reduces to the ℓ_1 SSR objective function when each element of x is a separate group. We use the SLEP [55] software package to solve the ℓ_1 and $\ell_2 - \ell_1$ problems. For comparison purposes we naively extend the ℓ_1 and $\ell_2 - \ell_1$ approaches to the MMV case by solving each MMV problem as L independent SMV problems.

We also compare our approach with Block-SBL (BSBL) [52, 44], which is a hierarchical Bayesian framework for solving the SMV block-sparse recovery problem. We naively extend BSBL to the MMV case by assuming that the outliers have stationary support, denoting the resulting algorithm as M-BSBL. In the context of the signal model in (2.1), M-BSBL corresponds to assuming *row sparsity* on E.

Table 2.1: Face classification accuracy results for $L = 1$ and $L = 5$ for various occlusion rates

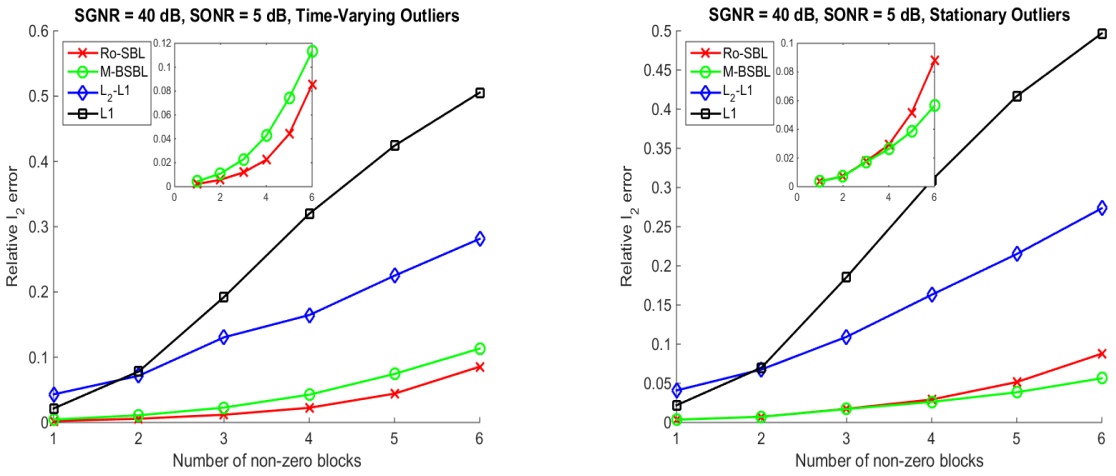
	10%		20%		30%		40%		50%	
	$L = 1$	$L = 5$	$L = 1$	$L = 5$	$L = 1$	$L = 5$	$L = 1$	$L = 5$	$L = 1$	$L = 5$
SRC [3]	89.72	100	83.61	100	71.29	97.81	54.24	92.54	35.97	67.98
$P_{\ell_2-\ell_1}$ [41]	91.60	100	84.50	100	69.90	96.93	54.08	93.42	39.07	72.37
M-BSBL [44]	—	100	—	100	—	99.12	—	94.30	—	76.75
Ro-SBL (proposed)	**91.92**	**100**	**91.11**	**100**	**83.52**	**100**	**65.58**	**100**	**41.44**	**91.22**

Simulation results for a 5 measurement SSR problem with 40 dB SGNR and 5 dB SONR are shown in Fig. 2.1a. M-BSBL and Ro-SBL drastically outperform the ℓ_1 and $\ell_2 - \ell_1$ SSR approaches, which shows that hierarchical Bayesian approaches outperform deterministic methods even in the challenging SSR setup considered. In addition, since the Bayesian approaches explicitly model the MMV nature of the problem, this result shows that significant improvements can be achieved by incorporating the prior knowledge that the support of x^i does not change with i in the SSR algorithm. We observe a $25\% - 51\%$ improvement in relative ℓ_2 error from Ro-SBL compared to M-BSBL for $s \leq 6$. This suggests that Ro-SBL is better able to capture outliers due to its superior model.

For illustrative purposes, we conducted the same synthetic data experiment, with the exception that the outliers were constrained to be the same for all i. The results are shown in Fig. 2.1b. As expected, the performance gains of the hierarchical Bayesian approaches over the deterministic approaches carries over into the stationary outlier scenario. It is important to note that M-BSBL slightly outperforms Ro-SBL because, in this case, the M-BSBL signal model is better fitted to estimate the outliers since M-BSBL assumes that the outliers are stationary and enjoys the advantages of MMV modeling, even for outliers.

2.3.2 FR Results

In this section, we present results demonstrating the efficacy of the proposed method in a FR task. We use the Extended Yale B Database [56], which consists of 2441 images of 38



(a) Time-Varying outliers

(b) Stationary outliers

Figure 2.1: Comparison of SSR algorithms on synthetic data

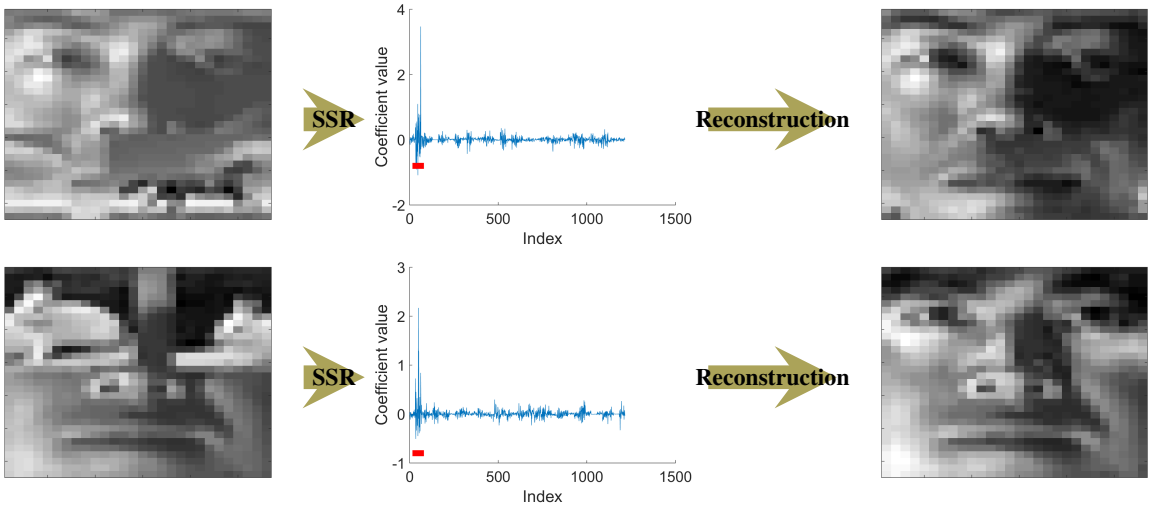


Figure 2.2: Example of face reconstruction using Ro-SBL. The red bar denotes coefficients corresponding to the true subject class.

subjects under various illumination conditions. Each 192×168 image is a frontal perspective of the subject's face and has been cropped such that only the face can be seen. We randomly split the database into training and test sets. Following the SRC framework, we downsample the images in the training set by $\frac{1}{12}$, vectorize the result, and concatenate the vectors to form the dictionary D:

$$D = \begin{bmatrix} D[:, \mathscr{T}^1] & D[:, \mathscr{T}^2] & \cdots & D[:, \mathscr{T}^{38}] \end{bmatrix}$$

where $D[:, \mathscr{T}^k]$ consists of training images from the k'th subject. Note that this automatically introduces a block structure in D. As before, we replace D by $\tilde{D}$.

In the testing phase, we seek to identify a given subject from L images of that subject's face. To simulate time-varying occlusions, we occlude each image by one of 10 animal images, choosing the location of the occluding image randomly. Given L observations, we use one of the SSR algorithms to estimate $\hat{X}$ and $\hat{E}$. Similar to the person re-identification classifier presented in [48], we label the test images using

$$k^* = \arg\min_k \sum_{i=1}^{L} \left\| y^i - D\left(\phi^k \odot \hat{x}^i\right) - \hat{e}^i \right\|_2^2$$

where $\phi^k[m] = 1 \; \forall m \in \mathscr{T}^k$ and $\phi^k[\mathscr{T}^{k'}] = 0 \; \forall k' \neq k$.

For each test subject, around 30 test images were available and the classification experiment was run $\left\lfloor \frac{30}{L} \right\rfloor$ times. We report averaged classification results in Table 2.1 for two cases: $L = 1$, i.e. an SMV FR problem considered in [42, 3, 41], and $L = 5$, i.e. an MMV problem considered in person re-identification [48]. Note that our proposed algorithm becomes equivalent to M-BSBL for $L = 1$, hence we do not report M-BSBL results for this case in Table 2.1. It is evident from the results that, for every algorithm, the $L = 5$ case leads to much better classification accuracy compared to $L = 1$, which corroborates the well known result that MMV modeling is superior to SMV in harsh conditions.

In all cases, Ro-SBL performs better than all other competing algorithms. For low

occlusion rates (10%), all of the competing algorithms perform comparably, despite the fact that only Ro-SBL explicitly models non-stationary occlusions. This result can be explained by the fact that 10% occlusion represents a minor fraction of the overall image area and a significant amount of facial features remain un-occluded. On the other hand, Ro-SBL drastically outperforms the other algorithms at high occlusion rates (50%) with $L = 5$, which can be attributed to the fact that a large portion of the facial features in each image are occluded and the SSR algorithm is forced to use all 5 measurements to jointly recover $\hat{X}$. Since Ro-SBL models outliers more accurately than the other SSR methods, it is better able to approximate the identity of the occluded subject.

Finally, as a visual example of how the proposed method performs face classification, we show that the occlusion can be removed from the test image by considering Dx^i as the estimate of the original, un-occluded test image. The face reconstruction result is shown in Fig. 2.2 (results are generated using $L = 5$ and a downsampling factor of $\frac{1}{6}$ was used for visualization purposes), which shows that the proposed method removes much of the occluding image and provides a relatively good reconstruction of the original face. Moreover, the coefficient plots show that the dominant coefficients all reside in the block corresponding to the test subject's index in D.

2.4 Conclusion

In this chapter, we have proposed a novel robust sparse recovery algorithm based on the well known SBL framework to recover simultaneous block sparse signals in presence of time varying outliers. Along with validating our method on synthetic data, we show the efficacy of our approach in simultaneous FR in the presence of time varying outliers.

Chapter 2, in full, is a reprint of material published in the article Igor Fedorov, Ritwik Giri, Bhaskar D. Rao, and Truong Q. Nguyen, "Robust Bayesian Method for Simultaneous Block Sparse Signal Recovery with Applications to Face Recognition," IEEE International Conference on Image Processing, 2016. I was the primary author and B. D. Rao supervised the research.

Chapter 3

A Unified Framework for Sparse Non-Negative Least Squares using Multiplicative Updates and the Non-Negative Matrix Factorization Problem

3.1 Introduction

Least squares problems occur naturally in numerous research and application settings. At a high level, given an observation $y \in \mathbb{R}^N$ of $x \in \mathbb{R}^M$ through a linear system $D \in \mathbb{R}^{N \times M}$, the least squares problem refers to

$$\arg\min_x \|y - Dx\|_2^2 . \tag{3.1}$$

Quite often, prior information about x is known. For instance, x may be known to be non-negative. Non-negative data occurs naturally in many applications, including text mining [57], image processing [58], speech enhancement [59], and spectral decomposition [60][61]. In this case, (3.1) is modified to

$$\arg\min_{x\geq 0} \|y - Dx\|_2^2 \tag{3.2}$$

where $x \geq 0$ refers to the elements of x being constrained to be non-negative and (3.2) is referred to as the non-negative least squares (NNLS) problem. A solution to (3.2) can be obtained using the well-known active set Lawson-Hanson algorithm [62] or one of its many variants [63]. In this work, we are interested in a specific flavor of NNLS problems where $M > N$. Under this constraint, the linear system in (3.2) is underdetermined and admits an infinite number of solutions. To constrain the set of possible solutions, a sparsity constraint on x can be added, leading to a sparse NNLS (S-NNLS) formulation:

$$\arg\min_{x\geq 0, \|x\|_0 \leq s} \|y - Dx\|_2^2 \tag{3.3}$$

where $\|\cdot\|_0$ refers to the ℓ_0 pseudo-norm, which counts the number of non-zero entries. Solving (3.3) directly is difficult because the ℓ_0 pseudo-norm is non-convex. In fact, solving (3.3) requires a combinatorial search and has been shown to be NP-hard [4]. Therefore, greedy methods have been adopted to approximate the solution [4, 64]. One effective approach, called reverse sparse NNLS (rsNNLS) [65], first finds an x such that $\|y - Dx\|_2^2 \leq \varepsilon$ using the active-set Lawson-Hanson algorithm and then prunes x with a greedy procedure until $\|x\|_0 \leq s$, all while maintaining $x \geq 0$. Other approaches include various relaxations of the ℓ_0 pseudo-norm in (3.3) using the ℓ_1 norm [66] or a combination of the ℓ_1 and ℓ_2 norms [67], leading to easier optimization problems.

The purpose of this work is to address the S-NNLS problem in a setting often encountered by practitioners, i.e. when several S-NNLS problems must be solved simultaneously. We are

primarily motivated by the problem of sparse non-negative matrix factorization (S-NMF). NMF falls under the category of dictionary learning algorithms. Dictionary learning is a common ingredient in many signal processing and machine learning algorithms [11, 68, 69, 18]. In NMF, the data, the dictionary, and the encoding of the data under the dictionary are all restricted to be non-negative. Constraining the encoding of the data to be non-negative leads to the intuitive interpretation of the data being decomposed into an additive combination of dictionary atoms [70, 71, 72]. More formally, let $Y \in \mathbb{R}_+^{N\times L}$ be a matrix representing the given data, where each column of Y, $y^i \in \mathbb{R}_+^N, 1 \leq i \leq L$, is a data vector. The goal of NMF is to decompose Y into two matrices $D \in \mathbb{R}_+^{N\times M}$ and $X \in \mathbb{R}_+^{M\times L}$. When $M < N$, NMF is often stated in terms of the optimization problem

$$\arg\min_{\theta\geq 0} \|Y - DX\|_F^2 \tag{3.4}$$

where $\theta = \{D, X\}$, D is called the dictionary, X is the encoding of the data under the dictionary, and $\theta \geq 0$ is short-hand for the elements of D and X being constrained to be non-negative. Optimizing (3.4) is difficult because it is not convex in θ [73]. Instead of performing joint optimization, a block coordinate descent method [74] is usually adopted where the algorithm alternates between holding D fixed while optimizing X and vice versa [70, 72, 73, 75, 76]:

$$\text{Update } D \text{ given } X \tag{3.5}$$

$$\text{Update } X \text{ given } D. \tag{3.6}$$

Note that (3.5) and (3.6) are a collection of N and L NNLS problems, respectively, which motivates the present work. The block coordinate descent method is advantageous because (3.5) and (3.6) are convex optimization problems for the objective function in (3.4), so that any number of techniques can be employed within each block. One of the most widely used optimization techniques, called the multiplicative update rules (MUR's), performs (3.5)-(3.6) using simple

element-wise operations on D and X [70, 72]:

$$D^{t+1} = D^t \odot \frac{YX^T}{D^t XX^T} \tag{3.7}$$

$$X^{t+1} = X^t \odot \frac{D^T Y}{D^T DX^t} \tag{3.8}$$

where $\odot$ denotes element-wise multiplication, A/B denotes element-wise division of matrices A and B, and t denotes the iteration index. The MUR's shown in (3.7)-(3.8) are guaranteed to not increase the objective function in (3.4) [70, 72] and, due to their simplicity, are widely used in the NMF community [77, 78, 79]. The popularity of NMF MUR's persists despite the fact that there is no guarantee that the sequence $\{D^t, X^t\}_{t=0}^{\infty}$ generated by (3.7)-(3.8) will converge to a local minimum [80] or even a stationary point [73, 80] of (3.4).

Unlike traditional NMF methods [70, 72], this work considers the scenario where D is overcomplete, i.e. $M \gg N$. Overcomplete dictionaries have much more flexibility to represent diverse signals [37] and, importantly, lead to effective sparse and low dimensional representations of the data [71, 37]. As in NNLS, the concept of sparsity has an important role in NMF because when X is overcomplete, (3.4) is not well-posed without some additional regularization. Sparsity constraints limit the set of possible solutions of (3.4) and, in some cases, lead to guarantees of uniqueness [81]. The S-NMF problem can be stated as the solution to

$$\underset{\theta \geq 0, \|X\|_0 \leq s}{\arg\min} \|Y - DX\|_F^2 \tag{3.9}$$

where $\|X\|_0 \leq s$ is shorthand for $\left\{ \left\|x^i\right\|_0 \leq s \right\}_{i=1}^{L}$. One classical approach to S-NMF relaxes the ℓ_0 constraint and appends a convex, sparsity promoting ℓ_1 penalty to the objective function [66]:

$$\underset{\theta \geq 0}{\arg\min} \|Y - DX\|_F^2 + \lambda \|X\|_1 \tag{3.10}$$

where $\|X\|_1$ is shorthand for $\sum_{i=1}^{L} \left\|x^i\right\|_1$. As shown in [66], (3.10) can be iteratively minimized

through a sequence of multiplicative updates where the update of D is given by (3.7) and the update of X is given by

$$X^{t+1} = X^t \odot \frac{D^T Y}{D^T D X^t + \lambda}. \tag{3.11}$$

We also consider an extension of S-NMF where a sparsity constraint is placed on D [67]

$$\underset{\theta \geq 0, \|X\|_0 \leq s_X, \|D\|_0 \leq s_D}{\arg\min} \|Y - DX\|_F^2 \tag{3.12}$$

which encourages basis vectors that explain localized features of the data [67]. We refer to (3.12) as S-NMF-D.

The motivation of this work is to develop a maximum a-posteriori (MAP) estimation framework to address the S-NNLS and S-NMF problems. We build upon the seminal work in [49] on Sparse Bayesian Learning (SBL). The SBL framework places a sparsity-promoting prior on the data [27] and has been shown to give rise to many models used in the compressed sensing literature [82]. It will be shown that the proposed framework provides a general class of algorithms that can be tailored to the specific needs of the user. Moreover, inference can be done through a simple MUR for the general model considered and the resulting S-NNLS algorithms admit convergence guarantees.

The key contribution of this work is to detail a unifying framework that encompasses a large number of existing S-NNLS and S-NMF approaches. Therefore, due to the very nature of the framework, many of the algorithms presented in this work are not new. Nevertheless, there is value in the knowledge that many of the algorithms employed by researchers in the S-NNLS and S-NMF fields are actually members of the proposed family of algorithms. In addition, the proposed framework makes the process of formulating novel task-specific algorithms easy. Finally, the theoretical analysis of the proposed framework applies to any member of the family of proposed algorithms. Such an analysis has value to both existing S-NNLS and S-NMF approaches

like [83, 84], which do not perform such an analysis, as well as to any future approaches which fall under the umbrella of the proposed framework. It should be noted that several authors have proposed novel sets of MUR's with provable convergence guarantees for the NMF problem in (3.4) [85] and S-NMF problem in (3.10) [86]. In contrast to [86], the proposed framework does not use the ℓ_1 regularization function to solve (3.9). In addition, since the proposed framework encompasses the update rules used in existing works, the analysis presented here applies to works from existing literature, including [83, 84].

3.1.1 Contributions of the Paper

- A general class of rectified sparsity promoting priors is presented and it is shown that the computational burden of the resulting inference procedure is handled by a class of simple, low-complexity MUR's.
- A monotonicity guarantee for the proposed class of MUR's is provided, justifying their use in S-NNLS and S-NMF algorithms.
- A convergence guarantee for the proposed class of S-NNLS and S-NMF-D algorithms is provided.

3.1.2 Notation

Bold symbols are used to denote random variables and plain font to denote a particular realization of a random variable. MATLAB notation is used to denote the (i, j)'th element of the matrix X as $x^i[m]$ and the i'th column of X as x^i. We use X^t to denote the matrix X at iteration t of a given algorithm and $(X)^z$ to denote the matrix X with each element raised to the power z.

Table 3.1: Distributions used throughout this work, where $\exp(a) = e^a$.

Distribution	pdf
Rectified Gaussian	$p^{\mathsf{RG}}(x;\gamma) = \sqrt{\frac{2}{\pi\gamma}}\exp\left(-\frac{x^2}{2\gamma}\right)u(x)$
Exponential	$p^{\mathsf{Exp}}(x;\gamma) = \gamma\exp(-\gamma x)\,u(x)$
Inverse Gamma	$p^{\mathsf{IGa}}(x;a,b) = \frac{b^a}{\Gamma(a)}x^{-a-1}\exp\left(-\frac{b}{x}\right)u(x)$
Gamma	$p^{\mathsf{Ga}}(x;a,b) = \frac{1}{\Gamma(a)b^a}x^{a-1}\exp(-xb)\,u(x)$
Rectified Student's-t	$p^{\mathsf{RST}}(x;\tau) = \frac{2\Gamma\left(\frac{\tau+1}{2}\right)}{\sqrt{\tau\pi}\Gamma\left(\frac{\tau}{2}\right)}\left(1+\frac{x^2}{\tau}\right)^{-\frac{(\tau+1)}{2}}u(x)$
Rectified Generalized Double Pareto	$p^{\mathsf{RGDP}}(x;a,b,\tau) = 2\eta\left(1+\frac{x^b}{\tau a^b}\right)^{-\left(\tau+\frac{1}{b}\right)}u(x)$

3.2 Sparse Non-Negative Least Squares Framework Specification

The S-NNLS signal model is given by

$$\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X} + \boldsymbol{V} \tag{3.13}$$

where the columns of $\boldsymbol{V}$, the noise matrix, follow a $\mathsf{N}(0,\sigma^2\mathsf{I})$ distribution. To complete the model, a prior on the columns of $\boldsymbol{X}$, which are assumed to be independent and identically distributed, must be specified. This work considers separable priors of the form

$$p\left(x^i\right) = \prod_{m=1}^{M} p\left(x^i[m]\right) \tag{3.14}$$

where $p\left(x^i[m]\right)$ has a scale mixture representation [87, 88]:

$$p\left(x^i[m]\right) = \int_0^\infty p\left(x^i[m]|\gamma^i[m]\right)p\left(\gamma^i[m]\right)d\gamma^i[m]. \tag{3.15}$$

Separable priors are considered because, in the absence of prior knowledge, it is reasonable to assume independence amongst the coefficients of X. The case where dependencies amongst

the coefficients exist is considered in Section 3.5. The proposed framework extends the work on power exponential scale mixtures [7, 89] to rectified priors and uses the Rectified Power Exponential (RPE) distribution for the conditional density of $\boldsymbol{x^i}[\boldsymbol{m}]$ given $\boldsymbol{\gamma^i}[\boldsymbol{m}]$:

$$p^{\mathsf{RPE}}\left(x^i[m]|\gamma^i[m];z\right) = \frac{ze^{-\left(\frac{x^i[m]}{\gamma^i[m]}\right)^z}}{\gamma^i[m]\Gamma\left(\frac{1}{z}\right)}u\left(x^i[m]\right)$$

where $u(\cdot)$ is the unit-step function, $0 < z \leq 2$, and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$. The RPE distribution is chosen for its flexibility. In this context, (3.15) is referred to as a rectified power exponential scale mixture (RPESM).

The advantage of the scale mixture prior is that it introduces a Markovian structure of the form

$$\boldsymbol{\gamma^i} \rightarrow \boldsymbol{x^i} \rightarrow \boldsymbol{y^i} \tag{3.16}$$

and inference can be done in either the $\boldsymbol{x}$ or $\boldsymbol{\gamma^i}$ domains. This work focuses on doing MAP inference in the $\boldsymbol{x}$ domain, which is also known as Type 1 inference, whereas inference in the $\boldsymbol{\gamma}$ domain is referred to as Type 2. The scale mixture representation is flexible enough to represent most heavy-tailed densities [24, 20, 90, 23, 22], which are known to be the best sparsity promoting priors [49, 50]. One reason for the use of heavy-tailed priors is that they are able to model both the sparsity and large non-zero entries of $\boldsymbol{x}$.

The RPE encompasses many rectified distributions of interest. For instance, the RPE reduces to a Rectified Gaussian by setting $z = 2$, which is a popular prior for modeling non-negative data [91, 88] and results in a Rectified Gaussian Scale Mixture in (3.15). Setting $z = 1$ corresponds to an Exponential distribution and leads to an Exponential Scale Mixture in (3.15) [92]. Table 3.2 shows that many rectified sparse priors of interest can be represented as a RPESM. Distributions of interest are summarized in Table 3.1.

Table 3.2: RPESM representation of rectified sparse priors.

z	$p\left(\gamma^i[m]\right)$	$p\left(x^i[m]\right)$
2	$p^{\mathsf{Exp}}\left(\gamma^i[m];\frac{\tau^2}{2}\right)$	$p^{\mathsf{Exp}}\left(x^i[m];\tau\right)$
2	$p^{\mathsf{IGa}}\left(\gamma^i[m];\frac{\tau}{2},\frac{\tau}{2}\right)$	$p^{\mathsf{RST}}\left(x^i[m];\tau\right)$
1	$p^{\mathsf{Ga}}\left(\gamma^i[m];\tau,\tau\right)$	$p^{\mathsf{RGDP}}\left(x^i[m];1,1,\tau\right)$

3.3 Unified MAP Inference Procedure

In the MAP framework, X is directly estimated from Y by minimizing

$$L(X) = -\log\left(\prod_{i=1}^{L} p\left(x^i|y^i\right)\right). \tag{3.17}$$

We have made the dependence of the negative log-likelihood on X and D implicit for brevity. Minimizing (3.17) in closed form is intractable for most priors, so the proposed framework resorts to an Expectation-Maximization (EM) approach [22]. In the E-step, the expectation of the negative complete data log-likelihood with respect to the distribution of γ, conditioned on the remaining variables, is formed:

$$Q\left(X,\bar{X}^t\right) \doteq \|Y-DX\|_F^2 + \lambda\left(\sum_{i=1,m=1}^{L,M}\left(x^i[m]\right)^z\left\langle\frac{1}{\left(\gamma^i[m]\right)^z}\right\rangle - \log u\left(x^i[m]\right)\right) \tag{3.18}$$

where $\langle\cdot\rangle$ refers to the expectation with respect to the density $p\left(\gamma^i[m]|\bar{x}^i[m]\right)$, t refers to the iteration index, $\bar{X}^t$ denotes the estimate of X at the t'th EM iteration, and $\doteq$ refers to dropping terms that do not influence the M-step and scaling by $\lambda = 2\sigma^2$. The last term in (3.18) acts as a barrier function against negative values of X. The function $Q\left(X,\bar{X}^t\right)$ is separable in the columns of X. In an abuse of notation, we use $Q\left(x^i,\bar{x}^{i,t}\right)$ to refer to the dependency of $Q\left(X,\bar{X}^t\right)$ on x^i.

In order to compute the expectation in (3.18), a similar method to the one used in [7, 24] is employed, with some minor adjustments due to non-negativity constraints. Let $p\left(x^i[m]\right) = p^R\left(x^i[m]\right)u\left(x^i[m]\right)$, where $p^R\left(x^i[m]\right)$ is the portion of $p\left(x^i[m]\right)$ that does not in-

clude the rectification term, and let $p^R\left(x^i[m]\right)$ be differentiable on $[0,\infty)$. Then,

$$\left\langle \frac{1}{(\gamma^i[m])^z} \right\rangle = -\frac{\partial \log p^R\left(\bar{x}^{i,t}[m]\right)}{\partial \bar{x}^{i,t}[m]} \frac{1}{z\left(\bar{x}^{i,t}[m]\right)^{z-1}}. \tag{3.19}$$

Turning to the M-step, the proposed approach employs the Generalized EM (GEM) M-step [22]:

$$\text{Choose } \bar{X}^{t+1} \text{ such that } Q(\bar{X}^{t+1},\bar{X}^t) \leq Q(\bar{X}^t,\bar{X}^t). \qquad \text{(GEM M-step)}$$

In particular, $Q(X,\bar{X}^t)$ is minimized through an iterative gradient descent procedure. As with any gradient descent approach, selection of the learning rate is critical in order to ensure that the objective function is decreased and the problem constraints are met. Following [72, 70], the learning rate is selected such that the gradient descent update is guaranteed to generate non-negative updates and can be implemented as a low-complexity MUR, given by

$$X^{s+1} = X^s \odot \frac{D^T Y}{D^T D X^s + \lambda \Omega^t \odot (X^s)^{z-1}} \tag{3.20}$$

$$\Omega^t[m,i] = -\frac{1}{\left(\bar{x}^{i,t}[m]\right)^{z-1}} \frac{\partial \log p^R\left(\bar{x}^{i,t}[m]\right)}{\partial \bar{x}^{i,t}[m]} \tag{3.21}$$

where s denotes the gradient descent iteration index (not to be confused with the EM iteration index t). The resulting S-NNLS algorithm is summarized in Algorithm 1, where ζ denotes the specific MUR used to update X, which is (3.20) in this case.

3.3.1 Extension to S-NMF

We now turn to the extension of our framework to the S-NMF problem. As before, the signal model in (3.13) is used as well as the RPESM prior on X. To estimate D and X, the

Algorithm 1 S-NNLS Algorithm

Require: $Y, D, \bar{X}^0, \lambda, \zeta, S, t^\infty$
  Initialize $t = 0, \mathscr{Z} = \{(m,i)\}_{m=1,i=1}^{M,L}$
  while $\mathscr{Z} \neq \emptyset$ **do**
    Form Ω^t and initialize $X^1 = \bar{X}^t$, $\mathscr{J} = \mathscr{Z}$
    for $s = 1$ to S **do**
      Generate $x^{i,s+1}[m]$ using update rule ζ for $(m,i) \in \mathscr{J}$
      Set $x^{i,s+1}[m] = x^{i,s}[m]$ for any $(m,i) \notin \mathscr{J}$
      $\mathscr{J} \leftarrow \mathscr{J} \setminus \left\{(m,i) : x^{i,s+1}[m] = 0 \text{ or } x^{i,s+1}[m] = x^{i,s}[m]\right\}$
    end for
    Set $\bar{X}^{t+1} = X^{S+1}$ and $\mathscr{Z} \leftarrow \mathscr{Z} \setminus \left\{(m,i) : \bar{x}^{i,t+1}[m] = x^{i,t}[m] \text{ or } \bar{x}^{i,t+1} = 0\right\}$
    $t \leftarrow t + 1$
    if $t = t^\infty$ **then**
      Break
    end if
  end while
  Return $\bar{X}^t$

proposed framework seeks to find

$$\arg\min_{\theta} L^{NMF}(\theta), \;\; L^{NMF}(\theta) = -\log p(D, X|Y) \tag{3.22}$$

where $\theta = \{D, X\}$. The random variables $\boldsymbol{D}$ and $\boldsymbol{X}$ are assumed independent and a non-informative prior over the positive orthant is placed on $\boldsymbol{D}$ for S-NMF. For S-NMF-D, a separable prior from the RPESM family is assumed for $\boldsymbol{D}$. In order to solve (3.22), the block-coordinate descent optimization approach in (3.5)-(3.6) is employed. For each one of (3.5) and (3.6), the GEM procedure described above is used.

The complete S-NMF/S-NMF-D algorithm is given in Algorithm 2. Due to the symmetry between (3.5) and (3.6) and to avoid unnecessary repetition, heavy use of Algorithm 1 in Algorithm 2 is made. Note that $\zeta_X =$ (3.20), $\zeta_D =$ (3.8) for S-NMF, and $\zeta_D =$ (3.20) for S-NMF-D.

Algorithm 2 S-NMF/S-NMF-D Algorithm

Require: $Y, \lambda, S, \zeta_D, \zeta_X, t^\infty$
  Initialize $d^{i,0}[n] = 1, x^{i,0}[m] = 1 \ \ \forall n, m, t = 0,$
  while $t \neq t^\infty$ and $\left(\bar{X}^{t+1} \neq \bar{X}^t \text{ or } \bar{D}^{t+1} \neq \bar{D}^t\right)$ **do**
    $\bar{D}^{t+1} = \left(\text{Algorithm1}(Y^T, (\bar{X}^t)^T, (\bar{D}^t)^T, \lambda, \zeta_D, S, 1)\right)^T$
    $\bar{X}^{t+1} = \text{Algorithm1}(Y, \bar{D}^{t+1}, \bar{X}^t, \lambda, \zeta_X, S, 1)$
    $t \leftarrow t + 1$
  end while

3.4 Examples of S-NNLS and S-NMF Algorithms

In the following, evidence of the utility of the proposed framework is provided by detailing several specific algorithms which naturally arise from (3.20) with different choices of prior. It will be shown that the algorithms described in this section are equivalent to well-known S-NNLS and S-NMF algorithms, but derived in a completely novel way using the RPESM prior. The S-NMF-D algorithms described are, to the best of our knowledge, novel. In Section 3.5, it will be shown that the proposed framework can be easily used to define novel algorithms where block-sparsity is enforced.

3.4.1 Reweighted l_2

Consider the prior $\boldsymbol{x^i}[\boldsymbol{m}] \sim p^{\mathsf{RST}}\left(x^i[m]; \tau\right)$. Given this prior, (3.20) becomes

$$X^{s+1} = X^s \odot \frac{D^T Y}{D^T D X^s + \frac{2\lambda(\tau+1)X^s}{\tau + (\bar{X}^t)^2}}. \tag{3.23}$$

Given this choice of prior on $x^i[m]$ and a non-informative prior on $w^i[n]$, it can be shown that $L^{NMF}(\theta)$ reduces to

$$\|Y - DX\|_F^2 + \tilde{\lambda} \sum_{m=1,i=1}^{M,L} \log\left(\left(x^i[m]\right)^2 + \tau\right) \tag{3.24}$$

over $X \in \mathbb{R}_+^{M\times L}$ and $D \in \mathbb{R}_+^{N\times M}$ (i.e. the $\log u(\cdot)$ terms have been omitted for brevity), where $\tilde{\lambda} = 2\sigma^2(\tau+1)$. The sparsity-promoting regularization term in (3.24) was first studied in [9] in the context of vector sparse coding (i.e. without non-negativity constraints). Majorizing the sparsity promoting term in (3.24), it can be shown that (3.24) is upper-bounded by

$$\|Y - DX\|_2^2 + \tilde{\lambda} \left\| \frac{X}{E^t} \right\|_F^2 \tag{3.25}$$

where $e^{i,t}[m] = \bar{x}^{i,t}[m] + \tau$. Note that this objective function was also used in [93], although it was optimized using a heuristic approach based on the Moore-Penrose pseudoinverse operator. Letting

$$R = X/E^t \tag{3.26}$$

and $\tilde{\lambda} \to 0$, (3.25) becomes

$$\left\|Y - D\left(E^t \odot R\right)\right\|_2^2 \tag{3.27}$$

which is exactly the objective function that is iteratively minimized in the NUIRLS algorithm [84] if we let $\tau \to 0$. Although [84] gives a MUR for minimizing (3.27), the MUR can only be applied for each column of X individually. It is not clear why the authors of [84] did not give a matrix based update rule for minimizing (3.27), which can be written as

$$R^{s+1} = R^s \odot \frac{D^T Y}{D^T D\left(E^t \odot R^s\right)}.$$

This MUR is identical to (3.23) in the setting $\lambda, \tau \to 0$. Although [84] makes the claim that NUIRLS converges to a local minimum of (3.27), this claim is not proved. Moreover, nothing is said regarding convergence with respect to the actual objective function being minimized

(i.e. (3.24) as opposed to the majorizing function in (3.27)). As the analysis in Section 3.6 will reveal, using the update rule in (3.23) within Algorithm 1, the iterates are guaranteed to converge to a stationary point of (3.24). We make no claims regarding convergence with respect to the majorizing function in (3.25) or (3.27).

3.4.2 Reweighted ℓ_1

Assuming $\boldsymbol{x^i}[\boldsymbol{m}] \sim p^{\mathsf{RGDP}}\left(x^i[m]; 1, 1, \tau\right)$, (3.20) reduces to

$$X^{s+1} = X^s \odot \frac{D^T Y}{D^T D X^s + \frac{\lambda(\tau+1)}{\tau+\bar{X}^t}}. \tag{3.28}$$

Plugging the RGDP prior into (3.22) and assuming a non-informative prior on $w^i[n]$ leads to the Lagrangian of the objective function considered in [8] for unconstrained vector sparse coding (after omitting the barrier function terms):

$$\|Y - DX\|_F^2 + \tilde{\lambda} \sum_{m=1,i=1}^{M,L} \log\left(x^i[m] + \tau\right). \tag{3.29}$$

Interestingly, this objective function is a special case of the block sparse objective considered in [83] (where the Itakura-Saito reconstruction loss is used instead of the Frobenius norm loss) if each $x^i[m]$ is considered a separate block. The authors of [83] did not offer a convergence analysis of their algorithm, in contrast with the present work. To the best of our knowledge, the reweighted ℓ_1 formulation has not been considered in the S-NNLS literature.

3.4.3 Reweighted ℓ_2 and Reweighted ℓ_1 for S-NMF-D

Using the reweighted ℓ_2 or reweighted ℓ_1 formulations to promote sparsity in D is straightforward in the proposed framework and involves setting ζ_D to (3.23) or (3.28), respectively, in Algorithm 2.

3.5 Extension to Block Sparsity

As a natural extension of the proposed framework, we now consider the block sparse S-NNLS problem. This section will focus on the S-NNLS context only because the extension to S-NMF is straightforward. Block sparsity arises naturally in many contexts, including speech processing [77, 94], image denoising [95], and system identification [96]. The central idea behind block-sparsity is that D is assumed to be divided into disjoint blocks and each y^i is assumed to be a linear combination of the elements of a *small number of blocks*. This constraint can be easily accommodated by changing the prior on x^i to a block rectified power exponential scale mixture:

$$p\left(x^i\right) = \prod_{k=1}^{K} \underbrace{\int_0^\infty \prod_{m \in \mathscr{T}^k} p\left(x^i[m] | \gamma^i[k]\right) p\left(\gamma^i[k]\right) d\gamma^i[k]}_{p\left(x^i\left[\mathscr{T}^k\right]\right)} \tag{3.30}$$

(3.31)

where

$$\cup_{k=1}^{K} \mathscr{T}^k = [M] \text{ and } \mathscr{T}^k \cap \mathscr{T}^{k'} = \emptyset \; \forall k' \neq k \tag{3.32}$$

and $x^i\left[\mathscr{T}^k\right]$ is a vector consisting of the elements of x^i whose indices are members of $\mathscr{T}^k$. To find the MAP estimate of X given Y, the same GEM procedure as before is employed, with the exception that the computation of the weights in (3.18) is modified to:

$$\left\langle \frac{1}{\left(\gamma^i[k]\right)^z} \right\rangle = -\frac{\partial \log p^R\left(\bar{x}^i\left[\mathscr{T}^k\right]\right)}{\partial \bar{x}^i[m]} \frac{1}{z\left(\bar{x}^i[m]\right)^{z-1}}$$

where $m \in \mathscr{T}^k$. It can be shown that the MUR for minimizing $Q(X, \bar{X}^t)$ in (3.20) can be modified to account for the block prior in (3.30) to

$$X^{s+1} = X^s \odot \frac{D^T Y}{D^T D X^s + \lambda \Phi^t \odot (X^s)^{z-1}} \tag{3.33}$$

$$\phi^{i,t}[m] = -\frac{1}{(\bar{x}^{i,t}[m])^{z-1}} \frac{\partial \log p^R\left(\bar{x}^{i,t}\left[\mathscr{T}^k\right]\right)}{\partial \bar{x}^{i,t}[m]} \text{ for any } m \in T^k. \tag{3.34}$$

Next, we show examples of block S-NNLS algorithms that arise from our framework.

3.5.1 Example: Reweighted ℓ_2 Block S-NNLS

Consider the block-sparse prior in (3.30), where $p\left(x^i[m]|\gamma^i[k]\right)$, $m \in \mathscr{T}^k$, is a RPE with $z = 2$ and $\gamma^i[k] \sim p^{\mathsf{IGa}}\left(\gamma^i[k]; \tau/2, \tau/2\right)$. The resulting density $p\left(x^i\right)$ is a block RST (BRST) distribution:

$$p\left(x^i\right) = \left(\prod_{k=1}^{K} \frac{2\Gamma\left(\frac{\tau+1}{2}\right)}{\sqrt{\pi\tau}\Gamma\left(\frac{\tau}{2}\right)} \left(1 + \frac{\left\|x^i\left[\mathscr{T}^k\right]\right\|_2^2}{\tau}\right)^{-\frac{(\tau+1)}{2}} \right) \prod_{m=1}^{M} u\left(x^i[m]\right).$$

The MUR for minimizing $Q(X, \bar{X}^t)$ under the BRST prior is given by:

$$X^{s+1} = X^s \odot \frac{D^T Y}{D^T D X^s + \frac{2\lambda(\tau+1)X^s}{\tau + S^t}} \tag{3.35}$$

where $s^{i,t}[m] = \left\|\bar{x}^{i,t}\left[\mathscr{T}^k\right]\right\|_2^2$ for all $m \in \mathscr{T}^k$.

3.5.2 Example: Reweighted ℓ_1 Block S-NNLS

Consider the block-sparse prior in (3.30), where $p\left(x^i[m]|\gamma^i[k]\right)$, $m \in T^k$, is a RPE with $z = 1$ and $\gamma^i[k] \sim p^{\mathsf{Ga}}\left(\gamma^i[k]; \tau, \tau\right)$. The resulting density $p\left(x^i\right)$ is a block rectified generalized

double pareto (BRGDP) distribution:

$$p\left(x^i\right) = \left(\prod_{k=1}^{K} 2\eta \left(1 + \frac{\left\|x^i\left[\mathscr{T}^k\right]\right\|_1}{\tau}\right)^{-(\tau+1)}\right) \prod_{m=1}^{M} u\left(x^i[m]\right).$$

The MUR for minimizing $Q(X, \bar{X}^t)$ under the BRGDP prior is given by:

$$X^{s+1} = X^s \odot \frac{D^T Y}{D^T D X^s + \frac{\lambda(\tau+1)}{\tau + V^t}} \tag{3.36}$$

where $v^{i,t}[m] = \left\|\bar{x}^{i,t}\left[\mathscr{T}^k\right]\right\|_1$ for all $m \in T^k$.

3.5.3 Relation To Existing Block Sparse Approaches

Block sparse coding algorithms are generally characterized by their block-sparsity measure. The analog of the ℓ_0 sparsity measure for block-sparsity is the $\ell_2 - \ell_0$ measure

$$\sum_{k=1}^{K} 1_{\left\|x^i\left[\mathscr{T}^k\right]\right\|_2 > 0}, \tag{3.37}$$

which simply counts the number of blocks with non-zero energy. This sparsity measure has been studied in the past and block versions of the popular MP and OMP algorithms have been extended to Block-MP (BMP) and Block-OMP (BOMP) [38]. Extending BOMP to non-negative BOMP (NNBOMP) is straightforward, but details are omitted due to space considerations. One commonly used block sparsity measure in the NMF literature is the $\log - \ell_1$ measure [83]:

$$\sum_{k=1}^{K} \log\left(\left\|x^i\left[\mathscr{T}^k\right]\right\|_1 + \tau\right). \tag{3.38}$$

This sparsity measure arises naturally in the proposed S-NNLS framework when the BRGDP prior is plugged into (3.17). We are not aware of any existing algorithms which use the sparsity

measure induced by the BRST prior:

$$\sum_{k=1}^{K} \log\left(\left\|x^i\left[\mathscr{T}^k\right]\right\|_2^2 + \tau\right). \tag{3.39}$$

3.6 Analysis

In this section, important properties of the proposed framework are analyzed. First, the properties of the framework as it applies to S-NNLS are studied. Then, the proposed framework is studied in the context of S-NMF and S-NMF-D.

3.6.1 Analysis in the S-NNLS Setting

We begin by confirming that (GEM M-step) does not have a trivial solution at $x^i[m] = \infty$ for any (m,i) because

$$\left\langle \left(\gamma^i[m]\right)^{-z} \right\rangle \geq 0, \tag{3.40}$$

since it is an expectation of a non-negative random variable. In the following discussion, it will be useful to work with distributions whose functional dependence on $x^i[m]$ has a power function form:

$$f\left(x^i[m], z, \tau, \alpha\right) = \left(\tau + \left(x^i[m]\right)^z\right)^{-\alpha} \tag{3.41}$$

where $\tau, \alpha > 0$ and $0 < z \leq 2$. Note that the priors considered in this work have a power function form.

Monotonicity of $Q(X,\bar{X}^t)$ under (3.20)

The following theorem states one of the main contributions of this work, validating the use of (3.20) in (GEM M-step).

Theorem 1. *Let $z \in \{1,2\}$ and the functional dependence of $p^R\left(x^i[m]\right)$ on $x^i[m]$ have a power function form. Consider using the update rule stated in* (3.20) *to update $x^{i,s}[m]$ for all $(m,i) \in \mathscr{J} = \{(m,i) : x^{i,s}[m] > 0\}$. Then, the update rule in* (3.20) *is well defined and $Q(X^{s+1},\bar{X}^t) \leq Q(X^s,\bar{X}^t)$.*

Proof. Proof provided in 3.9.1. ■

Theorem 1 also applies to the block-sparse MUR in (3.33).

Local Minima of $L(X)$

Before proceeding to the analysis of the convergence of Algorithm 1, it is important to consider the question as to whether the local minima of $L(X)$ are desirable solutions from the standpoint of being sparse.

Theorem 2. *Let X^* be a local minimum of* (3.17) *and let the functional dependence of $p^R\left(x^i[m]\right)$ on $x^i[m]$ have a power function form. In addition, let one of the following conditions be satisfied: 1) $z \leq 1$ or 2) $z > 1$ and $\tau \to 0$. Then, $\left\|x^{i,*}\right\|_0 \leq N$.*

Proof. Proof provided in 3.9.2. ■

Convergence of Algorithm 1

First, an important property of the cost function in (3.17) can be established.

Theorem 3. *The function $-\log p\left(x^i[m]\right)$ is coercive for any member of the RPESM family.*

Proof. The proof is provided in 3.9.3. ■

Theorem 3 can then be used to establish the following corollary.

Corollary 1. *Assume the signal model in* (3.13) *and let* $p\left(x^i[m]\right)$ *be a member of the RPESM family. Then, the cost function* $L(X)$ *in* (3.17) *is coercive.*

Proof. This follows from the fact that $\|Y - DX\|_F^2 \geq 0$ and the fact that $-\log p\left(x^i[m]\right)$ is coercive due to Theorem 3. ■

The coercive property of the cost function in (3.17) allows us to establish the following result concerning Algorithm 1.

Corollary 2. *Let* $z \in \{1,2\}$ *and the functional dependence of* $p^R\left(x^i[m]\right)$ *on* $x^i[m]$ *have a power function form. Then, the sequence* $\{\bar{X}^t\}_{t=1}^{\infty}$ *produced by Algorithm 1 with S, the number of inner loop iterations, set to* 1 *admits at least one limit point.*

Proof. The proof is provided in 3.9.4. ■

We are now in a position to state one of the main contributions of this paper regarding the convergence of Algorithm 1 to the set of stationary points of (3.17). A stationary point is defined to be any point satisfying the Karush-Kuhn-Tucker (KKT) conditions for a given optimization problem [97].

Theorem 4. *Let* $z \in \{1,2\}$, $\zeta =$ (3.20), $t^{\infty} = \infty$, $S = 1$, *the functional dependence of* $p^R\left(x^i[m]\right)$ *on* $x^i[m]$ *have a power function form, the columns of D and Y have bounded norm, and D be full rank. In addition, let one of the following conditions be satisfied: (1)* $z = 1$ *and* $\tau \leq \lambda / \max_{m,i}\left(D^T Y\right)[m,i]$ *or (2)* $z = 2$ *and* $\tau \to 0$. *Then the sequence* $\{\bar{X}^t\}_{t=1}^{\infty}$ *produced by Algorithm 1 is guaranteed to converge to the set of stationary points of* $L(X)$. *Moreover,* $\{L(\bar{X}^t)\}_{t=1}^{\infty}$ *converges monotonically to* $L(\bar{X}^*)$, *for stationary point* $\bar{X}^*$.

Proof. The proof is provided in 3.9.5. ■

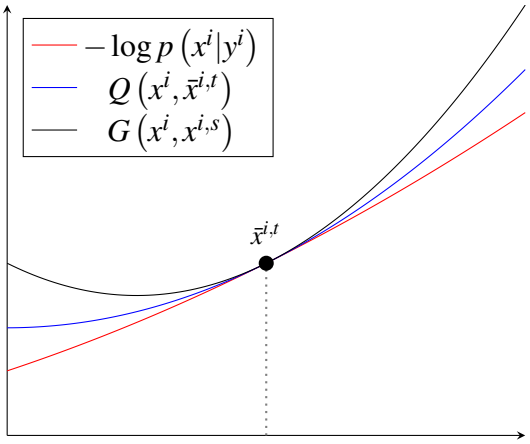


Figure 3.1: Visualization of Algorithm 1

The reason that $S = 1$ is specified in Theorem 4 is that it allows for providing convergence guarantees for Algorithm 1 without needing any convergence properties of the sequence generated by (3.20). Theorem 4 also applies to Algorithm 1 when the block-sparse MUR in (3.33) is used. To see the intuition behind the proof of Theorem 4 (given in 3.9.5), consider the visualization of Algorithm 1 shown in Fig. 3.1. The proposed framework seeks a minimum of $-\log p\left(x^i|y^i\right)$, for all i, through an iterative optimization procedure. At each iteration, $-\log p\left(x^i|y^i\right)$ is bounded by the auxiliary function $Q\left(x^i,\bar{x}^{i,t}\right)$ [97][22]. This auxiliary function is then bounded by another auxiliary function, $G\left(x^i,x^{i,s}\right)$, defined in (3.44). Therefore, the proof proceeds by giving conditions under which (GEM M-step) is guaranteed to reach a stationary point of $-\log p\left(x^i|y^i\right)$ by repeated minimization of $Q\left(x^i,\bar{x}^{i,t}\right)$ and then finding conditions under which $Q\left(x^i,\bar{x}^{i,t}\right)$ can be minimized by minimization of $G\left(x^i,x^{i,s}\right)$ through the use of (3.20).

3.6.2 Analysis in S-NMF and S-NMF-D Settings

We now extend the results of Section 3.6.1 to the case where D is unknown and is estimated using Algorithm 2. For clarity, let (z_D,τ_D) and (z_X,τ_X) refer to the distributional parameters of the priors over $\boldsymbol{D}$ and $\boldsymbol{X}$, respectively. As before, $\tau_D,\tau_X > 0$ and $0 < z_D,z_X \leq 2$. First, it is confirmed that Algorithm 2 exhibits the same desirable optimization properties as the

NMF MUR's (3.7)-(3.8).

Corollary 3. *Let $z_D, z_X \in \{1,2\}$ and the functional dependence of $p^R\left(x^i[m]\right)$ on $x^i[m]$ have a power function form. If performing S-NMF-D, let the functional dependence of $p^R\left(d^i[n]\right)$ on $d^i[n]$ have a power function form. Consider using Algorithm 2 to generate $\{\bar{D}^t, \bar{X}^t\}_{t=0}^{\infty}$. Then, the update rules used in Algorithm 2 are well defined and $L^{NMF}\left(\bar{D}^{t+1}, \bar{X}^{t+1}\right) \leq L^{NMF}\left(\bar{D}^t, \bar{X}^t\right)$.*

Proof. The proof is shown in 3.9.6. ■

Therefore, the proposed S-NMF framework maintains the monotonicity property of the original NMF MUR's, with the added benefit of promoting sparsity in X (and D, in the case of S-NMF-D).

Unfortunately, it is not clear how to obtain a result like Theorem 4 for Algorithm 2 in the S-NMF setting. The reason that such a result cannot be shown is because it is not clear that if a limit point, $(\bar{D}^{\infty}, \bar{X}^{\infty})$, of Algorithm 2 exists, that this point is a stationary point of $L^{NMF}(\cdot,\cdot)$. Specifically, if there exists (n,i) such that $\bar{D}^{\infty}[n,i] = 0$, the KKT condition $-\left(Y - \bar{D}^{\infty}\bar{X}^{\infty}\right)\left(\bar{X}^{\infty}\right)^T \geq 0$ cannot be readily verified. This deficiency is unrelated to the size of D and X and is, in fact, the reason that convergence guarantees for the original update rules in (3.7)-(3.8) do not exist. Interestingly, if Algorithm 2 is considered in S-NMF-D mode, this difficulty is alleviated.

Corollary 4. *Let $z_D, z_X \in \{1,2\}$, $S = 1$, and the functional dependence of $p^R\left(x^i[m]\right)$ on $x^i[m]$ and of $p^R\left(d^i[n]\right)$ on $d^i[n]$ have power function forms. Then, the sequence $\{\bar{X}^t, \bar{D}^t\}_{t=1}^{\infty}$ produced by Algorithm 2 admits at least one limit point.*

Proof. The objective function is now coercive with respect to D and X as a result of the application of Theorem 3 to $-\log p^R\left(x^i[m]\right)$ and $-\log p^R\left(d^i[n]\right)$. Since $\{L^{NMF}(\bar{D}^t, \bar{X}^t)\}_{t=1}^{\infty}$ is a non-increasing sequence, the proof for Corollary 2 in 3.9.4 can be applied to obtain the stated result. ■

Corollary 5. *Let* $\{\bar{D}^t, \bar{X}^t\}_{t=1}^{\infty}$ *be a sequence generated by Algorithm 2 with* $\zeta_D =$ (3.20)*. Let* $z_X, z_D \in \{1,2\}$*, the functional dependence of* $p^R\left(x^i[m]\right)$ *on* $x^i[m]$ *have a power function form, the functional dependence of* $p^R\left(d^i[n]\right)$ *on* $d^i[n]$ *have a power function form, the columns and rows of* Y *have bounded norm, the columns of* $\bar{D}^{\infty}$ *have bounded norm, the rows of* $\bar{X}^{\infty}$ *have bounded norm, and* $\bar{D}^{\infty}$ *and* $\bar{X}^{\infty}$ *be full rank. Let one of the following conditions be satisfied: (1x)* $z_X = 1$ *and* $\tau_X \leq \lambda / \max_{m,i}\left(\left(\bar{D}^{\infty}\right)^T Y\right)[m,i]$ *or (2x)* $z_X = 2$ *and* $\tau_X \to 0$*. In addition, let one of the following conditions be satisfied: (1d)* $z_D = 1$ *and* $\tau_D \leq \lambda / \max_{m,i}\left(\bar{X}^{\infty} Y^T\right)[m,i]$ *or (2d)* $z_D = 2$ *and* $\tau_D \to 0$*. Then,* $\{\bar{D}^t, \bar{X}^t\}_{t=1}^{\infty}$ *is guaranteed to converge to set of stationary points of* $L^{NMF}(\cdot,\cdot)$*.*

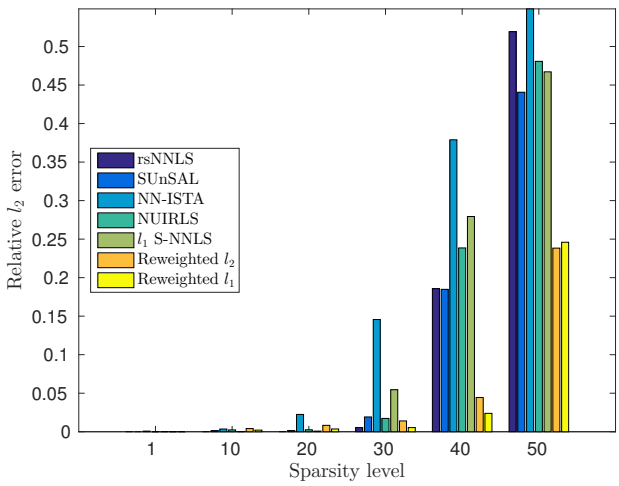
Proof. The proof is provided in 3.9.7. ■

3.7 Experimental Results

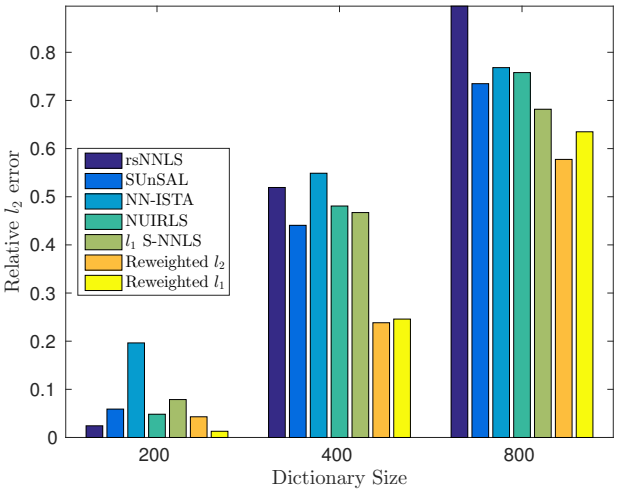
In the following, experimental results for the class of proposed algorithms are presented. The experiments performed were designed to highlight the main properties of the proposed approaches. First, the accuracy of the proposed S-NNLS algorithms on synthetic data is studied. Then, experimental validation for claims made in Section 3.6 regarding the properties of the proposed approaches is provided. Finally, the proposed framework is shown in action on real-world data by learning a basis for a database of face images.

3.7.1 S-NNLS Results on Synthetic Data

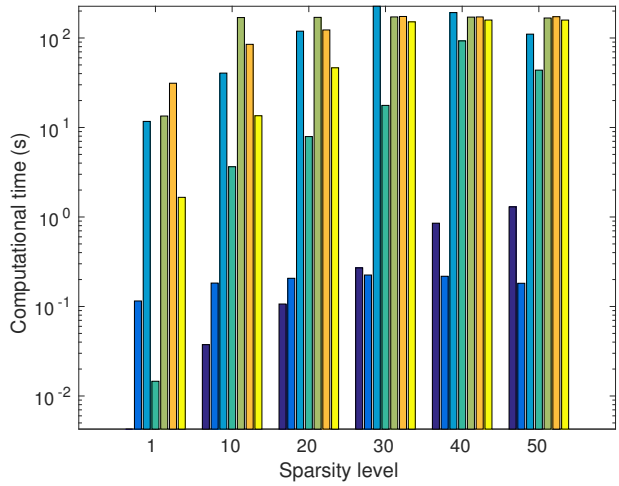
In order to compare the described methods, a sparse recovery experiment was undertaken. First, a dictionary $D \in \mathbb{R}_+^{100\times M}$ is generated, where each element of M is drawn from the $\mathsf{RG}(0,1)$ distribution. The columns of M are then normalized to have unit ℓ_2 norm. The matrix $X \in \mathbb{R}_+^{M\times 100}$ is then generated by randomly selecting s coefficients of x^i to be non-zero and drawing the non-zero values from a $\mathsf{RG}(0,1)$ distribution. The columns of X are normalized to have unit ℓ_2 norm.



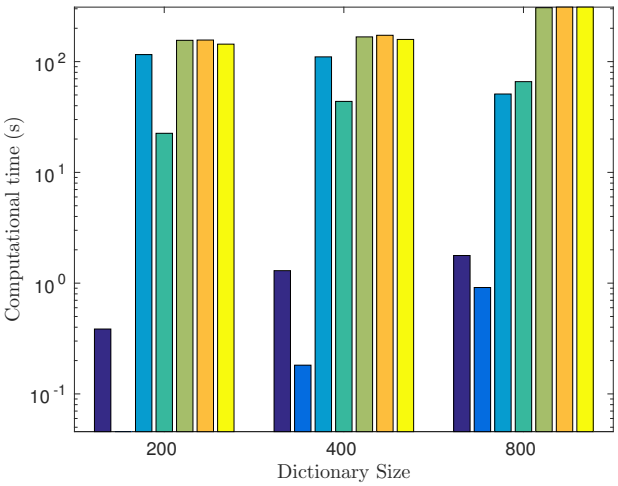
(a) Average relative ℓ_2 error as a function of sparsity level for $M = 400$.



(b) Average relative ℓ_2 error as a function of M for sparsity level 50.



(c) Average computational time as a function of sparsity level for $M = 400$.



(d) Average computational time as a function of M for sparsity level 50.

Figure 3.2: S-NNLS results on synthetic data. The legends for (c) and (d) have been omitted, but are identical to the legends in (a) and (b).

We then feed $Y = DX$ and D to the S-NNLS algorithm and approximate x^i with $\hat{x}^i$. Note that this is a noiseless experiment. The distortion of the approximation is measured using the relative Frobenius norm error,

$$\frac{\left\|X - \hat{X}\right\|_F}{\|X\|_F}. \tag{3.42}$$

A total of 50 trials are run and averaged results are reported.

We use Algorithm 1 to generate recovery results for the proposed framework, with the number of inner-loop iterations, S, of Algorithm 1 set to 2000 and the outer EM loop modified to run a maximum of 50 iterations. For reweighted ℓ_2 S-NNLS, the same annealing strategy for τ as reported in [9] is employed, where τ is initialized to 1 and decreased by a factor of 10 (up to a pre-specified number of times) when the relative ℓ_2 difference between $\bar{x}^{i,t+1}$ and $\bar{x}^{i,t}$ is below $\sqrt{\tau}/100$ for each i. Note that this strategy does not influence the convergence properties described in Section 3.6 for the reweighted ℓ_2 approach since τ can be viewed as fixed after a certain number of iterations. For reweighted ℓ_1 S-NNLS, we use $\tau = 0.1$. The regularization parameter λ is selected using cross-validation by running the S-NNLS algorithms on data generated using the same procedure as the test data.

We compare our results with rsNNLS [65], the SUnSAL algorithm for solving (3.10) [98], the non-negative ISTA (NN-ISTA) algorithm[1] for solving (3.10) [99], NUIRLS, and ℓ_1 S-NNLS [67] (i.e (3.11)). Since rsNNLS requires k as an input, we incorporate knowledge of k into the tested algorithms in order to have a fair comparison. This is done by first thresholding $\hat{x}^i$ by zeroing out all of the elements except the largest k and then executing (3.8) until convergence.

The S-NNLS results are shown in Fig. 3.2. Fig. 3.2a shows the recovery results for $M = 400$ as a function of the sparsity level s. All of the tested algorithms perform almost equally well up to $s = 30$, but the reweighted approaches dramatically outperform the competing methods

[1] We modify the soft-thresholding operator to $max\left(0, |x^i| - \beta\right)$

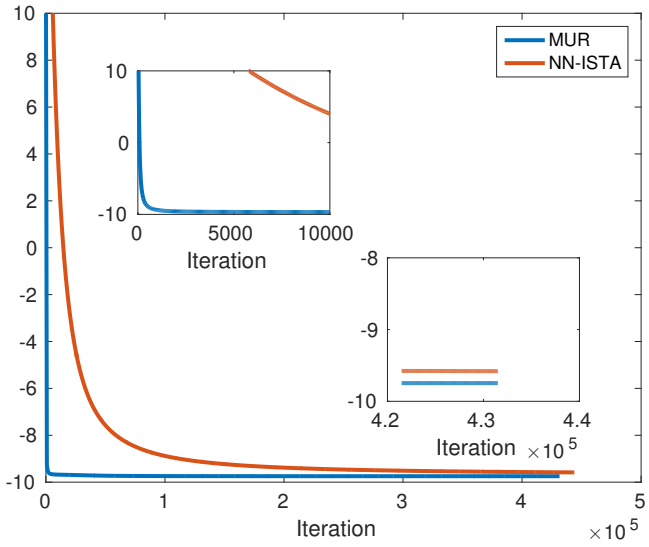


Figure 3.3: Evolution of $L(X)$ for the reweighted ℓ_1 formulation in Section 3.4.2 using Algorithm 1 and a baseline approach employing the NN-ISTA algorithm.

for $s = 40$ and $s = 50$. Fig. 3.2b shows the recovery results for $s = 50$ as a function of M. All of the tested algorithms perform relatively well for $M = 200$, but the reweighted approaches separate themselves for $M = 400$ and $M = 800$. Fig. 3.2c and 3.2d show the average computational time for the algorithms tested as a function of sparsity level and dictionary size, respectively.

Two additional observations from the results in Fig. 3.2a can be made. First, the reweighted approaches perform slightly worse for sparsity levels $s \leq 20$. We believe that this is a result of suboptimal parameter selection for the reweighted algorithms and using a finer grid during cross-validation would improve the result. This claim is supported by the observation that NUIRLS performs at least as well or better than the reweighted approaches for $s \leq 20$ and, as argued in Section 3.4.1, NUIRLS is equivalent to reweighted ℓ_2 S-NNLS in the limit $\lambda, \tau \to 0$. The second observation is that the reweighted ℓ_2 approach consistently outperforms NUIRLS at high values of k. This suggests that the strategy of allowing $\lambda > 0$ and annealing τ, instead of setting it to 0 as in NUIRLS [84], is much more robust.

In addition to displaying superior S-NNLS performance, the proposed class of MUR's also exhibits fast convergence. Fig. 3.3 compares the evolution of the objective function $L(X)$

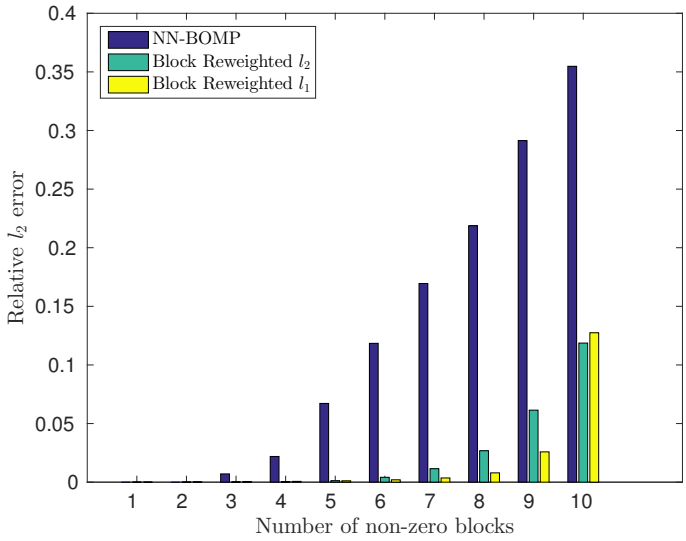


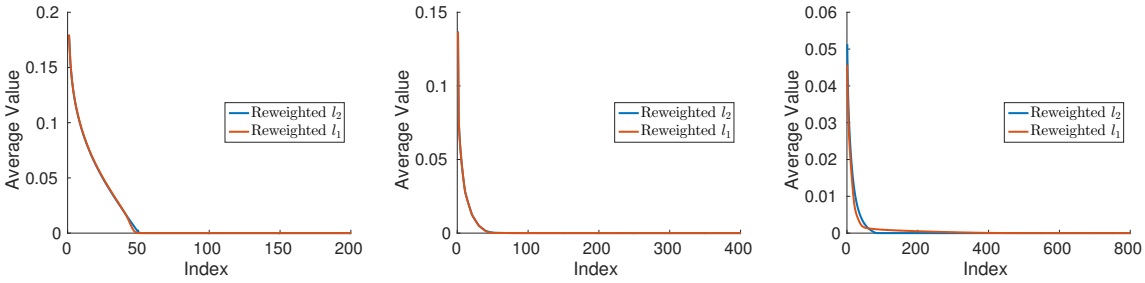
Figure 3.4: Block sparse recovery results

under the RGDP signal prior (i.e. the reweighted ℓ_1 formulation of Section 3.4.2) for Algorithm 1, with $S = 1$, with a baseline approach. The baseline employs the NN-ISTA algorithm to solve the reweighted ℓ_1 optimization problem which results from bounding the regularization term by a linear function of $x^i[m]$ (similar to (3.25), but with $\|X/E^t\|_F^2$ replaced by $\|X/E^t\|_1$). The experimental results show that the MUR in (3.28) achieves much faster convergence as well as a lower objective function value compared to the baseline.

3.7.2 Block S-NNLS Results on Synthetic Data

In this experiment, we first generate $D \in \mathbb{R}_+^{80\times 160}$ by drawing its elements from a $\mathsf{RG}(0,1)$ distribution. We generate the columns of $X \in \mathbb{R}_+^{160\times 100}$ by partitioning each column into blocks of size 8 and randomly selecting s blocks to be non-zero. The non-zero blocks are filled with elements drawn from a $\mathsf{RG}(0,1)$ distribution. We then attempt to recover X from $Y = DX$. The relative Frobenius norm error is used as the distortion metric and results averaged over 50 trials are reported.

The results are shown in Fig. 3.4. We compare the greedy NN-BOMP algorithm with the reweighted approaches. The reweighted approaches consistently outperform the ℓ_0 based method,



(a) $M = 200$ (b) $M = 400$ (c) $M = 800$

Figure 3.5: Average sorted coefficient value for S-NNLS with $D = \mathbb{R}_+^{100\times M}$. The value at index m represents the average value of the m'th largest coefficient in $\hat{x}^i$, averaged over all i.

showing good recovery performance even when the number of non-zero elements of each column of X is equal to the dimensionality of the column.

3.7.3 A Numerical Study of the Properties of the Proposed Methods

In this section, we seek to provide experimental verification for the claims made in the Section 3.6. First, the sparsity of the solutions obtained for the synthetic data experiments described in Section 3.7.1 is studied. Fig. 3.5 shows the magnitude of the m'th largest coefficient in $\hat{x}^i$ for various sizes of D, averaged over all 50 trials, all i, and all sparsity levels tested. The statement in Theorem 2 claims that the local minima of the objective function being optimized are sparse (i.e. that the number of nonzero entries is at most $N = 100$). In general, the proposed methods cannot be guaranteed to converge to a local minimum as opposed to a saddle point, so it cannot be expected that every solution produced by Algorithm 1 is sparse. Nevertheless, Fig. 3.5 shows that for $M = 200$ and $M = 400$, both reweighted approaches consistently find solutions with sparsity levels much smaller than 100. For $M = 800$, the reweighted ℓ_2 approach still finds solutions with sparsity smaller than 100, but the reweighted ℓ_1 method deviates slightly from the general trend.

Next, we test the claim made in Theorem 4 that the proposed approaches reach a stationary point of the objective function by monitoring the KKT residual norm of the scaled objective func-

Table 3.3: Normalized KKT residual for S-NNLS algorithms on synthetic data. For all experiments, $N = 100$ and $s = 10$.

M	200	400	800
Reweighted ℓ_2	$10^{-9.3}$	$10^{-9.4}$	$10^{-9.6}$
Reweighted ℓ_1	$10^{-9.9}$	$10^{-10.1}$	$10^{-10.4}$

Table 3.4: Normalized KKT residual for S-NMF-D algorithms on CSBL face dataset.

	D	X
Reweighted ℓ_2	$10^{-3.9}$	$10^{-5.3}$
Reweighted ℓ_1	10^{-5}	$10^{-7.3}$

tion. Note that, as in 3.9.5, the $-\log u\left(x^i[m]\right)$ terms are omitted from $L(X)$ and the minimization of $L(X)$ is treated as a constrained optimization problem when deriving KKT conditions. For instance, for reweighted ℓ_1 S-NNLS, the KKT conditions can be stated as

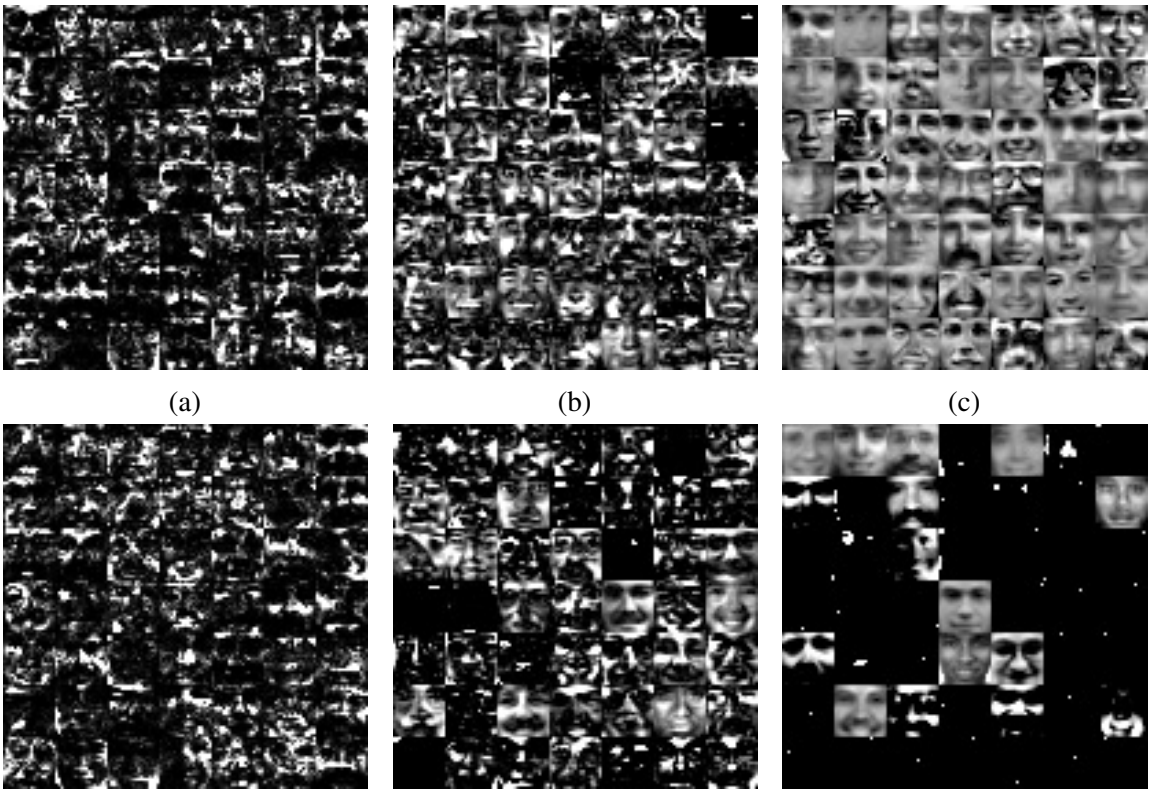
$$\min\left(X, D^T DX - D^T Y + \lambda\frac{\tau+1}{\tau+X}\right) = 0 \tag{3.43}$$

and the norm of the left-hand side, averaged over all of the elements of X, can be viewed as a measure of how close a given X is to being stationary [80]. Table 3.3 shows the average KKT residual norm of the scaled objective function for the reweighted approaches for various problem sizes. The reported values are very small and provide experimental support for Theorem 4.

3.7.4 Learning a Basis for Face Images

In this experiment, we use the proposed S-NMF and S-NMF-D frameworks to learn a basis for the CBCL face image dataset [2] [67, 85]. Each dataset image is a 19×19 grayscale image. We used $M = 3N$ and learned D by running S-NMF with reweighted-ℓ_1 regularization on X and S-NMF-D with reweighted-ℓ_1 regularization on D and X. We used $\tau_D, \tau_X = 0.1$ and ran all algorithms to convergence. Due to a scaling indeterminacy, D is normalized to have unit ℓ_2

[2] Available at `http://cbcl.mit.edu/cbcl/software-datasets/FaceData.html`



(a) (b) (c)

(d) (e) (f)

Figure 3.6: Visualization of random subset of learned atoms of D for CBCL dataset. 3.6a-3.6c: S-NMF with reweighted ℓ_1 regularization on X, $\lambda = 1e-3, 1e-2, 1e-1$, respectively. 3.6d-3.6f: S-NMF-D with reweighted ℓ_1 regularization on X and D, $\lambda = 1e-3, 1e-2, 1e-1$, respectively.

column norm at each iteration. A random subset of the learned basis vectors for each method with various levels of regularization is shown in Fig. 3.6. The results show the flexibility offered by the proposed framework. Fig. 3.6a-3.6c show that decreasing λ encourages S-NMF to learn high level features, whereas high values of λ force basis vectors to resemble images from the dataset. Fig. 3.6d-3.6f show a similar trend for S-NMF-D, although introducing a sparsity promoting prior on $\boldsymbol{D}$ tends to discourage basis vectors from resembling dataset images. It is difficult to verify Theorem 5 experimentally because D must be normalized at each iteration to prevent scaling instabilities and there is no guarantee that a given stationary point D^* has unit column norms. Nevertheless, the normalized KKT residual for the tested S-NMF-D algorithms with D normalization at each iteration on the CSBL face dataset is reported in Table 3.4.

3.7.5 Computational Issues

One of the advantages of using the proposed MUR's is that inference can be performed on the entire matrix simultaneously in each block of the block-coordinate descent procedure with relatively simple matrix operations. In fact, the computational complexity of the MUR's in (3.23), (3.28), (3.35), and (3.36) is equivalent to that of the original NMF MUR given in (3.8) (which is $O(NLr)$ where $r \leq \min(N, L)$ [85]). In other words, the proposed framework allows for performing S-NNLS and S-NMF without introducing computational complexity issues. Another benefit of this framework is that the operations required are simple matrix-based computations which lend themselves to a graphics processing unit (GPU) implementation. For example, a 9-fold speed-up is achieved in computing 500 iterations of (3.20) on a GPU compared to a CPU.

3.8 Conclusion

We presented a unified framework for S-NNLS and S-NMF algorithms. We introduced the RPESM as a sparsity promoting prior for non-negative data and provided details for a general

class of S-NNLS algorithms arising from this prior. We showed that low-complexity MUR's can be used to carry out the inference, which are validated by a monotonicity guarantee. In addition, it was shown that the class of algorithms presented is guaranteed to converge to a set of stationary points, and that the local minima of the objective function are *sparse*. This framework was then extended to a block coordinate descent technique for S-NMF and S-NMF-D. It was shown that the proposed class of S-NMF-D algorithms is guaranteed to converge to a set of stationary points.

3.9 Appendix

3.9.1 Proof of Theorem 1

Due to the assumption on the form of $p^R\left(x^i[m]\right)$, the functional dependence of $\left\langle \left(\gamma^i[m]\right)^{-z} \right\rangle$, and hence $\Omega^t[m,i]$, on $\bar{x}^{i,t}[m]$ has the form $\left(\tau + \left(\bar{x}^{i,t}[m]\right)^z\right)^{-1}$ up to a scaling constant, which is well-defined for all $\tau > 0$ and $\bar{x}^{i,t}[m] \in [0,\infty)$. As a result, (3.20) is well defined for all (m,i) such that $x^{i,s}[m] > 0$.

To show that $Q(X,\bar{X}^t)$ is non-increasing under MUR (3.20), a proof which follows closely to [66, 70] is presented. We omit the $-\log u\left(x^i[m]\right)$ term in $Q(X,\bar{X}^t)$ in our analysis because it has no contribution to $Q(X,\bar{X}^t)$ if $X \geq 0$ and the update rules are guaranteed to keep $x^i[m]$ non-negative.

First, note that $Q(X,\bar{X}^t)$ is separable in the columns of X, x^i, so we focus on minimizing $Q(X,\bar{X}^t)$ for each x^i separately. For the purposes of this proof, let y and x represent columns of Y and X, respectively, and let $Q(x)$ denote the dependence of $Q(X,\bar{X}^t)$ on one of the columns of X, with the dependency on $\bar{X}^t$ being implicit. Then,

$$Q(x) = \|y - Dx\|_2^2 + \lambda \sum_{m=1}^{M} q[m]\,(x[m])^z$$

where q represents the non-negative weights in (3.18). Let $G(x,x^s)$ be

$$G(h,h^s) = Q(x^s) + (x-x^s)^T \nabla Q(x^s) + \frac{(x-x^s)^T K(x^s)(x-x^s)}{2} \tag{3.44}$$

where $K(x^s) = \text{diag}\left(\left(D^T Dx^s + \lambda zq \odot (x^s)^{z-1}\right)/x^s\right)$. For reference,

$$\nabla Q(x^s) = D^T dx^s - D^T y + \lambda zq \odot (x^s)^{z-1} \tag{3.45}$$

$$\nabla^2 Q(x^s) = D^T D + \lambda z(z-1)\text{diag}\left(q \odot (x^s)^{z-2}\right). \tag{3.46}$$

It will now be shown that $G(x,x^s)$ is an auxiliary function for $Q(x)$. Trivially, $G(x,x) = Q(x)$. To show that $G(x,x^s)$ is an upper-bound for $Q(x)$, we begin by using the fact that $Q(x)$ is a polynomial of order 2 to rewrite $Q(x)$ as

$$Q(x) = Q(x^s) + (x-x^s)^T \nabla Q(x^s) + 0.5(x-x^s)^T \nabla^2 Q(x^s)(x-x^s). \tag{3.47}$$

It then follows that $G(x,x^s)$ is an auxiliary function for $Q(x)$ if and only if the matrix

$$A = K(x^s) - \nabla^2 Q(x^s) \tag{3.48}$$

is positive semi-definite (PSD). The matrix X can be decomposed as

$$A = A_1 + A_2, \tag{3.49}$$

where

$$A_1 = \text{diag}\left(\frac{D^T Dx^s}{x^s}\right) - D^T D \tag{3.50}$$

$$A_2 = \lambda z(2-z)\text{diag}\left(q \odot (x^s)^{z-2}\right). \tag{3.51}$$

The matrix A_1 was shown to be PSD in [70]. The matrix A_2 is a diagonal matrix with the (m,m)'th entry being $\lambda z(2-z)q[m]\left(h^s[m]\right)^{z-2}$. Since $q[m]\left(h^s[m]\right)^{z-2} \geq 0$ and $z \leq 2$, A_2 has non-negative entries on its diagonal and, consequently, is PSD. Since the sum of PSD matrices is PSD, it follows that A is PSD and $G\left(x,x^s\right)$ is an auxiliary function for $Q\left(x\right)$. Since $G\left(x,x^s\right)$ as an auxiliary function for $Q\left(x\right)$, $Q\left(x\right)$ is non-increasing under the update rule [70]

$$x^{s+1} = \arg\min_{x} G\left(x,x^s\right). \tag{3.52}$$

The optimization problem in (3.52) can be solved in closed form, leading to the MUR shown in (3.20). The multiplicative nature of the update rule in (3.20) guarantees that the sequence $\{X^s\}_{s=1}^{\infty}$ is non-negative.

3.9.2 Proof of Theorem 2

This proof is an extension of (Theorem 1 [100]) and (Theorem 8 [101]). Since $L(X)$ is separable in the columns of X, consider the dependence of $L(X)$ on a single column of X, denoted by $L(x)$. The function $L(x)$ can be written as

$$\|y-Dx\|_2^2 - 2\sigma^2 \sum_{m=1}^{M} \log p\left(x[m]\right). \tag{3.53}$$

Let x^* be a local minimum of $L(x)$. We observe that x^* must be non-negative. Note that $-\log p\left(x[m]\right) \to \infty$ when $x[m] < 0$ since $p\left(x[m]\right) = 0$ over the negative orthant. As such, if one of the elements of x^* is negative, x^* must be a global maximum of $L(x)$. Using the assumption on the form of $p^R\left(x[m]\right)$, (3.53) becomes

$$\|y-Dx\|_2^2 + \sum_{m=1}^{M} 2\sigma^2 \left(\alpha \log\left(\tau + \left(x[m]\right)^z\right) - \log u\left(x[m]\right)\right) + c \tag{3.54}$$

where constants which do not depend on h are denoted by c. By the preceding argument, $\log u\left(x[m]^*\right) = 0$, so the $\log u\left(x[m]^*\right)$ term makes no contribution to $L(x^*)$. The vector x^* must be a local minimum of the constrained optimization problem

$$\min_{y=Dx+v^*} \underbrace{\sum_{m=1}^{M} \log\left(\tau + (x[m])^z\right)}_{\phi(x)} \tag{3.55}$$

where $v^* = y - Dx^*$ and $\phi(\cdot)$ is the diversity measure induced by the prior on $\boldsymbol{X}$. It can be shown that $\phi(\cdot)$ is concave under the conditions of Theorem 2. Therefore, under the conditions of Theorem 2, the optimization problem (3.55) satisfies the conditions of (Theorem 8 [101]). It then follows that the local minima of (3.55) are basic feasible solutions, i.e they satisfy $y = Dx + v^*$ and $\|x\|_0 \leq N$. Since x^* is one of the local minima of (3.55), $\|x^*\|_0 \leq N$.

3.9.3 Proof of Theorem 3

It is sufficient to show that

$$\lim_{x^i[m]\to\infty} p\left(x^i[m]\right) = 0. \tag{3.56}$$

Consider the form of $p\left(x^i[m]\right)$ when it is a member of the RPESM family:

$$p\left(x^i[m]\right) = \int_0^\infty p\left(x^i[m]|\gamma^i[m]\right) p\left(\gamma^i[m]\right) d\gamma^i[m] \tag{3.57}$$

where $\boldsymbol{x^i[m]}|\boldsymbol{\gamma^i[m]} \sim p^{RPE}\left(x^i[m]|\gamma^i[m];z\right)$. Note that

$$\left|p^{RPE}\left(x^i[m]|\gamma^i[m]\right) p\left(\gamma^i[m]\right)\right| \leq \left|p^{RPE}\left(0|\gamma^i[m];z\right) p\left(\gamma^i[m]\right)\right|.$$

Coupled with the fact that $p\left(x^i[m]|\gamma^i[m]\right)$ is continuous over the positive orthant, the dominated convergence theorem can be applied to switch the limit with the integral in (3.57):

$$\lim_{x^i[m]\to\infty}\int_0^\infty p\left(x^i[m]|\gamma^i[m]\right)p\left(\gamma^i[m]\right)d\gamma^i[m]=\int_0^\infty \lim_{x^i[m]\to\infty} p\left(x^i[m]|\gamma^i[m]\right)p\left(\gamma^i[m]\right)d\gamma^i[m] \quad (3.58)$$

$$=0. \quad (3.59)$$

3.9.4 Proof of Corollary 2

This proof follows closely to the first part of the proof of (Theorem 1, [86]). Let

$$\mathscr{S}_0=\{X\in\mathbb{R}_+^{M\times L}|L(X)\leq L(\bar{X}^0)\}. \quad (3.60)$$

Lemma 1 established that $L(X)$ is coercive. In addition, $L(X)$ is a continuous function of X over the positive orthant. Therefore, $\mathscr{S}_0$ is a compact set (Theorem 1.2, [102]). The sequence $\{L(\bar{X}^t)\}_{t=1}^\infty$ is non-increasing as a result of Theorem 1, such that $\{\bar{X}^t\}_{t=1}^\infty\in\mathscr{S}_0$. Since $\mathscr{S}_0$ is compact, $\{\bar{X}^t\}_{t=1}^\infty$ admits at least one limit point.

3.9.5 Proof of Theorem 4

From Lemma 2, the sequence $\{\bar{X}^t\}_{t=1}^\infty$ admits at least one limit point. What remains is to show that every limit point is a stationary point of (3.17). The sufficient conditions for the limit points to be stationary are (Theorem 1, [26])

1. $Q(X,\bar{X}^t)$ is continuous in both X and $\bar{X}^t$,

2. At each iteration t, one of the following is true

$$Q\left(x^{i,t+1},\bar{x}^{i,t}\right) < Q\left(\bar{x}^{i,t},\bar{x}^{i,t}\right) \tag{3.61}$$

$$\bar{x}^{i,t+1} = \underset{x^i \geq 0}{\arg\min}\, Q\left(x^i,\bar{x}^{i,t}\right). \tag{3.62}$$

The function $Q(X,\bar{X}^t)$ is continuous in X, trivially, and in $\bar{X}^t$ if the functional dependence of $p^R\left(\bar{x}^{i,t}[m]\right)$ on $\bar{x}^{i,t}[m]$ has the form (3.41).

In order to show that the descent condition is satisfied, we begin by noting that $Q\left(x^i,\bar{x}^{i,t}\right)$ is strictly convex with respect to x^i if the conditions of Theorem 4 are satisfied. This can be seen by examining the expression for the Hessian of $Q\left(x^i,\bar{x}^{i,t}\right)$ in (3.46). If D is full rank, then D^TD is positive definite. In addition, the matrix

$$\lambda z(z-1)\mathrm{diag}\left(\Omega^t[:,i]\odot\left(x^{i,s}\right)^{z-2}\right) \tag{3.63}$$

is PSD because $z \geq 1$. Therefore, the Hessian of $Q\left(x^i,\bar{x}^{i,t}\right)$ is positive definite if the conditions of Theorem 4 are satisfied.

Since $S=1$, $\bar{x}^{i,t+1}$ is generated by (3.20) with X^s replaced by $\bar{X}^t$. This update has two possibilities:

1. $\bar{x}^{i,t+1} \neq \bar{x}^{i,t}$, or

2. $\bar{x}^{i,t+1} = \bar{x}^{i,t}$.

If condition (1) is true, then (3.61) is satisfied because of the strict convexity of $Q\left(x^i,\bar{x}^{i,t}\right)$ and the monotonicity guarantee of Theorem 1.

It will now be shown that if condition (2) is true, then $\bar{x}^{i,t+1}$ must satisfy (3.62). Since $Q\left(x^i,\bar{x}^{i,t}\right)$ is convex, any $\bar{x}^{i,t+1}$ which satisfies the Karush-Kuhn-Tucker (KKT) conditions associ-

[3] As in the proof of Theorem (1), we omit the $-\log u\left(x^i[m]\right)$ term from $Q\left(x^i,\bar{x}^{i,t}\right)$ and explicitly enforce the non-negativity constraint on x^i.

ated with (3.62) must be a solution to (3.62) [97]. The KKT conditions associated with (3.62) are given by [85]:

$$x^i[m] \geq 0 \tag{3.64}$$

$$\left(\nabla Q\left(x^i, \bar{x}^{i,t}\right)\right)[m] \geq 0 \tag{3.65}$$

$$x^i[m]\left(\nabla Q\left(x^i, \bar{x}^{i,t}\right)\right)[m] = 0 \tag{3.66}$$

for all m. The expression for $\nabla Q\left(x^i, \bar{x}^{i,t}\right)$ is given in (3.45). For any m such that $\bar{x}^{i,t+1}[m] > 0$, $\left(D^T X\right)[m,i] = \left(D^T D\bar{X}^{t+1}\right)[m,i] + \lambda\Omega^t[m,i]\left(\bar{X}^{t+1}_{(i,j)}\right)^{z-1}$ because $\bar{X}^{t+1}$ was generated by (3.20). This implies that

$$\left(\nabla Q\left(x^i, \bar{x}^{i,t}\right)\Big|_{x^i=\bar{x}^{i,t+1}}\right)[m] = 0$$

for all m such that $\bar{x}^{i,t+1}[m] > 0$. Therefore, all of the KKT conditions are satisfied.

For any i such that $\bar{x}^{i,t+1}[m] = 0$, (3.64) and (3.66) are trivially satisfied. To see that (3.65) is satisfied, first consider the scenario where $z = 1$. In this case,

$$\begin{aligned}
&\lim_{\bar{x}^{i,t+1}[m]\to 0}\left(\nabla Q\left(x^i, \bar{x}^{i,t}\right)\Big|_{x^i=\bar{x}^{i,t+1}}\right)[m] \\
&=^1 \lim_{\bar{x}^{i,t+1}[m]\to 0}\left(D^T D\bar{x}^{t+1}\right)[m] + \frac{\lambda\left(\bar{x}^{i,t+1}[m]\right)^0}{\tau + \left(\bar{x}^{i,t+1}[m]\right)^1} - \left(D^T X\right)[m] \\
&= c + \frac{\lambda}{\tau} - \left(D^T x\right)[m] \geq^2 0
\end{aligned}$$

where $c \geq 0$, (1) follows from the assumption on $p^R\left(x^i[m]\right)$ having a power exponential form, and (2) follows from the assumptions that the elements of $D^T x$ are bounded and $\tau \leq \lambda / \max_{m,i}\left(D^T X\right)[m,i]$.

When $z = 2$,

$$\begin{aligned}
&\lim_{\bar{x}^{i,t+1}[m]\to 0} \lim_{\tau\to 0} \left(\nabla Q\left(x^i, \bar{x}^{i,t}\right) \Big|_{x^i=\bar{x}^{i,t+1}} \right)_i \\
&=^1 \lim_{\bar{x}^{i,t+1}[m]\to 0} \left(D^T D\bar{x}^{t+1}\right)[m] + \frac{2\lambda}{\bar{x}^{i,t+1}[m]} - \left(D^T x\right)[m] \\
&\geq^2 0
\end{aligned}$$

where (1) follows from the assumption on $p^R\left(x^i[m]\right)$ having a power exponential form and (2) follows from the assumption that the elements of $D^T X$ are bounded. Therefore, (3.65) is satisfied for all m such that $\bar{x}^{i,t+1}[m] = 0$. To conclude, if $\bar{x}^{i,t+1}$ satisfies $\bar{x}^{i,t+1} = \bar{x}^{i,t}$, then it satisfies the KKT conditions and must be solution of (3.62).

3.9.6 Proof of Corollary 3

In the S-NMF setting ($\zeta_D =$ (3.8), $\zeta_X =$ (3.20)), this result follows from the application of (Theorem 1 [70]) to the D update stage of Algorithm 2 and the application of Theorem 1 to the X update stage of Algorithm 2. In the S-NMF-D setting ($\zeta_D =$ (3.20), $\zeta_X =$ (3.20)), the result follows from the application of Theorem 1 to each step of Algorithm 2. In both cases,

$$L^{NMF}(\bar{D}^t, \bar{X}^t) \geq L^{NMF}(\bar{D}^{t+1}, \bar{X}^t) \geq L^{NMF}(\bar{D}^{t+1}, \bar{X}^{t+1}).$$

3.9.7 Proof of Corollary 5

The existence of a limit point $(\bar{D}^\infty, \bar{X}^\infty)$ is guaranteed by Corollary 4. It is sufficient to show that $L^{NMF}(\cdot,\cdot)$ is stationary with respect to $\bar{D}^\infty$ and $\bar{X}^\infty$ individually. The result follows by application of Theorem 4 to $\bar{D}^\infty$ and $\bar{X}^\infty$.

Chapter 3, in full, is a reprint of material published in the article Igor Fedorov, Alican Nalci, Ritwik Giri, Bhaskar D. Rao, Truong Q. Nguyen, and Harinath Garudadri, “A Unified

Framework for Sparse Non-Negative Least Squares using Multiplicative Updates and the Non-Negative Matrix Factorization Problem," Signal Processing, 2018. I was the primary author and B. D. Rao supervised the research.

Chapter 4

Multimodal Sparse Bayesian Dictionary Learning

4.1 Introduction

Due to improvements in sensor technology, acquiring vast amounts of data has become relatively easy. Given the ability to harvest data, the task becomes how to extract relevant information from the data. The data is often multimodal, which introduces novel challenges in learning from it. Multimodal dictionary learning has become a popular tool for fusing multimodal information [103, 104, 105, 106, 30].

Let L and J denote the number of data points for each modality and number of modalities, respectively. Let $Y_j = \begin{bmatrix} y_j^1 & \dots & y_j^L \end{bmatrix} \in \mathbb{R}^{N_j \times L}$ denote the data matrix for modality j, where y_j^i denotes the i'th data point for modality j. We use uppercase symbols to denote matrices and lowercase symbols to denote the corresponding matrix columns. The multimodal dictionary learning problem consists of estimating dictionaries $D_j \in \mathbb{R}^{N_j \times M_j}$ given data Y_j such that $Y_j \approx D_j X_j \; \forall j$. We focus on overcomplete dictionaries because they are more flexible in the range of signals they can represent [37]. Since y_j^i admits an infinite number of representations under

overcomplete D_j, we seek sparse x_j^i [10].

Without any further constraints, the multimodal dictionary learning problem can be viewed as J independent unimodal problems. To fully capture the multimodal nature of the problem, the learning process must be adapted to encode the prior knowledge that each set of points $\boldsymbol{y^i} = \left\{y_j^i\right\}_{j=1}^J$ is generated by a common source which is measured J different ways, where $[J] = \{1, \cdots, J\}$. For any variable x, $\boldsymbol{x}$ denotes $\left\{x_j\right\}_{j=1}^J$. For instance, in [107], low and high resolution image patches are modeled as y_1^i and y_2^i, respectively, and the association between y_1^i and y_2^i is enforced by the constraint $x_1^i = x_2^i$. The resulting multimodal dictionary learning problem, referred to here as ℓ_1DL, is to solve [17]

$$\begin{aligned} &\underset{\tilde{D},X}{\arg\min} \left\|\tilde{Y} - \tilde{D}X\right\|_F^2 + \lambda \|X\|_1 \\ &\tilde{Y} = \begin{bmatrix} Y_1^T & \cdots & Y_J^T \end{bmatrix}^T, \tilde{D} = \begin{bmatrix} D_1^T & \cdots & D_J^T \end{bmatrix}^T, \end{aligned} \tag{4.1}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\|X\|_1 = \sum_{i\in[L]} \left\|x^i\right\|_1$, and the ℓ_1-norm is used as a convex proxy to the ℓ_0 sparsity measure. In a classification framework, (4.1) can be viewed as learning a multimodal feature extractor, where the optimizer is the multimodal representation of $\boldsymbol{y^i}$ that is fed into a classifier [103, 108, 104]. There are 4 significant deficiencies associated with using ℓ_1DL for multimodal dictionary learning:

D1 While using the same sparse code for each modality establishes an explicit relationship between the dictionaries for each modality, the same coefficient values may not be able to represent different modalities well.

D2 Some data modalities are often less noisy than others and the algorithm should be able to leverage the clean modalities to learn on the noisy ones. Since (4.1) constrains λ to be the same for all modalities, it is unclear how the learning algorithm can incorporate prior knowledge about the noise level of each modality.

D3 The formulation in (4.1) constrains $M_j = M \; \forall j$, for some M. Since dimensionality can vary across modalities, it is desirable to allow M_j to vary.

D4 The choice of λ is central to the success of approaches like (4.1). If λ is chosen too high, the reconstruction term is ignored completely, leading to poor dictionary quality. If λ is chosen too low, the sparsity promoting term is effectively eliminated from the objective function. Extensive work has been done to approach this hyperparameter selection problem from various angles. Two popular approaches include treating the hyperparameter selection problem as an optimization problem of its own [109, 110] and grid search, with the latter being the prevailing strategy in the dictionary learning community [17, 13]. In either case, optimization of λ involves evaluating (4.1) at various choices of λ, which can be computationally intensive and lead to suboptimal results in practice.

Next, we review relevant works from the dictionary learning literature, highlighting each method's benefits and drawbacks in light of **D1**-**D4**. In past work, $M_j = M \; \forall j$, thus exhibiting **D3**. One of the seminal unimodal dictionary learning algorithms is K-SVD, which optimizes[10]

$$\underset{D,\, \{\|x^i\|_0 \leq s\}_{i=1}^{L}}{\arg\min} \; \|Y - DX\|_F^2 \,, \tag{4.2}$$

where s denotes the desired sparsity level and modality subscripts have been omitted for brevity. The K-SVD algorithm proceeds in a block-coordinate descent fashion, where D is optimized while holding X fixed and vice-versa. The update of X involves a greedy ℓ_0 pseudo-norm minimization procedure [5]. In a multimodal setting, K-SVD can be naively adapted, where Y is replaced by $\tilde{Y}$ and D by $\tilde{D}$ in (4.2), as in (4.1).

One recent approach, referred to here as Joint ℓ_0 Dictionary Learning (Jℓ_0DL), builds

upon K-SVD for the multimodal dictionary learning problem and proposes to solve [106]

$$\underset{\left\{\left\{\chi_j^i=\chi^i\right\}_{j=1}^{J},|\chi^i|\leq s\right\}_{i=1}^{L}}{\arg\min} \sum_{j=1}^{J} \lambda_j \left\|Y_j - D_j X_j\right\|_F^2 \tag{4.3}$$

where χ_j^i denotes the support of x_j^i. The Jℓ_0DL algorithm tackles **D1**-**D2** by establishing a correspondence between the supports of the sparse codes for each modality and by allowing modality specific regularization parameters, which allow for encoding prior information about the noise level of y_j^i. On the other hand, Jℓ_0DL does not address **D3** and presents even more of a challenge than ℓ_1DL with respect to **D4** since the size of the grid search needed to find $\boldsymbol{\lambda}$ grows exponentially with J. Another major drawback of Jℓ_0DL is that, since (4.3) has an ℓ_0 type constraint, solving it requires a greedy algorithm which can produce poor solutions, especially if some modalities are much noisier than others [111].

The multimodal version of (4.1), referred to here as Joint ℓ_1 DL (Jℓ_1DL), seeks [32]

$$\underset{\boldsymbol{D},\boldsymbol{X}}{\arg\min} \frac{1}{2} \sum_{i\in[L],j\in[J]} \left\|y_j^i - D_j x_j^i\right\|_2^2 + \lambda \sum_{i\in[L]} \left\|\Pi^i\right\|_{12} \tag{4.4}$$

where $\Pi^i = \begin{bmatrix} x_1^i & \cdots & x_J^i \end{bmatrix}$, $\left\|\Pi^i\right\|_{12} = \sum_{m\in[M]} \left\|\Pi^i[m,:]\right\|_2$, and $\Pi^i[m,:]$ denotes the m'th row of Π^i. The ℓ_{12}-norm in (4.4) promotes row sparsity in Π^i, which promotes $\boldsymbol{x^i}$ that share the same support. Like all of the previous approaches, Jℓ_1DL adopts a block-coordinate descent approach to solving (4.4), where an alternating direction method of multipliers algorithm is used to compute the sparse code update stage [112]. While Jℓ_1DL makes progress toward addressing **D1**, it does so at the cost of sacrificing the hard constraint that $\boldsymbol{x^i}$ share the same support. The authors of [32] attempt to address **D2** by relaxing the constraint on the support of $\boldsymbol{x^i}$ even more, but we will not study this approach here because it moves even further from the theme of this work. In addition, Jℓ_1DL does not address **D3**-**D4**.

One desirable property of dictionary learning approaches is that they be scalable with

respect to the size of the dictionary as well as the dataset. When L becomes large, the algorithm must be able to learn in a stochastic manner, where only a subset of the data samples need to be processed at each iteration. Stochastic learning strategies have been studied in the context of ℓ_1DL [17, 12] and Jℓ_1DL [32], but not for K-SVD or Jℓ_0DL. Likewise, the algorithm should be able to accommodate large N_j.

When the dictionary learning algorithm is to be used as a building block in a classification framework, class information can be incorporated within the learning process. In a supervised setting, the input to the algorithm is $\{\boldsymbol{Y}, H\}$, where $H = \begin{bmatrix} h^1 & \cdots & h^L \end{bmatrix} \in \mathbb{B}^{C\times L}$ is the binary class label matrix for the dataset and C is the number of classes. Each h^i is the label for the i'th data point in a one-of-C format. This type of dictionary learning is referred to as task-driven [32, 12], label consistent [13], or discriminative [14] and the goal is to learn a dictionary D_j such that x_j^i is indicative of the class label. For instance, discriminative K-SVD (D-KSVD) optimizes [14]

$$\underset{D,W,\{\|x^i\|_0\leq s\}_{i=1}^L}{\arg\min} \|Y - DX\|_F^2 + \lambda_{su}\|H - WX\|_F^2 \tag{4.5}$$

where W can be viewed as a linear classifier for x^i.

Task-driven ℓ_1DL (TD-ℓ_1DL) optimizes [12]

$$\underset{D,W}{\arg\min}\, E_x\left[l_{su}(h, W, x^*(y, D)\right] + \nu\|W\|_F^2 \tag{4.6}$$

where $E_x[\cdot]$ denotes the expectation with respect to $p(x)$, $l_{su}(\cdot)$ denotes the supervised loss function[1], $x^*(y, D)$ denotes the solution of (4.1) with the dictionary fixed to D, and the last term provides regularization on W to avoid over-fitting.

[1]Examples of supervised losses include the squared loss in (4.5), logistic loss, and hinge loss [12].

Task-driven Jℓ_1DL (TD-Jℓ_1DL) optimizes [32]

$$\underset{\boldsymbol{D},\boldsymbol{W}}{\arg\min} E_{\boldsymbol{x}} \left[\sum_{j\in[J]} l_{su}(h_j, W_j, x_j^*(y_j, D_j) \right] + \nu \sum_{j\in[J]} \left\| W_j \right\|_F^2$$

where $x_j^*(y_j, D_j)$ denotes the j'th modality sparse code in the optimizer of (4.4) with fixed dictionaries.

4.1.1 Contributions

We present the multimodal sparse Bayesian dictionary learning algorithm (MSBDL). MSBDL is based on a hierarchical sparse Bayesian model, which was introduced in the context of Sparse Bayesian Learning (SBL) [29, 49, 27] as well as dictionary learning [29], and has since been extended to various structured learning problems [33, 25, 88]. We presented initial work on this approach in [30], where we made some progress towards tackling **D1**-**D2**. Here, we go beyond our preliminary work in a number of significant ways. We address **D1**-**D2** to a fuller extent by offering scalable and task-driven variants of MSBDL. More importantly, we tackle **D3**-**D4**, where our solution to **D4** is crucial to the ability of MSBDL to address **D1** and resolves a major hyperparameter tuning issue in [30]. By resolving **D3**, the present work represents a large class of algorithms that contains [30] as a special case. In summary:

1. We extend MSBDL to address **D3**. To the best of our knowledge, MSBDL is the first dictionary learning algorithm capable of learning differently sized dictionaries.

2. We extend MSBDL to the task-driven learning scenario.

3. We present scalable versions of MSBDL.

4. We optimize algorithm hyperparameters during learning, obviating the need for grid search, and conduct a theoretical analysis of the approach.

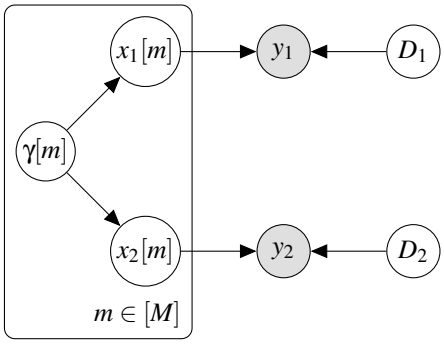


Figure 4.1: Graphical model for two modality MSBDL.

5. We show that multimodal dictionary learning offers provable advantages over unimodal dictionary learning.

4.2 Proposed Approach

The graphical model for MSBDL is shown in Fig. 4.1. The signal model is given by

$$y_j = D_j x_j + v_j, \;\; v_j \sim \mathsf{N}\left(0, \sigma_j^2 \mathsf{I}\right) \tag{4.7}$$

where $\mathsf{N}(\cdot)$ denotes a Gaussian distribution and the noise variance is allowed to vary among modalities. In order to promote sparse x_j^i, we assume a Gaussian Scale Mixture (GSM) prior on each element of x_j [49, 27, 50, 113]. The GSM prior is a popular class of distributions whose members include many sparsity promoting priors, such as the Laplacian and Student's-t [49, 27, 50, 113, 24, 7]. The remaining task is to specify the conditional density of x_j given γ. One option is to use what we refer to as the one-to-one prior [30]:

$$p\left(x_j|\gamma\right) = \prod_{m\in[M]} p\left(x_j[m]|\gamma[m]\right) = \prod_{m\in[M]} \mathsf{N}\left(0, \gamma[m]\right) \tag{4.8}$$

where $x_j[m]$ denotes the m'th element of x_j and the choice of $p(\gamma[m])$ determines the marginal density of $x_j[m]$. We assume a non-informative prior on $\gamma_j[m]$ [49]. As will be shown in Section

4.2.2, the conditional distribution in (4.8) represents a Bayesian realization of the constraint that $\boldsymbol{x}$ share the same support. The prior in (4.8) still constrains $M_j = M \; \forall j$, but this restriction will be lifted in Section 4.5. When $J = 1$, this model is identical to the one used in [29].

4.2.1 Inference Procedure

We adopt an empirical Bayesian inference strategy to estimate $\theta = \left\{ \boldsymbol{D}, \{\gamma^i\}_{i=1}^L \right\}$ by optimizing [35]

$$\arg\max_{\theta} \log p\left(\boldsymbol{Y}|\theta\right) = \arg\max_{\theta} \sum_{j\in[J]} \log p\left(Y_j|\theta\right) \tag{4.9}$$

$$p\left(\boldsymbol{Y}|\theta\right) = \prod_{i\in[L], j\in[J]} p\left(y_j^i|\theta\right), p\left(y_j^i|\theta\right) = \mathsf{N}(0, \Sigma_y^i) \tag{4.10}$$

$$\Sigma_{y,j}^i = \sigma_j^2 \mathsf{I} + D_j \Gamma^i D_j^T, \;\; \Gamma^i = \mathrm{diag}(\gamma^i). \tag{4.11}$$

We use Expectation-Maximization (EM) to maximize (4.9), where $\{\boldsymbol{X}, \boldsymbol{Y}, \theta\}$ and $\boldsymbol{X}$ are treated as the complete and nuisance data, respectively [22]. At iteration t, the E-step computes $Q\left(\theta, \theta^t\right) = \left\langle \log p\left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \{\gamma^i\}_{i=1}^L\right) \right\rangle$, where $\langle\cdot\rangle$ denotes the expectation with respect to $p\left(\boldsymbol{X}|\boldsymbol{Y}, \theta^t\right)$, and θ^t denotes the estimate of θ at iteration t. Due to the conditional independence properties of the model, the posterior factors over i and

$$p\left(x_j^i|y_j^i, \theta\right) = \mathsf{N}\left(\mu_j^i, \Sigma_{x,j}^i\right) \tag{4.12}$$

$$\Sigma_{x,j}^i = \left(\sigma_j^{-2} D_j^T D_j + \left(\Gamma^i\right)^{-1}\right)^{-1} \tag{4.13}$$

$$\mu_j^i = \sigma_j^{-2} \Sigma_{x,j}^i D_j^T y_j^i. \tag{4.14}$$

In the M-step, $Q(\theta,\theta^t)$ is maximized with respect to θ. In general, the M-step depends on the choice of $p(\boldsymbol{x}|\boldsymbol{\gamma})$. For the choice in (4.8), the M-step becomes

$$\left(\gamma^i[m]\right)^{t+1} = J^{-1}\sum_{j=1}^{J}\Sigma^i_{x,j}[m,m] + \left(\mu^i_j[m]\right)^2 \tag{4.15}$$

$$D_j^{t+1} = Y_j U_j^T\left(U_j U_j^T + \sum_{i\in[L]}\Sigma^i_{x,j}\right)^{-1} \tag{4.16}$$

$$U_j = \begin{bmatrix}\mu_j^1 & \cdots & \mu_j^L\end{bmatrix}. \tag{4.17}$$

4.2.2 How does MSBDL solve deficiency D1?

One consequence of the GSM prior is that many of the elements of $\boldsymbol{\gamma}^i$ converge to 0 during inference [27]. When $\boldsymbol{\gamma}^i[m]=0$, $p\left(x^i_j[m]|y^i_j,\boldsymbol{\gamma}^i\right)$ reduces to $\delta\left(x_j[m]\right)$ for all j, where $\delta(\cdot)$ denotes the Dirac-delta function [27]. Since the only role of $\boldsymbol{x}$ in the inference procedure is in the E-step, where we take the expectation of the complete data log-likelihood with respect to $p\left(x^i_j|y^i_j,\boldsymbol{\gamma}^i\right)$, the effect of having $p\left(x^i_j[m]|y^i_j,\boldsymbol{\gamma}^i\right)=\delta\left(x^i_j[m]\right)$ is that the E-step reduces to evaluating the complete data log-likelihood at $x^i_j[m]=0,\forall j$. Therefore, upon convergence, the proposed approach exhibits the property that $\boldsymbol{x^i}$ share the same support.

4.2.3 Connection to Jℓ_1DL

Suppose that the conditional distribution in (4.8) is used with $\boldsymbol{\gamma}[m]\sim \mathsf{Ga}(J/2,1)\ \forall m$, where $\mathsf{Ga}(\cdot)$ refers to the Gamma distribution. It can be shown that [114]

$$p\left(\boldsymbol{x^i}\right) = c\prod_{m\in[M]}K_0\left(\left\|\begin{bmatrix}x^i_1[m] & \cdots & x^i_J[m]\end{bmatrix}\right\|_2\right),$$

where c is a normalization constant and $K_0(\cdot)$ denotes the modified Bessel function of the second kind and order 0. For large x, $K_0(x)\approx\pi\exp(-x)/\sqrt{2\pi x}$ [115]. In the following, we replace $K_0(x)$

by its approximation for purposes of exposition. Under the constraint $\sigma_j = 0.5\lambda \; \forall j$, the MAP estimate of $\{\boldsymbol{D}, \boldsymbol{X}\}$ is given by

$$\begin{aligned} \underset{\boldsymbol{D},\boldsymbol{X}}{\arg\min} \sum_{i\in[L],j\in[J]} \left\|y_j^i - D_j x_j^i\right\|_2^2 + \lambda \sum_{i\in[L]} \left\|\Pi^i\right\|_{12} + \\ 0.5\lambda \sum_{i\in[L],m\in[M]} \log \left\|\Pi^i[m,:]\right\|_2 . \end{aligned} \tag{4.18}$$

This analysis exposes a number of similarities between MSBDL and Jℓ_1DL. If we ignore the last term in (4.18), Jℓ_1DL becomes the MAP estimator of $\{\boldsymbol{D}, \boldsymbol{X}\}$ under the one-to-one prior in (4.8). If we keep the last term in (4.18), the effect is to add a Generalized Double Pareto prior on the ℓ_2 norm of the rows of Π^i [7, 116].

At the same time, there are significant differences between MSBDL and Jℓ_1DL. The Jℓ_1DL objective function assumes σ_j is constant across modalities, which can lead to a strong mismatch between data and model when the dataset consists of sources with disparate noise levels. In contrast to Jℓ_1DL, MSBDL enjoys the benefits of evidence maximization [117, 35, 7], which naturally embodies "Occam's razor" by penalizing unnecessarily complex models and searches for areas of large posterior mass, which are more informative than the mode when the posterior is not well-behaved.

4.3 Complete Algorithm

So far, it has been assumed that $\boldsymbol{\sigma}$ is known. Although it is possible to include $\boldsymbol{\sigma}$ in θ and estimate it within the evidence maximization framework in (4.9), we experimental observe that $\boldsymbol{\sigma}$ decays very quickly and the algorithm tends to get stuck in poor stationary points. An alternative approach is described next. Consider the MAP estimate of x_j given y_j, D_j:

$$\underset{x_j}{\arg\min} \left\|y_j - D_j x_j\right\|_2^2 - 2\sigma_j^2 \log p\left(x_j\right). \tag{4.19}$$

The estimator in (4.19) shows that σ_j can be thought of as a regularization parameter which controls the trade-off between sparsity and reconstruction error. As such, we propose to anneal σ_j. The motivation for annealing σ_j is that the quality of D_j increases with t, so giving too much weight to the reconstruction error term early on can force EM to converge to a poor stationary point.

Let $\sigma_j^0 > \sigma^\infty \geq 0, \alpha_\sigma < 1$, $\tilde{\sigma}_j^{t+1} = \max(\sigma^\infty, \alpha_\sigma \sigma_j^t)$. The proposed annealing strategy can then be stated as

$$\sigma_j^{t+1} = \begin{cases} \tilde{\sigma}_j^{t+1} & \text{if } \log p\left(Y_j|\theta^{t+1}, \tilde{\sigma}_j^{t+1}\right) > \log p\left(Y_j|\theta^{t+1}, \sigma_j^t\right) \\ \sigma_j^t & \text{else.} \end{cases} \tag{4.20}$$

Although it may seem that we have replaced the task of selecting $\boldsymbol{\sigma}$ with that of selecting $\left\{\boldsymbol{\sigma}^{\mathbf{0}}, \alpha_\sigma, \sigma^\infty\right\}$, we claim that the latter is easy to select and provide both theoretical (Section 4.7) and experimental (Section 4.8) validation for this claim. The main benefit of the proposed approach is that it essentially traverses a continuous space of candidate $\boldsymbol{\sigma}$ without explicitly performing a grid search, which would be intractable as J grows. The parameter σ^∞ can be set arbitrarily small and $\boldsymbol{\sigma}^{\mathbf{0}}$ can be set arbitrarily large. The only recommendation we make is to set $\sigma_j^0 > \sigma_{j'}^0$ if modality j is a-priori believed to be more noisy than modality j'.

Section 4.7 studies the motivation for and properties of the annealing strategy in greater detail. In practice, the computation of $p\left(Y_j|\theta^{t+1}, \tilde{\sigma}_j^{t+1}\right)$ is costly because it requires the computation of the sufficient statistics in (4.11) for all L data points and J modalities. Instead, we replace the condition in (4.20) by checking whether decreasing σ_j should increase $p\left(Y_j|\theta^{t+1}, \sigma^t\right)$. This check is performed by checking the sign of the first derivative of $p\left(Y_j|\theta^{t+1}, \sigma^t\right)$, replacing (4.20)

Require: $\boldsymbol{Y}, \boldsymbol{\sigma}^0, \sigma^\infty, \alpha_\sigma$
1: **while** $\boldsymbol{\sigma}$ not converged **do**
2: **while** $\boldsymbol{D}$ not converged **do**
3: **for** $i \in [L]$ **do**
4: Update $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{\boldsymbol{i}}, \boldsymbol{\mu}^{\boldsymbol{i}}, \boldsymbol{\gamma}^{\boldsymbol{i}}$ using (4.13), (4.14), and (4.15)
5: **end for**
6: $\left\{ \text{ Update } D_j \text{ using (4.16) if } \sigma_j \text{ not converged} \right\}_{j=1}^{J}$
7: **end while**
8: $\left\{ \text{Update } \sigma_j \text{ using (4.21) if } \sigma_j \text{ not converged} \right\}_{j=1}^{J}$
9: **end while**

Figure 4.2: MSBDL algorithm for the one-to-one prior in (4.8).

with

$$\sigma_j^{t+1} = \begin{cases} \tilde{\sigma}_j^{t+1} & \text{if } \partial \log p\left(Y_j | \theta^{t+1}, \sigma_j^t\right) / \partial \sigma_j^t < 0 \\ \sigma_j^t & \text{else.} \end{cases} \tag{4.21}$$

It can be shown that (4.21) can be computed essentially for free by leveraging the sufficient statistics computed in the θ update step[2]. The complete MSBDL algorithm is summarized in Fig. 4.2. In practice, each D_j is normalized to unit ℓ_2 column norm at each iteration to prevent scaling instabilities.

4.3.1 Dictionary Cleaning

We adopt the methodology in [10] and "clean" each D_j every T iterations. Cleaning D_j means removing atoms which are aligned with one or more other atoms and replacing the removed atom with the element of Y_j which has the poorest reconstruction under D_j. A given atom is also replaced if it does not contribute to explaining Y_j, as measured by the energy of the corresponding row of U_j.

[2]See Supplemental Material.

4.4 Scalable Learning

When L is large, it is impractical to update the sufficient statistics for all data points at each EM iteration. To draw a parallel with stochastic gradient descent (SGD), when the objective function is a sum of functions of individual data points, one can traverse the gradient with respect to a randomly chosen data point at each iteration instead of computing the gradient with respect to every sample in the dataset. In the dictionary learning community, SGD is often the optimization algorithm of choice because the objective function is separable over each data point [17, 12, 32]. In addition, as N_j grows, the computation of the sufficient statistics in (4.13)-(4.14) can become intractable. In the following, we propose to address these issues using a variety of modifications of the MSBDL algorithm from Section 4.2. The proposed methods also apply to priors other than the one in (4.8) [3].

4.4.1 Scalability with respect to the size of the dataset

In the following, we present two alternatives to the EM MSBDL algorithm to achieve scalability with respect to L. The first proposed approach, referred to here as Batch EM, computes sufficient statistics only for a randomly chosen subset $\phi = \{i^1, \cdots, i^{L_0}\}$ at each EM iteration, where L_0 denotes the batch size. The M-step consists of updating $\{\gamma^i\}_{i\in\phi}$ using (4.15) and updating D_j using (4.16), with the exception that only the sufficient statistics from $i \in \phi$ are employed.

Another stochastic inference alternative is called Incremental EM, which is reviewed in the Appendix [118]. In the context of MSBDL, Incremental EM is tantamount to an inference procedure which, at each iteration, randomly selects a subset ϕ of points and updates the sufficient statistics in (4.13)-(4.14) for $i \in \phi$. During the M-step, the hyperparameters $\{\gamma^i\}_{i\in\phi}$ are updated. The dictionaries D_j are updated using (4.16), where the update rule depends on sufficient statistics computed for all L data points, only a subset of which have been updated during the given iteration.

[3] See Section 4.5

Table 4.1: CC, MC, L_0, and $\lceil N \rceil$ denote worst case computational complexity, worst case memory complexity, batch size, and a quantity which is upperbounded by N, respectively.

	EM Type	μ/Σ_x	CC	MC
MSBDL	Exact	Exact	LN^3	LM^2
MSBDL-1	Incremental	Exact	L_0N^3	LM^2
MSBDL-2	Incremental	Approximate	$L_0N^2\lceil N \rceil$	LM
MSBDL-3	Batch	Exact	L_0N^3	L_0M^2
MSBDL-4	Batch	Approximate	$L_0N^2\lceil N \rceil$	L_0M

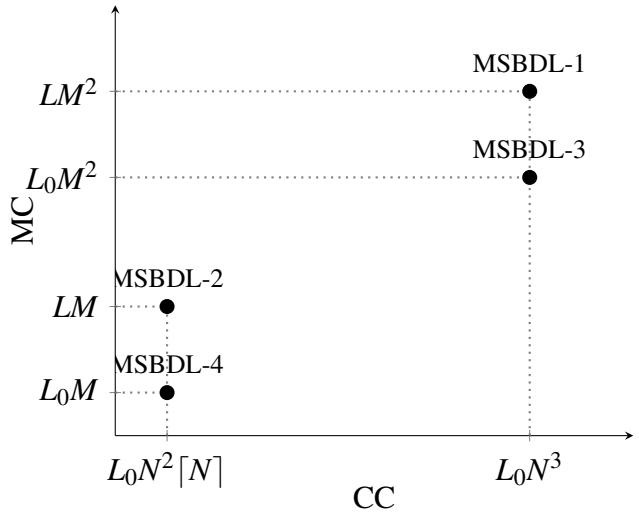


Figure 4.3: Visualization of worst-case computational and memory complexity per modality and EM iteration of the proposed approaches.

The taxonomy of MSBDL algorithms is visualized in Fig. 4.3.

4.4.2 Scalability with respect to the size of the dictionary

In order to avoid the inversion of the $N_j \times N_j$ matrix[4] required to compute (4.13), we use the conjugate gradient algorithm to compute μ_j^i and approximate $\Sigma_{x,j}^i$ by

$$\Sigma_{x,j}^i \approx \left(diag\left(\sigma^{-2} D_j^T D_j + \left(\Gamma^i\right)^{-1} \right) \right)^{-1} \tag{4.22}$$

where, in this case, $diag(\cdot)$ denotes setting the off-diagonal elements of the input to 0 [119, 120].

[4]Due to the matrix inversion lemma, (4.13) can be computed using $\Gamma^i - \Gamma^i D_j^T \left(\sigma_j^2 \mathrm{I} + D_j \Gamma^i D_j^T \right)^{-1} D_j \Gamma^i$.

Figure 4.4: Prototype branches for the atom-to-subspace (4.4a) and hierarchical sparsity (4.4b) models.

Table 4.1 shows a taxonomy of MSBDL algorithms considered using exact or incremental EM and exact or approximate computation of (4.13)-(4.14), along with the corresponding worst case computational and memory complexity per EM iteration. The Appendix provides a visualization of the difference between the proposed algorithms.

4.5 Modeling More Complex Relationships

The drawback of the one-to-one prior in (4.8) is that it constrains $M_j = M \; \forall j$. In the following, we propose two models which allow for M_j to be modality-dependent. For ease of exposition, we set $J = 2$, but the models we describe can be readily expanded to $J > 2$. We propose to organize $\boldsymbol{x}$ into a tree with K disjoint branches. We adopt the convention that the elements of x_1 and x_2 form the roots and leaves of the tree, respectively[5]. The root of the k'th branch is $x_1[k]$ and the leaves are indexed by $\mathscr{T}^k \subseteq [M_2]$. The defining property of the models we propose is the relationship between the sparsity pattern of the root and leaf levels.

4.5.1 Atom-to-subspace sparsity

The one-to-one prior in (4.8) can be viewed as linking the one-dimensional subspaces spanned by d_1^m and d_2^m for $m \in [M]$. Whenever d_1^m is used to represent y_1, d_2^m is used to represent

[5]We adopt this convention without loss of generality since the modalities can be re-labeled arbitrarily.

y_2, and vice-versa. The extension to the multi-dimensional subspace case stipulates that if d_1^k is used to represent y_1, then $\left\{d_2^m\right\}_{m\in\mathscr{T}^k}$ is used to represent y_2, and vice-versa. This model does not constrain $|\mathscr{T}^k|$ to be the same for all k, such that M_2 can be chosen independently of M_1.

Let $\gamma_B \in \mathbb{R}_+^K$, in contrast to Section 4.2 where $\gamma \in \mathbb{R}_+^M$. We encode the atom-to-subspace sparsity prior by assigning a single hyperparameter $\gamma_B[k]$ to each branch k. The distribution $p(\boldsymbol{x}|\gamma_B)$ is then given by

$$\prod_{k\in[K]} \mathsf{N}(x_1[k];0,\gamma_B[k]) \prod_{m\in\mathscr{T}^k} \mathsf{N}(x_2[m];0,\gamma_B[k]). \tag{4.23}$$

The marginal prior on $\boldsymbol{x}$ under $\gamma_B[k] \sim \mathsf{IGa}\left(\frac{\tau}{2},\frac{\tau}{2}\right)$ takes the form:

$$p(\boldsymbol{x}) = \prod_{k\in[K]} ST\left(\left\|\begin{bmatrix} x_1[k] \\ x_2\left[\mathscr{T}^k\right] \end{bmatrix}\right\|_2^2 ;\tau\right) \tag{4.24}$$

where $x_2\left[\mathscr{T}^k\right]$ is shorthand for the the elements of x_2 indexed by $\mathscr{T}^k$ and $ST(\cdot)$ denotes the Student's-t distribution.

Fig. 4.4a shows a prototype branch under the atom-to-subspace prior. Inference for the prior in (4.23) proceeds in much the same way as in Section 4.2.1. The form of the marginal likelihood in (4.10) and posterior in (4.12) remain the same, with the exception that $\Sigma_{y,j}^i$ and $\Sigma_{x,j}^i$ are re-defined to be

$$\begin{aligned} \Sigma_{y,j}^i &= \sigma_j^2 \mathsf{I} + D_j \Gamma_j^i D_j^T \\ \Sigma_{x,j}^i &= \left(\sigma_j^{-2} D_j^T D_j + \left(\Gamma_j^i\right)^{-1}\right)^{-1}, \Gamma_1^i = diag(\gamma_B^i), \end{aligned} \tag{4.25}$$

where Γ_2^i is a diagonal matrix whose $[m,m]$'th entry is $\gamma_B^i[k]$ for $m \in \mathscr{T}^k$. The update of γ_B^i is

given by

$$\frac{\Sigma^i_{x,1}[m,m] + \left(\mu^i_1[m]\right)^2 + \sum_{m\in\mathscr{T}^k} \Sigma^i_{x,2}[m,m] + \left(\mu^i_2[m]\right)^2}{1+|\mathscr{T}^k|},$$

while the update of D_j remains identical to (4.16). There are many ways to extend the atom-to-subspace prior for $J > 2$, depending on the specific application. One possibility is to simply append more branches to the root at $x_1[k]$ corresponding to coefficients from modalities $j > 2$.

One problem is that, for $|\mathscr{T}^k| > 1$, the atoms of D_2 indexed by $\mathscr{T}^k$ are not identifiable. The reason for the identifiability issue is that D_2 appears in the objective function in (4.10) only through the $D_2\Gamma_2 D_2^T$ term in (4.25), which can be written as

$$D_2\Gamma_2 D_2^T = \sum_{k\in[K]} \gamma_B[k] \sum_{m\in\mathscr{T}^k} d_2^m \left(d_2^m\right)^T. \tag{4.26}$$

Therefore, any $D_2^{'}$ which satisfies $\sum_{m\in\mathscr{T}^k} d_2'^m \left(d_2'^m\right)^T = \sum_{m\in\mathscr{T}^k} d_2^m \left(d_2^m\right)^T$ for all k achieves the same objective function value as D_2. Since the objective function is agnostic to the individual atoms of D_2, the performance of this model is severely upper-bounded in terms of the ability to recover D_2. In the following, we propose an alternative model which circumvents the identifiability problem.

4.5.2 Hierarchical Sparsity

In this section, we propose a model which allows the root of each branch to control the sparsity of the leaves, but not vice-versa. Specifically, we stipulate that if $x_1[k] = 0$, then $x_2[m] = 0\ \forall m \in \mathscr{T}^k$. Hierarchical sparsity was first studied in [121, 122, 123] and later incorporated into a unimodal dictionary learning framework in [124]. Later, Bayesian hierarchical sparse signal recovery techniques were developed, which form the basis for the following derivation [125, 126].

From an optimization point of view, hierarchical sparsity can be promoted through a

composite regularizer [123]. In this case, the regularizer could[6] take the form

$$\sum_{k\in[K],m\in\mathscr{T}^k}\left\|\begin{bmatrix}x_1[k] & x_2[m]\end{bmatrix}\right\|_2+|x_2[m]|. \tag{4.27}$$

As described in [123], the key to designing a composite regularizer for a given root-leaf pair is to measure the group norm of the pair along with the energy of the leaf alone. The combination of the group and individual norms serve two purposes which, jointly, promote hierarchical sparsity [123]:

1. it is possible that $x_2[m]=0, m\in\mathscr{T}^k$, without requiring $x_1[k]=0$, and
2. the infinitesimal penalty on $x_1[k]$ deviating from 0 tends to 0 for $|x_2[m]|>0, m\in T^k$.

In a Bayesian setting, we can mimic the effect of the regularizer in (4.27) through an appropriately defined prior on $\boldsymbol{x}$. Let

$$\tilde{x}_j=S_jx_j,\ \ S_1\in\mathbb{B}^{M_2\times M_1},\ \ S_2=\begin{bmatrix}\mathsf{I} & \mathsf{I}\end{bmatrix}^T\in\mathbb{B}^{2M_2\times M_1} \tag{4.28}$$

where S_1 is a binary matrix such that $S_1[m,k]=1$ if and only if $m\in\mathscr{T}^k$. Let R_j be a diagonal matrix such that $S_j^TS_jR_j=\mathsf{I}$[7] and define $\hat{x}_j=S_jR_jx_j$. Let $\gamma_j\in\mathbb{R}^{M_2}\ \forall j$ and

$$p(\boldsymbol{x}|\gamma)=\mathsf{N}(\hat{x}_1;0,\Gamma_1)\,\mathsf{N}(\hat{x}_2;0,\Gamma_2), \tag{4.29}$$

where $\Gamma_1=diag(\gamma_1)$ and $\Gamma_2=diag\left(\begin{bmatrix}\gamma_1^T & \gamma_2^T\end{bmatrix}^T\right)$. Marginalizing over $\boldsymbol{\gamma_j}$ for $\gamma_j[m]\sim\mathsf{IGa}\left(\frac{\tau}{2},\frac{\tau}{2}\right)$,

[6] The exact form of the regularizer depends on how the energy in a given group is measured.

[7] A diagonal R_j is guaranteed to exist because $S_j^TS_j$ is itself a diagonal matrix.

the prior on $\boldsymbol{x}$ turns out to be

$$p(\boldsymbol{x}) = \prod_{k\in[K],m\in\mathscr{T}^k} ST\left(\left\|\begin{bmatrix} x_1[k] \\ x_2[m] \end{bmatrix}\right\|_2^2;\tau\right) ST\left(x_2[m]^2;\tau\right). \tag{4.30}$$

A prototype branch for the hierarchical sparsity prior is shown in Fig. 4.4b. To see how this model leads to hierarchical sparsity, observe that

$$\begin{aligned} p(x_1[k]|\gamma) &= \mathsf{N}(0,\gamma_1[k]) \\ p(x_2[m]|\gamma) &= \mathsf{N}\left(0,\left(\gamma_1^{-1}[k]+\gamma_2^{-1}[m]\right)^{-1}\right) \end{aligned} \tag{4.31}$$

for $m \in \mathscr{T}^k$. If we infer that $\gamma_1[k] = 0$, then the prior on both $x_1[k]$ and $x_2[m], \forall m \in \mathscr{T}^k$, reduces to a dirac-delta function, i.e. if the root is zero, then the leaves must also be zero. On the other hand, if $\gamma_2[m]$ is inferred to be 0, only the prior on $\gamma_2[m]$ is affected, i.e. leaf sparsity does not imply root sparsity.

Inference for the tree-structured model proceeds in a similar fashion to that shown in Section 4.2.1, with a few variations. The goal is to optimize (4.9) through the EM algorithm. The difference here is that we use $\left\{\hat{\boldsymbol{X}},\boldsymbol{Y},\theta\right\}$ and $\hat{\boldsymbol{X}}$ as the complete and nuisance data, respectively. In order to carry out inference, we must first find the posterior density $p\left(\hat{\boldsymbol{X}}|\boldsymbol{Y},\theta\right)$. It is helpful to first derive the signal model in terms of $\hat{\boldsymbol{x}}$ [125]:

$$p\left(y_j|D_j,\hat{x}_j,\sigma_j\right) = \mathsf{N}\left(A_j S_j^T \hat{x}_j, \sigma_j^2 \mathsf{I}\right). \tag{4.32}$$

Using (4.32), it can be shown that

$$p\left(\hat{x}_j^i|y_j^i,\theta\right) = \mathsf{N}\left(\mu_{\hat{x},j}^i, \Sigma_{\hat{x},j}^i\right) \tag{4.33}$$

$$\Sigma_{\hat{x},j}^i = \left(\sigma_j^{-2} S_j D_j^T D_j S_j^T + \left(\Gamma_j^i\right)^{-1}\right)^{-1} \tag{4.34}$$

$$\mu_{\hat{x},j}^i = \sigma_j^{-2} \Sigma_{\hat{x},j}^i D_j S_j^T y_j. \tag{4.35}$$

The likelihood function itself is different from (4.10)-(4.11) and is given by $p\left(y_j^i|\theta\right) = \mathsf{N}\left(0, \Sigma_{\hat{y},j}^i\right)$, where $\Sigma_{\hat{y},j}^i = \sigma_j^2 \mathsf{I} + D_j S_j^T \Gamma_j^i S_j D_j^T$. The EM update rules are given by

$$\left(\gamma_j^i[m]\right)^{t+1} = \begin{cases} 0.5 \sum_{j'=1}^{2} \Sigma_{\hat{x},j'}^i[m,m] + \left(\mu_{\hat{x},j'}^i[m]\right)^2 & \text{if } m \leq M_2 \\ \Sigma_{\hat{x},j}^i[m,m] + \left(\mu_{\hat{x},j}^i[m]\right)^2 & \text{else.} \end{cases} \tag{4.36}$$

$$\left(D_j\right)^{t+1} = Y_j U_j^T S_j \left(S_j^T \left(U_j U_j^T + \sum_{i\in[L]} \Sigma_{\hat{x},j}^i\right) S_j\right)^{-1}. \tag{4.37}$$

Extending the hierarchical sparsity prior for $J > 2$ is straightforward and depends on the specific application being considered. It is possible to simply append more leaves to each root $x_1[k]$ corresponding to coefficients from modality $j > 2$. Another possibility is to treat each $x_2[m]$ as itself a root with leaves from x_3, assigning a hyperparameter to each x_2-x_3 root-leaf pair as well as to each x_3 leaf, and repeating the process until all modalities are incorporated into the tree.

4.5.3 Avoiding Poor Stationary Points

For both the atom-to-subspace and hierarchical sparsity models, we experimentally observe that MSBDL tends to get stuck in undesirable stationary points. In the following, we describe the behavior of MSBDL in these situations and offer a solution. Suppose that data is generated according to the atom-to-subspace model, where D_j denotes the true dictionary for modality j. In this scenario, we experimentally observe that MSBDL performs well when

Require: $D, U, \{\mathscr{T}^k\}_{k\in[K]}, M_p, S, \varepsilon$

Let $z[m] = \left\| \left(S^T U\right)[m,:] \right\|_2^2 \ \forall m$ and

$$v[k] = \underset{m_1,m_2\in\mathscr{T}^k \text{ s.t. } m_1\neq m_2}{\arg\max} \frac{\left|\left(d^{m_1}\right)^T d^{m_2}\right|}{\|d^{m_1}\|_2 \|d^{m_2}\|_2}$$

while $\exists k$ s.t. $v[k] > \varepsilon$ and $|\mathscr{T}^k| > 1$ and $M_p > 0$ **do**
    $k = \arg\max_k v[k]$
    Find $m = \arg\min_{m\in\mathscr{T}^k} z[m]$
    Remove column m from D and $\mathscr{T}^k$
    $M_p \leftarrow M_p - 1$
end while
while $M_p > 0$ **do**
    Find $m = \arg\min_{m\in\mathscr{T}^k \text{ s.t. } |\mathscr{T}^k|>1} z[m]$
    Remove column m from D and $\mathscr{T}^k$
    $M_p \leftarrow M_p - 1$
end while

Figure 4.5: Pruning algorithm for the learning strategy in Section 4.5.3. M_p denotes the number of columns to be pruned, S is given by the identity matrix for the atom-to-subspace model and by S_2 in (4.28) for the hierarchical sparsity model, and U denotes the matrix of first order sufficient statistics.

$|\mathscr{T}^k| = c \ \forall k$. On the other hand, if $|\mathscr{T}^k|$ varies as a function of k, MSBDL tends to get stuck in poor stationary points, where the quality of a stationary point is (loosely) defined next. Let $|\mathscr{T}^k| = 1$ for all k except k', for which $|\mathscr{T}^{k'}| = 2$, i.e. $M_2 = M_1 + 1$. In this case, MSBDL is able to recover D_1, but recovers only the atoms of D_2 which are indexed by $\mathscr{T}^k \ \forall k \neq k'$.

To avoid poor stationary points, we adopt the following strategy. If the tree describing the assignment of columns of D_2 to those of D_1 is unbalanced, i.e. $|\mathscr{T}^k|$ varies with k, then we first balance the tree by adding additional leaves[8]. Let $\hat{M}_2$ be the number of leaves in the balanced tree. We run MSBDL until convergence to generate $\hat{\boldsymbol{D}}$. Finally, we prune away $\hat{M}_2 - M_2$ columns of $\hat{D}_2$ using the algorithm in Fig. 4.5.

[8] A balanced tree is one which has the same number of leaves for each subtree, or $|\mathscr{T}^k| = c$.

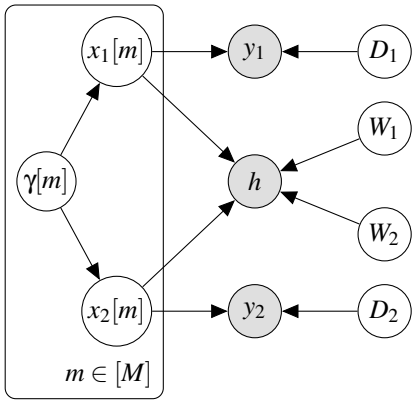


Figure 4.6: Graphical model for two modality TD-MSBDL.

4.6 Task Driven MSBDL (TD-MSBDL)

In the following, we describe a task-driven extension of the MSBDL algorithm. For purposes of exposition, we assume the one-to-one prior in (4.8), but the approach applies equally to the priors discussed in Section 4.5, as discussed in Section 4.8. To incorporate task-driven learning, we modify the MSBDL graphical model to the one shown in Fig. 4.6. We set $p\left(h|x_j,W_j\right) = \mathsf{N}\left(W_j x_j, \beta_j^2 \mathsf{I}\right)$, where β_j is the class label noise standard deviation for modality j. The class label noise standard deviation is modality dependent, affording the model an extra level of flexibility compared to [32, 13, 12]. The choice of the Gaussian distribution for the conditional density of h, as opposed to a multinomial or softmax, stems from the fact that the posterior $p\left(x_j|y_j,h,D_j,W_j\right)$ needed to perform EM remains computable in closed form, i.e. $p\left(x_j^i|y_j^i,h^i,D_j,W_j\right) = \mathsf{N}\left(\Sigma_{x,j}^{TD,i},\mu_j^{TD,i}\right)$ where

$$\Sigma_{x,j}^{TD,i} = \left(\sigma_j^{-2} D_j^T D_j + \beta_j^{-2} W_j^T W_j + \left(\Gamma^i\right)^{-1}\right)^{-1} \tag{4.38}$$

$$\mu_j^{TD,i} = \Sigma_{x,j}^{TD,i}\left(\sigma_j^{-2} D_j^T y_j^i + \beta_j^{-2} W_j^T h^i\right) \tag{4.39}$$

4.6.1 Inference Procedure

We employ EM to optimize

$$\arg\max_{\theta^{TD}} \log p\left(\boldsymbol{Y}, H|\theta^{TD}\right) \tag{4.40}$$

where $\theta^{TD} = \{\theta, \boldsymbol{W}\}$. It can be shown that $p\left(y_j, h|x_j, \theta^{TD}\right) = \mathsf{N}\left(\begin{bmatrix} y_j^T & h^T \end{bmatrix}^T; 0, \Sigma_{y,j}^{TD}\right)$ where

$$\Sigma_{y,j}^{TD} = \begin{bmatrix} \sigma_j^2 \mathsf{I} + D_j \Gamma D_j^T & 0 \\ 0 & \beta_j \mathsf{I} + W_j \Gamma W_j^T \end{bmatrix}. \tag{4.41}$$

The update rules for $\boldsymbol{D}$ and $\{\gamma^i\}_{i=1}^L$ [9] remain identical to (4.15) and (4.16), respectively, with the exception that the modified posterior statistics shown in (4.38)-(4.39) are used. The update of W_j is given by

$$W_j^{t+1} = H\left(U_j^{TD}\right)^T \left(U_j^{TD}\left(U_j^{TD}\right)^T + \sum_{i\in[L]} \Sigma_{x,j}^{TD,i}\right)^{-1} \tag{4.42}$$

$$U_j^{TD} = \begin{bmatrix} \mu_j^{TD,1} & \cdots & \mu_j^{TD,L} \end{bmatrix}. \tag{4.43}$$

We also find it useful to add regularization in the form of $\nu \sum_{j\in[J]} \|W_j\|_F^2$, leading to the update rule

$$W_j^{t+1} = \begin{bmatrix} H\left(U_j^{TD}\right)^T & 0 \end{bmatrix} \left(\begin{bmatrix} U_j^{TD}\left(U_j^{TD}\right)^T + \sum_{i\in[L]} \Sigma_{x,j}^{TD,i} & \sqrt{\nu}\mathsf{I} \end{bmatrix}\right)^{-1}. \tag{4.44}$$

TD-MSBDL has the same worst-case computational complexity as MSBDL with the benefit of supervised learning.

[9] Assuming that the prior in (4.8) is used.

4.6.2 Complete Algorithm

Supervised learning algorithms are ultimately measured by their performance on test data. While a given algorithm may perform well on training data, it may generalize poorly to test data[10]. To maintain the generalization properties of the model, it is common to split the training data into a training set $\{\boldsymbol{Y}, H\}$ and validation set $\left\{\boldsymbol{Y}^V, H^V\right\}$, where the number of training points L does not necessarily have to equal the number of validation points L^V. The validation set is then used during the training process as an indicator of generalization, as summarized by the following rule of thumb: Continue optimizing θ^{TD} until performance on the validation set stops improving. In the context of TD-MSBDL, the concept of generalization has a natural Bayesian definition: The parameter set $\left\{\theta^{TD}, \boldsymbol{\beta}\right\}$ which achieves optimal generalization solves

$$\underset{\theta^{TD}\in\mathscr{H},\boldsymbol{\beta}}{\arg\max} \prod_{j\in[J]} p\left(H^V|\theta^{TD},\beta_j\right), \tag{4.45}$$

where $\mathscr{H}$ denotes the set of solutions to (4.40). Note that $p\left(H^V|\theta^{TD},\beta_j\right) = p\left(H^V|W_j,\beta_j\right)$, which is intractable to compute since it requires integrating $p\left(h^{V,i}|W_j,\gamma^{V,i},\beta\right) = \mathsf{N}\left(0,\beta_j\mathsf{I}+W_j\Gamma^{V,i}W_j^T\right)$ over $\gamma^{V,i}$, where $\Gamma^{V,i} = diag(\gamma^{V,i})$. As such, we approximate $p\left(H^V|W_j,\beta_j\right)$ by $p\left(H^V|W_j,\gamma^{*,V,i},\beta_j\right)$, where $\gamma^{*,V,i}$ is the output of MSBDL with fixed D_j for input data $y_j^{V,i}$, leading to the tractable optimization problem

$$\underset{\theta^{TD}\in\mathscr{H},\boldsymbol{\beta}}{\arg\max} \prod_{j\in[J]} p\left(H^V|W_j,\left\{\gamma^{*,V,i}\right\}_{i\in[L^V]},\beta_j\right). \tag{4.46}$$

What remains is to select $\boldsymbol{\beta}$. As β_j decreases, TD-MSBDL fits the parameters θ^{TD} to the training data to a larger degree, i.e. the optimizers of (4.40) achieve increasing objective function values. Since direct optimization of (4.46) over $\boldsymbol{\beta}$ presents the same challenges as the optimization of $\boldsymbol{\sigma}$, we propose an annealing strategy which proposes progressively smaller values of β_j until the

[10] In the supervised learning community, lack of generalization to test data is commonly referred to as over-fitting the training data.

objective in (4.46) stops improving:

$$\beta_j^{t+1} = \begin{cases} \tilde{\beta}_j^{t+1} & \text{if } \log p\left(H^V | W_j, \left\{\gamma^{*,V,i}\right\}_{i\in[L^V]}, \tilde{\beta}_j^{t+1}\right) > \log p\left(H^V | W_j, \left\{\gamma^{*,V,i}\right\}_{i\in[L^V]}, \beta_j^t\right) \\ \beta_j^t & \text{else} \end{cases} \tag{4.47}$$

where $\beta_j^0 > \beta^\infty \geq 0, \alpha_\beta < 1, \tilde{\beta}_j^{t+1} = \max(\beta^\infty, \alpha_\beta \beta_j^t)$, and we make the same recommendations for setting $\boldsymbol{\beta^0}$ and β^∞ as $\boldsymbol{\sigma^0}$ and σ^∞ in Section 4.3. Computing (4.47) is computationally intensive because it requires running MSBDL, so we only update $\log p\left(H^V | W_j, \left\{\gamma^{*,V,i}\right\}_{i\in[L^V]}, \beta_j^t\right)$ every T^V iterations. The complete TD-MSBDL algorithm is shown in Fig. 4.7. Given test data Y_j^{test}, we first run MSBDL with D_j fixed and treat $\mu_j^{test,i}$ as an estimate of $x_j^{test,i}$. The data is then classified according to $\arg\max_{c\in[C]} \left(W_j \mu_j^{test,i}\right)[c]$, where e_c refers to the c'th standard basis vector [32].

4.7 Analysis

We begin by analyzing the convergence properties of MSBDL. Note that MSBDL is essentially a block-coordinate descent algorithm with blocks θ and $\boldsymbol{\sigma}$. Therefore, we first analyze the θ update block, then the $\boldsymbol{\sigma}$ update block, and finally the complete algorithm. Unless otherwise specified, we assume a non-informative prior on γ and the one-to-one prior[11]. While MSBDL relies heavily on EM, it is not strictly an EM algorithm because of the $\boldsymbol{\sigma}$ annealing procedure. It will be shown that MSBDL still admits a convergence guarantee and an argument will be presented for why annealing $\boldsymbol{\sigma}$ produces favorable results in practice. Although we focus specifically on MSBDL, the results can be readily extended to TD-MSBDL, with details omitted for brevity. Proofs for all results are shown in the Appendix.

[11] Extension of Theorems 6-10 to the atom-to-subspace and hierarchical priors is straightforward, but omitted for brevity.

Require: $\boldsymbol{Y}, \boldsymbol{Y}^V, \boldsymbol{\sigma}^{\mathbf{0}}, \boldsymbol{\beta}^{\mathbf{0}}, \sigma^{\infty}, \beta^{\infty}, H, H^V, \alpha_{\sigma}, \alpha_{\beta}, T^V$

while $\boldsymbol{\beta}$ not converged **do**
  while $\boldsymbol{D}$ and $\boldsymbol{W}$ not converged **do**
    for $i \in [L]$ **do**
      Update $\boldsymbol{\Sigma}_x^{TD,i}$ using (4.38)
      Update $\boldsymbol{\mu}^{TD,i}$ using (4.39)
      Update γ^i using (4.15)
    end for
    $\left\{ \text{Update } D_j \text{ using (4.16) if } \sigma_j \text{ not converged} \right\}_{j=1}^{J}$
    $\left\{ \text{Update } W_j \text{ using (4.44) if } \beta_j \text{ not converged} \right\}_{j=1}^{J}$
  end while
  $\left\{ \text{Update } \sigma_j \text{ using (4.21) if } \sigma_j \text{ not converged} \right\}_{j=1}^{J}$
  if $\text{modulo}(t, T^V) = 0$ **then**
    for $j \in [J]$ **do**
      if β_j not converged **then**
        $\{\gamma^{*,V,i}\}_{i \in L^V} = MSBDL\left(Y_j^V, \sigma_j^0, \sigma_j, \sigma^{\infty}, \alpha_{\sigma}\right)$
        Update β_j using (4.47)
      end if
    end for
  end if
end while

Figure 4.7: Complete TD-MSBDL algorithm for the prior in (4.8).

We first prove that the set of iterates $\{\theta^t\}_{t=1}^{\infty}$ produced by the inner loop of MSBDL converge to the set of stationary points of the log-likelihood with respect to θ. This is established by proving that the objective function is coercive (Theorem 5), which means that $\{\theta^t\}_{t=1}^{\infty}$ admits at least one limit point (Corollary 6), and then proving that limit points are stationary (Theorem 6).

Theorem 5. *Let $\{\sigma_j > 0\}_{j=1}^{J}$, let there be at least one j^* for each m such that $\left\|d_{j^*}^m\right\|_2 > 0$, and let $\exists i$ such that $\gamma^i[m] > 0$ for any choice of m. Then, $-\log p(\boldsymbol{Y}|\theta,\boldsymbol{\sigma})$ is a coercive function of $\{\theta,\boldsymbol{\sigma}\}$.*

Corollary 6. *Let the conditions of Theorem 5 be satisfied. Then, the sequence $\{\theta^t\}_{t=1}^{\infty}$ produced by the inner loop of the MSBDL algorithm admits at least one limit point.*

Theorem 6. *Let the conditions of Corollary 6 be satisfied, D_j^t be full rank for all t and j, $\boldsymbol{\sigma}$ be fixed, and generate $\{\theta^t\}_{t=1}^{\infty}$ using the inner loop of MSBDL. Then, $\{\theta^t\}_{t=1}^{\infty}$ converges to the set of stationary points of $\log p(\boldsymbol{Y}|\theta,\boldsymbol{\sigma})$. Moreover, $\{\log p(\boldsymbol{Y}|\theta,\boldsymbol{\sigma})\}_{t=1}^{\infty}$ converges monotonically to $\log p(\boldsymbol{Y}|\theta^*,\boldsymbol{\sigma})$, for stationary point θ^*.*

The requirement that D_j^t be full rank is easily satisfied in practice as long as $L > N_j$. Note that the SBL algorithm in [27] is a special case of MSBDL for fixed $\boldsymbol{D}$ and $J = 1$. To the best of our knowledge, Theorem 6 represents the first result in the literature guaranteeing the convergence of SBL. A similar result to Theorem 6 can be given in the stochastic EM regime.

Theorem 7. *Let $\boldsymbol{\sigma}$ be fixed, U_j^t be full rank for all t, j, and generate $\{\theta^t\}_{t=1}^{\infty}$ using the inner loop of MSBDL, only updating the sufficient statistics for a batch of points at each iteration (i.e. incremental EM). Then, the limit points of $\{\theta^t\}_{t=1}^{\infty}$ are stationary points of $\log p(\boldsymbol{Y}|\theta,\boldsymbol{\sigma})$.*

If we consider the entire MSBDL algorithm, i.e. including the update of $\boldsymbol{\sigma}$, we can still show that MSBDL is convergent.

Corollary 7. *Let the conditions of Theorems 5 and 6 be satisfied. Then, the sequence $\left\{\theta^t, \boldsymbol{\sigma}^t\right\}_{t=1}^{\infty}$ produced by the MSBDL algorithm admits at least one limit point.*

Although the preceding results establish that MSBDL has favorable convergence properties, the question still remains as to why we choose to anneal σ_j. At a high level, it can be argued that setting σ_j to a large value initially and then gradually decreasing it prevents MSBDL from getting stuck in poor stationary points with respect to θ at the beginning of the learning process. To motivate this intuition, consider the log likelihood function in (4.9), which decomposes into a sum of J modality-specific log-likelihoods. The curvature of the log-likelihood for modality j depends directly on σ_j. Setting a high σ_j corresponds to choosing a relatively flat log likelihood surface, which, from an intuitive point of view, has less stationary points. This intuition can be formalized in the scenario where D_j is constrained in a special way.

Theorem 8. *Let* $\sigma_j^1 > \sigma_j^2, \Psi_j = \left\{ D_j : D_j = \begin{bmatrix} \check{D}_j & | \end{bmatrix} \right\}$, *and*

$$\Omega_{\sigma_j,j} = \left\{ \Sigma_{y,j} : \Sigma_{y,j} = \sigma_j^2 \mathrm{I} + D_j \Gamma {D_j}^T, D_j \in \Psi_j \right\}. \tag{4.48}$$

Then, $\Omega_{\sigma_j^1,j} \subseteq \Omega_{\sigma_j^2,j}$.

Theorem 8 suggests that as σ_j gets smaller, the space over which the log-likelihood in (4.9) is optimized grows. As the optimization space grows, the number of stationary points grows as well[12]. As a result, it may be advantageous to slowly anneal σ_j in order to allow MSBDL to learn D_j without getting stuck in a poor stationary point.

If we constrain the space over which D_j is optimized to Ψ_j, as in Theorem 8, then we can establish a number of interesting convergence results.

Theorem 9. *Let* α_σ *be arbitrarily close to* 1, $\sigma^\infty = 0$, $\sigma_j^0 \geq \arg\max_{\sigma_j} \max_\theta \log p\left(Y_j|\theta,\sigma_j\right)$, $D_j \in \Psi_j$, $\theta^t = \arg\max_\theta \log p\left(Y_j|\theta,\sigma_j^{t-1}\right)$, *and consider updating* σ_j *using* (4.20). *Then,* $\sigma_j^t = \sigma_j^{t-1}$ *implies* $\sigma_j^t = \arg\max_{\sigma_j} \log p\left(Y_j|\theta^t,\sigma_j\right)$.

Theorem 9 states that, under certain conditions, annealing terminates for a given modality

[12] This holds only for the constrained scenario in Theorem 8.

j at a global maximum of the log-likelihood with respect to σ_j for fixed θ^t. The conditions of Theorem 9 ensure that (4.20) only terminates at stationary points of the log-likelihood.

Theorem 10. *Let the conditions of Theorems 6 and 9 and Corollary 7 be satisfied. Then, the sequence* $\left\{\theta^t, \boldsymbol{\sigma}^t\right\}_{t=1}^{\infty}$ *produced by MSBDL converges to the set of stationary points of* $\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right)$.

Convergence results like Theorem 10 cannot be established for K-SVD and Jℓ_0DL because they rely on greedy search techniques. In addition, no convergence results are presented in [32] for Jℓ_1DL.

Finally, we consider what guarantees can be given for dictionary recovery in the noiseless setting, i.e. $Y_j = D_j X_j \ \forall j$. We assume $M_j = M \ \forall j$ and that $\boldsymbol{x^i}$ share a common sparsity profile for all i. We do not claim that the following result applies to more general cases when $M_j \neq M \ \forall j$ or different priors. The question we seek to answer is: Under what conditions is the factorization $Y_j = D_j X_j$ unique? Due to the nature of dictionary learning, uniqueness can only be considered up to permutation and scale. Two dictionaries D^1 and D^2 are considered equivalent if $D^1 = D^2 Z$, where Z is a permutation and scaling matrix. Guarantees of uniqueness in the unimodal setting were first studied in [127]. The results relied on several assumptions about the data generation process.

Assumption 1. *Let* $s = \|x_j^i\|_0 \ \forall i, j$, *and* $s < \frac{spark(D_j)}{2}$, *where* $spark(D_j)$ *is the minimum number of columns of* D_j *which are linearly dependent [128]. Each* x_j^i *has exactly s non-zeros.*

Assumption 2. Y_j *contains at least* $s+1$ *different points generated from every combination of s atoms of* D_j.

Assumption 3. *The rank of every group of* $s+1$ *points generated by the same s atoms of* D_j *is exactly s. The rank of every group of* $s+1$ *points generated by different atoms of* D_j *is exactly* $s+1$.

Lemma 1 (Theorem 3, [127]). *Let Assumptions 1-3 be true. Then, Y_j admits a unique factorization $D_j X_j$. The minimum number of samples required to guarantee uniqueness is given by $(s+1)\binom{M}{s}$.*

Treating the multimodal dictionary learning problem as J independent unimodal dictionary learning problems, the following result follows from Lemma 1.

Corollary 8. *Let Assumptions 1-3 be true for all j. Then, Y_j admits a unique factorization $D_j X_j$ for all j. The minimum number of samples required to guarantee uniqueness is given by $J(s+1)\binom{M}{s}$.*

As the experiments in Section 4.8 will show, there are benefits to jointly learning multimodal dictionaries. It is therefore interesting to inquire whether or not there are *provable* benefits to the multimodal dictionary learning problem, at least from the perspective of the uniqueness of factorizations. To formalize this intuition, consider the scenario where some data points i do not have data available for all modalities. Let the Boolean matrix $P \in \mathbb{B}^{J \times L}$ be defined such that $P[j,i]$ is 1 if data for modality j is available for instance i and 0 else. The conditions on the amount of data needed to guarantee unique recovery of $\boldsymbol{D}$ by Corollary 8 can be restated as $L = (s+1)\binom{M}{s}$ and $P[j,i] = 1 \; \forall j,i$. The natural question to ask next is: Can uniqueness of factorization be guaranteed if $P[j,i] = 0$ for some (j,i)?

Theorem 11. *Let $\boldsymbol{x^i}$ share a common sparsity profile for all i and Assumptions 1-2 be true for all j. Let Assumption 3 be true for a single j^*. Let $G_j^k = \left\{ i : P[j,i] = 1 \text{ and } y_j^i \in \text{span}\left(D_j[:, \Upsilon^k]\right) \right\}$ where Υ^k is the k'th subset of size s of $[M]$. Let $|G_j^k \cap G_{j^*}^k| \geq s$ for all $j \neq j^*$ and k. Then, the factorization $Y_j = D_j X_j$ is unique for all j. The minimum total number of data points required to guarantee uniqueness is given by $J\left(s + \frac{1}{J}\right)\binom{M}{s}$.*

Theorem 11 establishes that the number of samples required to guarantee a unique solution to the multimodal dictionary learning problem is *strictly less* than in Corollary 8.

4.8 Results

4.8.1 Synthetic Data Dictionary Learning

To validate how well MSBDL is able to learn unimodal and multimodal dictionaries, we conducted a series of experiments on synthetic data. We adopt the setup from [129] and generate ground-truth dictionaries $D_j \in \mathbb{R}^{20\times 50}$ by sampling each element from a $\mathsf{N}(0,1)$ distribution and scaling the resulting matrices to have unit ℓ_2 column norm. We then generate x_j^i by randomly selecting $s = 5$ indices and generating the non-zero entries by drawing samples from a $\mathsf{N}(0,1)$ distribution. The supports of $\boldsymbol{x^i}$ are constrained to be the same, while the coefficients are not. The elements of v_j^i are generated by drawing samples from a $\mathsf{N}(0,1)$ distribution and scaling the resulting vector in order to achieve a specified Signal-to-Noise Ratio (SNR). We use $L = 1000$ and simulate both bimodal and trimodal datasets. The bimodal dataset consists of 30dB ($j = 1$) and 10dB ($j = 2$) SNR modalities. The trimodal dataset consists of 30 dB ($j = 1$), 20 dB ($j = 2$), and 10 dB ($j = 3$) SNR modalities. We use the empirical probability of recovering $\boldsymbol{D}$ as the measure of success, which is given by

$$\frac{1}{M_j}\sum_{m\in[M_j]} \mathbb{1}\left[\mathfrak{l}\left(d_j^m, \hat{D}_j\right) > 0.99\right] \tag{4.49}$$

$$\mathfrak{l}\left(d_j^m, \hat{D}_j\right) = \max_{1\le m'\le M_j} \frac{\left|\left(d_j^m\right)^T \hat{d}_j^{m'}\right|}{\left\|d_j^m\right\|_2 \left\|d_j^{m'}\right\|_2}, \tag{4.50}$$

where $\hat{D}_j$ denotes the output of the dictionary learning algorithm and $\mathbb{1}[\cdot]$ denotes the indicator function. The experiment is performed 50 times and averaged results are reported. We compare MSBDL with ℓ_1DL[13], K-SVD[14], Jℓ_0DL, and Jℓ_1DL. While code for Jℓ_1DL was publicly available, it could not be run on any of our Windows or Linux machines, so we used our own implementation.

[13] http://spams-devel.gforge.inria.fr/downloads.html
[14] http://www.cs.technion.ac.il/~ronrubin/software.html

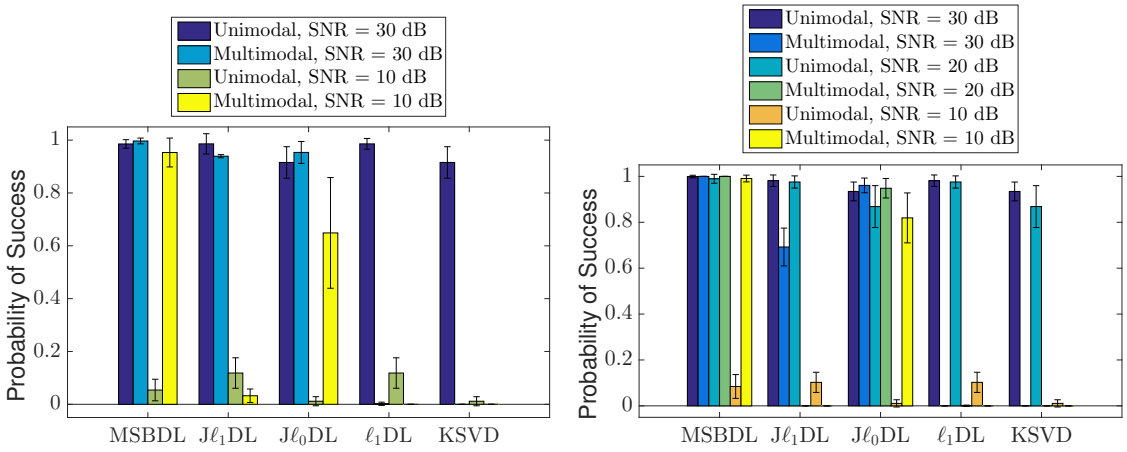


Figure 4.8: Bimodal (4.8a) and trimodal (4.8b) synthetic data results with one standard deviation error bars.

Code for Jℓ_0DL was not publicly available, so we used our own implementation. For all algorithms, the batch size was set to L.

For the bimodal setting, the parameters σ_1^0 and σ_2^0 were set to 1 and 10, respectively. For the trimodal setting, the parameters σ_1^0, σ_2^0, and σ_3^0 were set to $1, 1.5$, and 2, respectively. In both cases, we set $\alpha_\sigma = 0.995$ and $\sigma^\infty = 1e-3$, where this choice of σ^∞ corresponds to the lowest candidate λ in the cross-validation procedure for competing algorithms. It was experimentally determined that MSBDL is relatively insensitive to the choice of σ_j^0 as long as $\sigma_1^0 < \sigma_2^0 < \sigma_3^0$, thus obviating the need to cross-validate these parameters. The regularization parameters λ in (4.1) and $\boldsymbol{\lambda}$ in (4.3) were selected by a grid search over $\{1e-3, 1e-2, 1e-1, 1\}$ and both K-SVD and J0DL were given the true s. The parameter λ_1 was set to 1 for Jℓ_0DL across all experiments because the objective function in (4.3) depends only on the relative weighting of modalities. All algorithms were run until convergence.

The bimodal and trimodal dictionary recovery results are shown in Fig.'s 4.8a and 4.8b, respectively. For unimodal data, all of the algorithms recover the true dictionary almost perfectly when the SNR $= 30$ dB, with the exception of J0DL and K-SVD. All of the tested algorithms perform relatively well for data with 20 dB SNR and poorly on data with 10 dB SNR, although

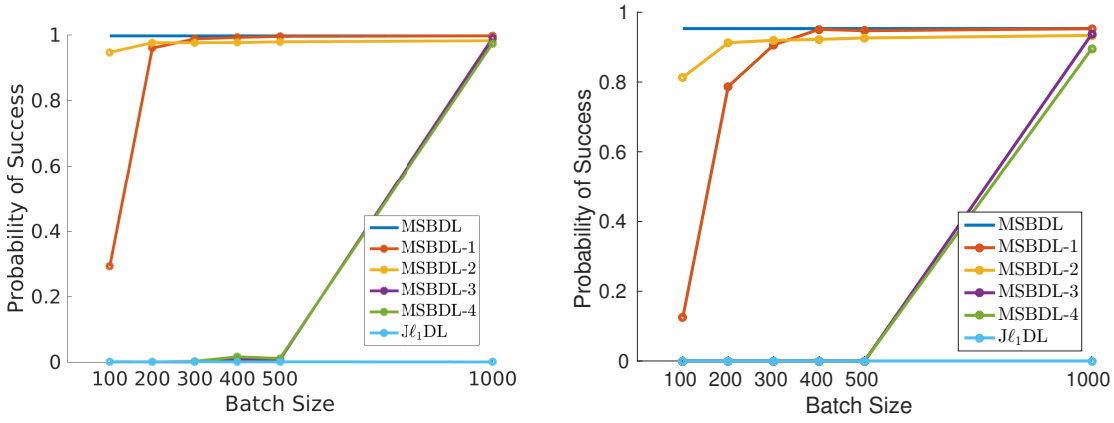


Figure 4.9: Bimodal synthetic data results using stochastic learning for 30 dB (Fig. 4.9a) and 10 dB (Fig. 4.9b) datasets.

MSBDL outperforms the other tested method in these scenarios. In the multimodal scenario, the proposed method clearly distinguishes itself from the other methods tested. For trimodal data, not only does MSBDL achieve 100% accuracy on the 30 dB data dictionary, but it achieves accuracies of 100% and 99.2% on the 20 dB and 10 dB data dictionaries, respectively. MSBDL outperforms the next best method by 17.2% on the 10 dB data recovery task[15]. Jℓ_0DL was able to capture some of the multimodal information in learning the 10 dB data dictionary, but the 10 dB data dictionary accuracy only reaches 81.9%. Jℓ_1DL performs even worse in recovering the 10 dB data dictionary, achieving 0% accuracy. Similar trends can be seen in the bimodal results.

Next, we evaluate the performance of the MSBDL algorithms in Table 4.1. We repeat the bimodal experiment and compare the proposed methods with Jℓ_1DL, which is the only competing multimodal dictionary learning algorithm that has a stochastic version. The dictionary recovery results are shown in Fig. 4.9. The results show that Jℓ_1DL is not able to recover any part of either the 30 dB nor 10 dB dataset dictionaries. In terms of the asymptotic performance as the batch size approaches L, MSBDL-1 exhibits negligible bias on both datasets, whereas the other MSBDL flavors incur a small bias, especially on the 10 dB dataset. On the other hand, it is interesting that

[15]Throughout this work, we report the improvement to the probability of success or the classification rate in absolute terms.

MSBDL-2 dramatically outperforms MSBDL-1 for batch sizes less than 300, which is unexpected since MSBDL-2 performs approximate sufficient statistic computations. The poor performance of MSBDL-3 and MSBDL-4 suggests that these algorithms should be considered only in extremely memory constrained scenarios. Finally, we report on the performance of the proposed annealing strategy for σ_j. For the bimodal dataset, one expects to σ_1 to converge to a smaller value than σ_2. Fig. 4.11 shows a histogram of the values to which $\boldsymbol{\sigma}$ converge to. The results align with expectations and lend experimental validation for the annealing strategy. We observe the same trend for MSBDL-1 with $L_0 < L$.

To validate the performance of MSBDL using the atom-to-subspace model, we run a number of synthetic data experiments. In all cases, we use $J = 2$ and $L = 1000$. We use MSBDL-1 with $L_0 = 500$ to highlight that the algorithm works in incremental EM mode. We simulate 4 scenarios, summarized in Table 4.2. For each scenario, we first generate the elements of the ground-truth dictionary $\boldsymbol{D}$ by sampling from a $\mathsf{N}(0,1)$ distribution and normalizing the resulting dictionaries to have unit ℓ_2 column norm. We then set $\mathscr{T}^k = \{k, k+M_1\}$ if $k+M_1 \leq M_2$ and k otherwise. This choice of $\left\{\mathscr{T}^k\right\}_{k\in[K]}$ represents the most uniform assignment of columns of D_2 to columns of D_1. We then generate x_1^i by randomly selecting $s = 5$ indices and generating the non-zero entries by drawing samples from a $\mathsf{N}(0,1)$ distribution. We use $\left\{\mathscr{T}^k\right\}_{k\in[K]}$ to find the support of x_2^i and generate the non-zero entries by drawing from a $\mathsf{N}(0,1)$ distribution. In order to assess the performance of the learning algorithm, we must first define the concept of distance between multi-dimensional subspaces. We follow [130] and compute the distance between $D_2[:,\mathscr{T}^k]$ and $\hat{D}_2$ using

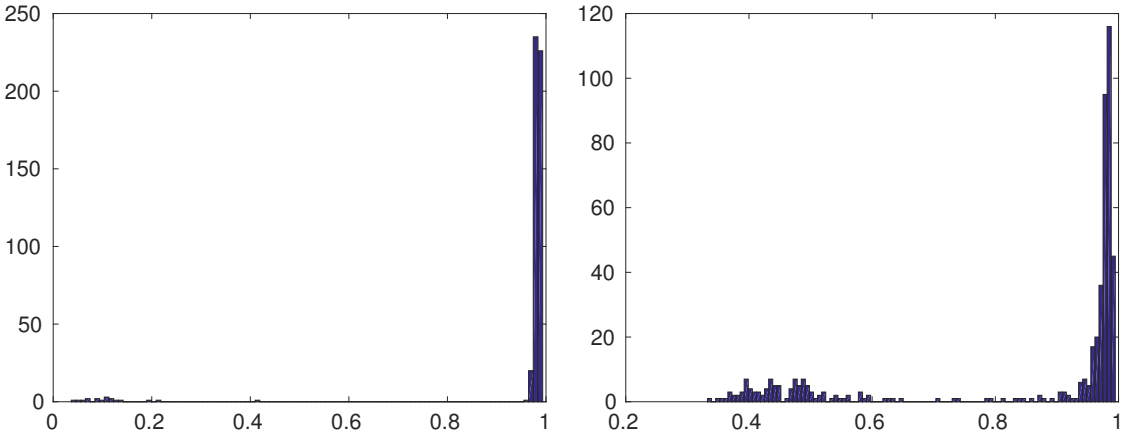
$$\imath\left(D_2[:,\mathscr{T}^k],\hat{D}_2\right) = \max_{1\leq k'\leq K} \sqrt{|V_1^T V_2 V_2^T V_1|},$$

where we use $D_2[:,\mathscr{T}^k]$ to denote the columns of D_2 indexed by $\mathscr{T}^k$, and V_1, V_2 denote orthonormal bases for $D_2[:,\mathscr{T}^k]$ and $\hat{D}_2[:,\mathscr{T}^{k'}]$, respectively. We then define the distance between D_2 and

$\hat{D}_2$ using the two quantities

$$\vartheta_1\left(D_2,\hat{D}_2\right) = c_1 \sum_{k\in\{k:|\mathscr{T}^k|=1\}} \mathbb{1}\left[\iota\left(D_2[:,\mathscr{T}^k],\hat{D}_2\right) > 0.99\right]$$
$$\vartheta_2\left(D_2,\hat{D}_2\right) = c_2 \sum_{k\in\{k:|\mathscr{T}^k|>1\}} \mathbb{1}\left[\iota\left(D_2[:,\mathscr{T}^k],\hat{D}_2\right) > 0.99\right]$$

where $c_1 = |\{k : |\mathscr{T}^k| = 1\}|^{-1}$ and $c_2 = |\{k : |\mathscr{T}^k| > 1\}|^{-1}$. We use MSBDL-1 to learn $\hat{\boldsymbol{D}}$, with accuracy results reported in Table 4.2. Histograms of results are provided in the Appendix for a higher resolution perspective into the performance of the proposed approach. Note that Table 4.2 reports the accuracy of MSBDL-1 in recovering both the atoms and subspaces of D_2. Test case A simulates the scenario where both modalities have a high SNR and $M_2 = 2M_1$. In other words, test case A tests if MSBDL-1 is able to learn in the atom-to-subspace model, without the complications that arise from added noise. The results show that MSBDL-1 effectively learns both the atoms of D_1 and the subspaces comprising D_2. Test case B simulates the scenario where $|\mathscr{T}^k| = 1$ for some k but not for others. In effect, test case B tests the pruning strategy described in Section 4.5.3 and summarized in Fig. 4.5. The results show that the pruning strategy is effective and allows MSBDL-1 to learn both the atoms of D_1 and the atoms and subspaces comprising D_2. Test case C is identical to test case B, but with noise added to modality 2. The results show that MSBDL-1 still effectively recovers D_1, but there is a drop in performance with respect to recovering the atoms of D_2 and a significant drop in recovering the subspaces of D_2. The histogram in Fig. 4.10a shows that the distribution of $\left\{\iota(D_2[:\mathscr{T}^k],\hat{D}_2)\right\}_{k\in K}$ is concentrated near 1 for test case C, suggesting that an alignment threshold of 0.99 is simply too strict in this case. Finally, test case D demonstrates that MSBDL-1 exhibits robust performance when the modality comprising the roots of the tree is noisy. To provide experimental evidence for the fact that the atom-to-subspace model is agnostic to the atoms of D_2, as discussed in Section 4.5.1, we show the ability of MSBDL-1 to recover the atoms of D_2 for test case A in Fig. 4.12. The results show that MSBDL-1 is not able to recover the atoms of D_2.



(a) (b)

Figure 4.10: Histograms of $\iota\left(D_2[:,\mathscr{T}^k],\hat{D}_2\right)$ for test case C in Table 4.2 (4.10a) and $\iota\left(D_2[:,m],\hat{D}_2\right)$ for test case C in Table 4.3 (4.10b).

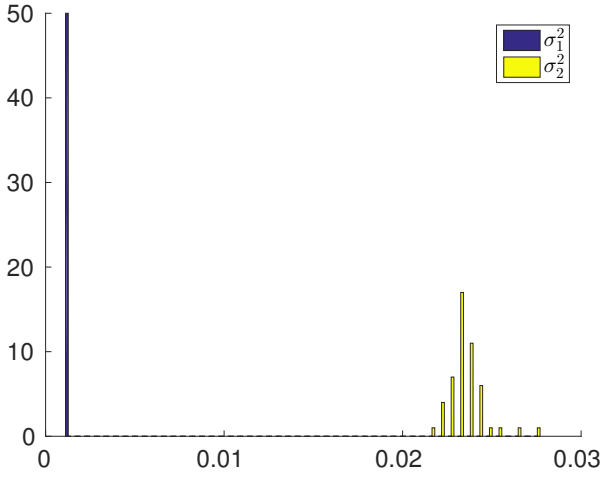


Figure 4.11: Histogram of σ_j^2 at convergence for a bimodal dataset consisting of 30 dB and 10 dB modalities.

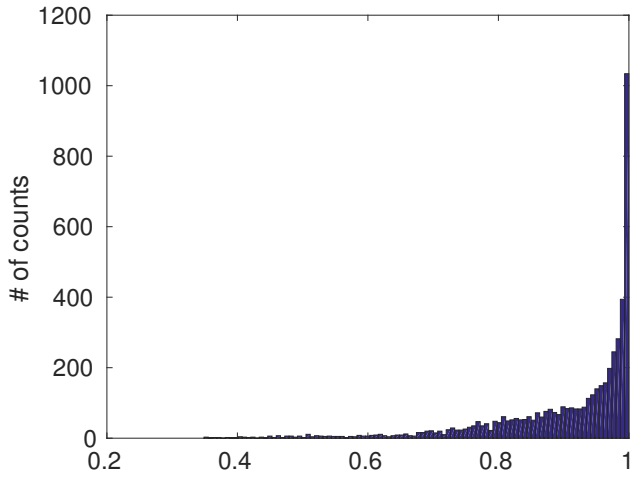


Figure 4.12: Histogram of $\iota\left(D_2[:,m],\hat{D}_2\right)\ \forall m$ for test case A.

Next, we present experimental results for dictionary learning under the hierarchical model, using the same setup as for the atom-to-subspace model. We simulate 4 scenarios, summarized in Table 4.3. To evaluate the performance of the proposed approach, we measure how well it is able to recover the atoms of D_1 and D_2, where we distinguish between the atoms of D_2 corresponding to $|\mathscr{T}^k| = 1$ and $|\mathscr{T}^k| > 1$ using

$$\rho_1\left(D_2,\hat{D}_2\right) = c_3 \sum_{k\in\{k:|\mathscr{T}^k|=1\}} \mathbb{1}\left[\iota\left(D_2[:,\mathscr{T}^k],\hat{D}_2\right) > 0.99\right]$$

$$\rho_2\left(D_2,\hat{D}_2\right) = c_4 \sum_{k\in\{k:|\mathscr{T}^k|=1\},m\in\mathscr{T}^k} \mathbb{1}\left[\iota\left(D_2[:,m],\hat{D}_2\right) > 0.99\right]$$

where $c_3 = |\{k : |\mathscr{T}^k| = 1\}|^{-1}$ and $c_4 = |\{k : |\mathscr{T}^k| > 1\}|^{-1}$. The recovery results are reported in Table 4.3. Histograms of the results are provided in the Appendix. Test case A demonstrates that MSBDL-1 is able to learn the atoms of D_1 and D_2 in the low noise scenario for $M_2 = 2M_1$. Test case B shows that MSBDL-1 is able to learn the atoms of D_1 and D_2 for $M_1 < M_2 < 2M_1$, i.e. highlighting that the pruning strategy in Fig. 4.5 is effective for the hierarchical sparsity model. Test case C adds a considerable amount of noise to the modality occupying the leaves of the tree. Although the recovery results in Table 4.3 suggest that MSBDL-1 does not perform well in

Table 4.2: Recovery results using atom-to-subspace model.

	d_1, M_1	d_2, M_2	SNR_1	SNR_2	$\vartheta_1(D_1, \hat{D}_1)$	$\vartheta_1(D_2, \hat{D}_2)$	$\vartheta_2(D_2, \hat{D}_2)$
A	$20, 50$	$40, 100$	30	30	99.8	—	92
B	$20, 50$	$30, 60$	30	30	99.5	100	97
C	$20, 50$	$30, 60$	30	10	99.9	68.05	3.6
D	$20, 50$	$30, 60$	10	30	73.64	97.7	75.4

Table 4.3: Recovery results using hierarchical model.

	d_1, M_1	d_2, M_2	SNR_1	SNR_2	$\rho_1\left(D_1, \hat{D}_1\right)$	$\rho_1\left(D_2, \hat{D}_2\right)$	$\rho_2\left(D_2, \hat{D}_2\right)$
A	$20, 50$	$40, 100$	30	30	99.3	—	96.6
B	$20, 50$	$30, 60$	30	30	100	94.7	82.7
C	$20, 50$	$30, 60$	30	10	100	22.1	3.3
D	$20, 50$	$30, 60$	10	30	5.2	93.1	61.1

recovering D_2 in this scenario, the histogram in Fig. 4.10b shows robust performance. Finally, test case D shows the scenario where a large amount of noise is added to the modality occupying the roots of each subtree.

4.8.2 Photo Tweet Dataset Classification

We validate the performance of TD-MSBDL on the Photo Tweet dataset [131]. The Photo Tweet dataset consists of 603 tweets covering 21 topics. Each tweet contains text and an image with an associated binary label indicating its sentiment. The dataset consists of 5 partitions. We use these partitions to perform 5 rounds of leave-one-out cross-validation, where, during each round, we use one partition as the test set, one as the validation set, and the remaining partitions as the training set. For each round, we process the images by first extracting a bag of SURF features [132] from the training set using the MATLAB computer vision system toolbox. We then encode the training, validation, and test sets using the learned bag of features, yielding a 500-dimensional representation for each image [133]. Finally, we compute the mean of each dimension across the training set, center the training, validation, and test sets using the computed

Table 4.4: Photo tweet dataset classification accuracy (%).

Feature Type	TD-MSBDL-2	TD-Jℓ_1DL	TD-ℓ_1DL	D-KSVD
Images	**65.6**	59.2	61.1	63.9
Text	**76**	73.7	74.1	69.4

means, and perform 10 component PCA to generate a 10-dimensional image representation. We process the text data by first building a 2688 dimensional bag of words from the training set using the scikit-learn Python library. We then encode the training, validation, and test sets using the learned bag of words, normalizing the resulting representations by the number of words in the tweet. We then center the data and perform 10 component PCA to yield a 10-dimensional text representation.

We run TD-MSBDL using incremental EM and approximate posterior sufficient statistic computations, referring to the resulting algorithm as TD-MSBDL-2 in accordance with the taxonomy in Table 4.1. Our convention is to refer to the text and image data as modalities 1 and 2 respectively. We use $M_j = 40 \forall j$, $L_0 = 200$, $\sigma_1^0 = 0.01$, $\sigma_1^0 = 0.2$, $\beta_j^0 = 100 \forall j$, $\sigma^\infty = 1e-4, \alpha_\sigma = 0.995$, $\beta^\infty = 1e-2, \alpha_\beta = 0.995$, and $T^V = 500$. We compare TD-MSBDL with several unimodal and multimodal approaches. We use TD-ℓ_1DL and D-KSVD to learn classifiers for images and text using unimodal data. For TD-ℓ_1DL we use the validation set to optimize for λ over the set $\{1e-3, 1e-2, 1e-1, 1\}$. We also compare with TD-Jℓ_1DL trained on the multimodal data, using the the same validation approach as for TD-ℓ_1DL. In all cases, we run training for a maximum of $15e3$ iterations.

The classification results are shown in Table 4.4. Comparing TD-MSBDL with the unimodal methods (i.e. TD-ℓ_1DL and D-KSVD), the results show that TD-MSBDL achieves higher performance for both feature types. Moreover, it is interesting that TD-Jℓ_1DL performs worse than TD-ℓ_1DL, suggesting that it is not capable of capturing the multimodal relationships which TD-MSBDL benefits from.

Finally, we show the efficacy of the priors in Section 4.5 in classifying the Photo Tweet

Table 4.5: Photo tweet dataset classification accuracy (%) using TD learning and priors from Section 4.5. Our convention is to designate text as modality 1 and images as modality 2.

	TD-MSBDL-2	TD-MSBDL-1	TD-MSBDL-1
Prior	One-to-one (4.8)	Hierarchical (4.29)	Atom-to-subspace (4.23)
$d_1 \times M_1 / d_2 \times M_2$	$10 \times 40 / 10 \times 40$	$10 \times 40 / 20 \times 80$	$10 \times 40 / 20 \times 80$
Images	65.6	69.2	69.5
Text	76	74.6	74.3

dataset. The goal is to show that, by allowing the number of atoms of the image and text dictionaries to be different, the atom-to-subspace and hierarchical sparsity priors lead to superior classification performance. We begin by extracting features from the text and image data as before, with the exception that we use 20 PCA components to represent images. We then set M_1, the text data dictionary size, to 40 and M_2, the image dictionary size, to 80, corresponding to an oversampling factor M_j/N_j of 4 for both modalities. We run the TD-MSBDL algorithm with the atom-to-subspace and hierarchical sparsity priors. For the atom-to-subspace prior, the only required modification is to change the update rule of $\boldsymbol{\gamma}^j$ to (4.15)[16]. For the hierarchical sparsity prior, the $\boldsymbol{\gamma}^j$, D_j, and W_j update rules are modified to (4.36), (4.37), and $\left(W_j\right)^{t+1} = H\left(U_j^{TD}\right)^T S_j \left(S_j^T \left(U_j^{TD}\left(U_j^{TD}\right)^T + \sum_{i\in[L]} \Sigma_{\hat{x},j}^{TD,i}\right) S_j\right)^{-1}$, respectively. Because there are significant dependencies among the elements in x_j a-priori, we use exact sufficient statistic computation, referring to the resulting algorithm as TD-MSBDL-1. The classification results are presented in Table 4.5, where the TD-MSBDL-2 results with one-to-one prior from Section 4.8.2 are shown for reference. The results show a significant improvement in image classification. Although text classification deteriorates slightly, the text classification rate for both atom-to-subspace and hierarchical priors is still higher than the competing methods in Table 4.4.

[16] We also found it necessary to introduce a post-processing step to the output of the TD-MSBDL algorithm with the atom-to-subspace prior. We output the W_j^t which corresponds to the maximum measured classification accuracy on the validation set during training.

4.9 Conclusion

We have detailed a sparse multimodal dictionary learning algorithm. Our approach incorporates the main features of existing methods, which establish a correspondence between the elements of the dictionaries for each modality, while addressing the major drawbacks of previous algorithms. Our method enjoys the theoretical guarantees and superior performance associated with the sparse Bayesian learning framework.

4.10 Appendix

4.10.1 Incremental EM

One stochastic inference alternative is called Incremental EM, which we review next for the case of $J = 1$, omitting modality subscripts for brevity [118]. Let $F(\tilde{p}, \theta) = E_{\tilde{p}}[\log p(X, Y, \theta)] + H(\tilde{p})$, where $H(\tilde{p})$ is the entropy of $\tilde{p}(\cdot)$. It can be shown that $p(X|Y, \theta)$ is the unique maximizer of $F(\tilde{p}, \theta)$, given θ. It can also be shown that $F(\tilde{p}, \theta) = \log p(Y|\theta)$ for $\tilde{p}(X) = p(X|Y, \theta)$. It then follows that the E and M steps of EM can be re-stated in the following form:

$$\tilde{p}^{t+1} = \arg\max_{\tilde{p}} F(\tilde{p}, \theta^t) \tag{4.51}$$

$$\theta^{t+1} = \arg\max_{\theta} F(\tilde{p}^{t+1}, \theta). \tag{4.52}$$

When the posterior factors over the data points in the dataset, it is reasonable to consider only distributions of the form $\tilde{p}(X) = \prod_{i \in [L]} \tilde{p}^i(x^i)$ in (4.51). Although the factorization constraint may seem restrictive, it should be noted that the maximizer of $F(\tilde{p}, \theta)$ must also factor [17] [118]. It then follows that $F(\tilde{p}, \theta^t) = \sum_{i \in [L]} F^i(\tilde{p}^i, \theta^t)$. This leads to a class of algorithms which perform

[17] In other words, if $\{p^*, \theta^*\}$ is the maximizer of $F(\cdot, \cdot)$, then p^* factors.

the E-step in an incremental fashion by modifying (4.51) to

$$\left(\tilde{p}^i\right)^{t+1} = \begin{cases} \arg\max_{\tilde{p}_i} F^i(\tilde{p}^i, \theta^t) \text{ if } i \in \phi \\ \left(\tilde{p}^i\right)^t \text{ else.} \end{cases} \tag{4.53}$$

4.10.2 Computation of (4.21)

It can be shown that $\partial \log p\left(Y_j|\theta^t, \sigma_j^t\right)/\partial\sigma_j^t$ is given by

$$-\sum_{i\in[L]} \left(y_j^i\right)^T V_j^i \left(\Lambda_j^i\right)^{-2} \left(V_j^i\right)^T y_j^i + \sum_{n\in[N_j]} \left(\Lambda_j[n,n] + \sigma_j^2\right)^{-1}$$

where $V_j^i \Lambda_j^i \left(V_j^i\right)^T$ is the eigen-decomposition of $\Sigma_{y,j}^i$ and $\left(\Lambda_j^i\right)^{-2}$ represents a diagonal matrix whose $[n,n]$'th entry is $\left(\Lambda_j^i[n,n]\right)^{-2}$.

4.10.3 Additional Results for Synthetic Data Experiments

Fig. 4.13 shows histograms of results for the synthetic data experiments in Section 4.8 under the atom-to-subspace prior. Fig. 4.14 shows histograms of results for the synthetic data experiments in Section 4.8 under the hierarchical sparsity prior.

4.10.4 Proof of Theorem 5

By definition, the log-likelihood function is coercive if

$$\lim_{\|\{\theta,\boldsymbol{\sigma}\}\|\to\infty} -\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right) = \infty \tag{4.54}$$

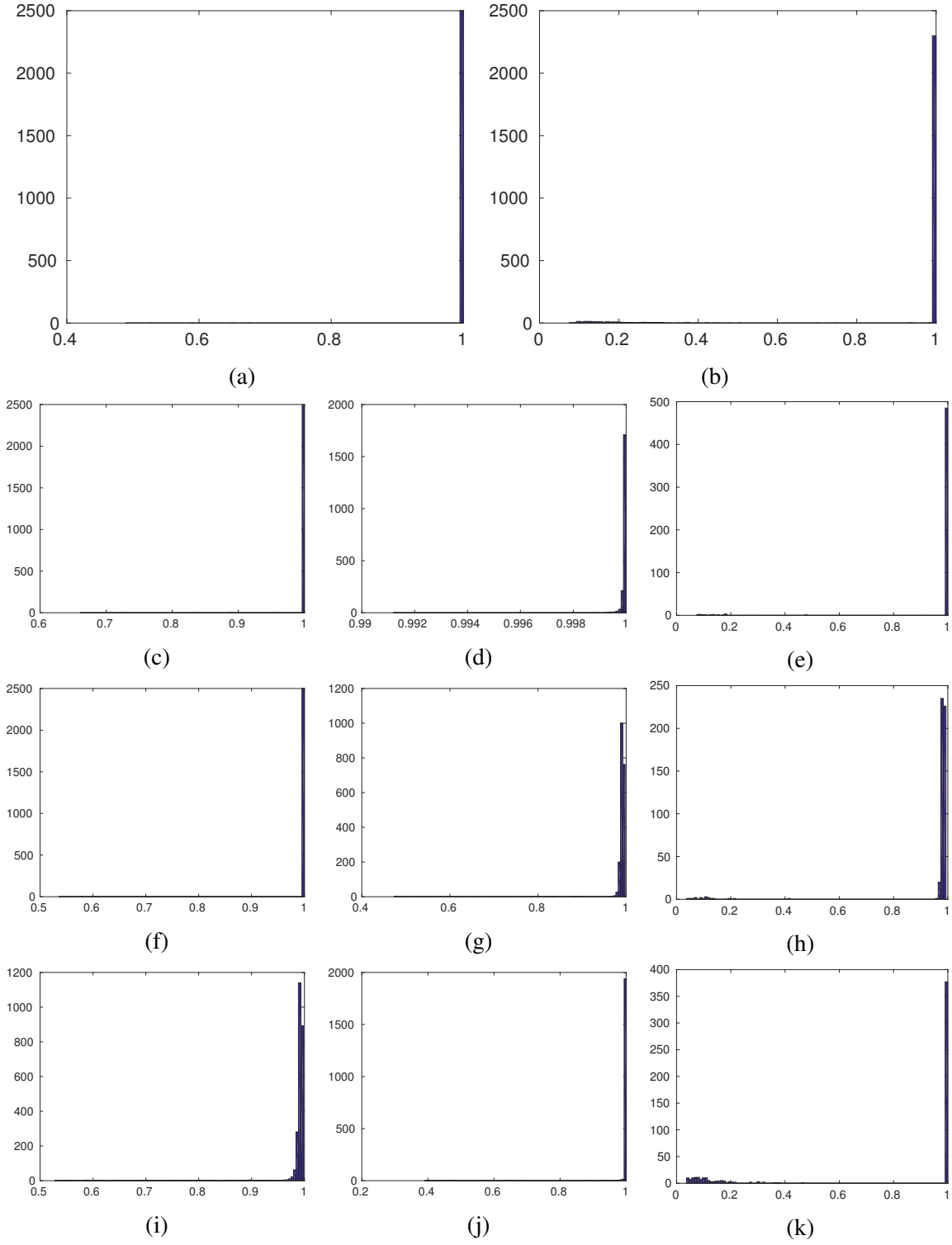


Figure 4.13: Histograms of recovery results for atom-to-subspace model and test cases in Table 4.2. (Fig.'s 4.13a, 4.13c, 4.13f, 4.13i): $\iota\left(D_1[:,m],\hat{D}_1\right)\ \forall m$ for cases A-D. (Fig.'s 4.13d, 4.13g, 4.13j): $\iota\left(D_2[:,k],\hat{D}_2\right), |\mathscr{T}^k| = 1$ for cases B-D. (Fig.'s 4.13b, 4.13e, 4.13h, 4.13k): $\iota\left(D_2[:,k],\hat{D}_2\right), |\mathscr{T}^k| > 1$ for cases A-D.

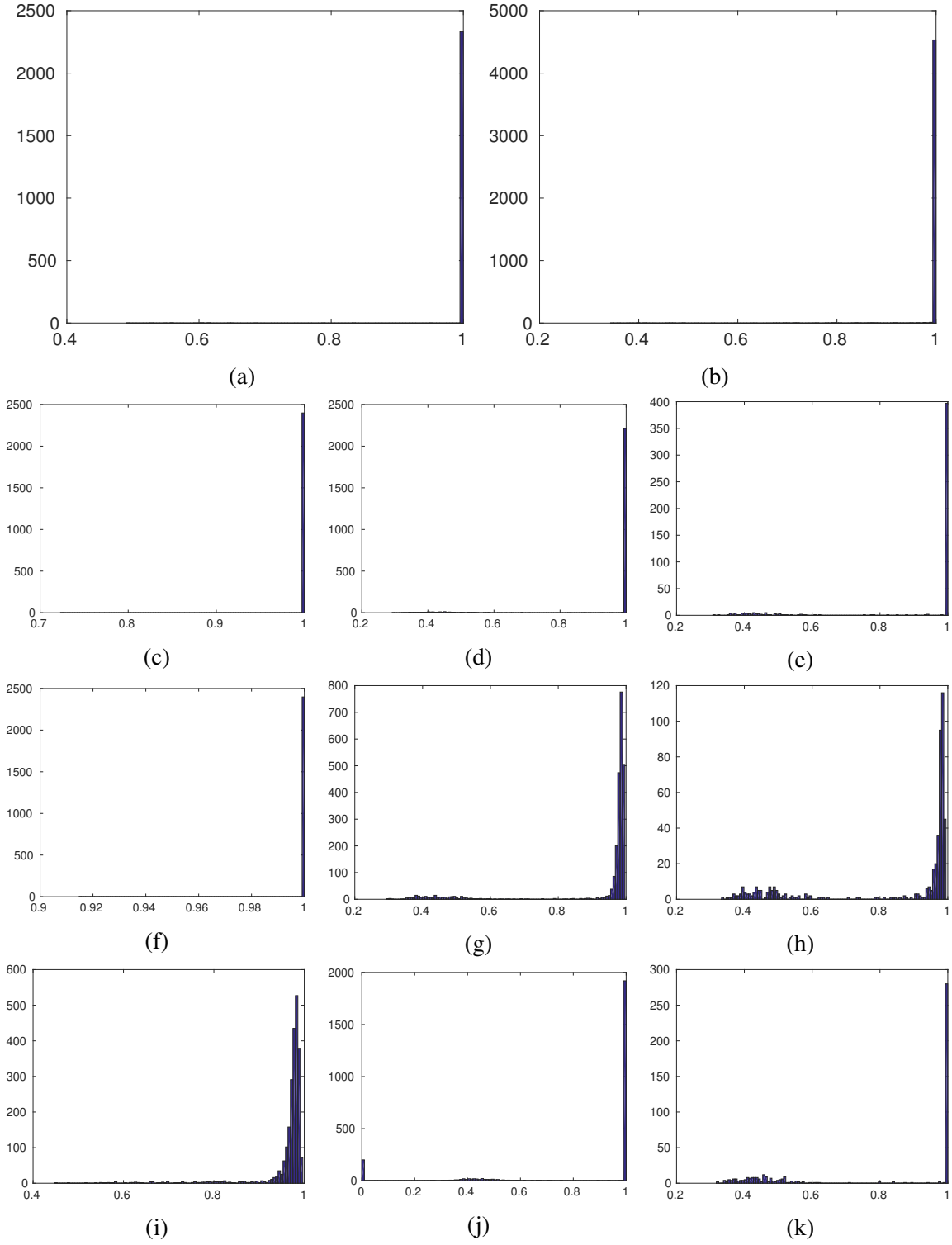


Figure 4.14: Histograms of recovery results for hierarchical model and test cases in Table 4.3. (Fig. 4.14a, 4.14c, 4.14f, 4.14i): $\iota\left(D_1[:,m],\hat{D}_1\right)\ \forall m$ for cases A-D. (Fig. 4.14d, 4.14g, 4.14j): $\iota\left(D_2[:,k],\hat{D}_2\right), |\mathscr{T}^k| = 1$ for cases B-D. (Fig. 4.14b, 4.14e, 4.14h, 4.14k): $\iota\left(D_2[:,m],\hat{D}_2\right)\ \forall m \in T^k : |\mathscr{T}^k| > 1$ for cases A-D.

where we define the norm of $\{\theta, \boldsymbol{\sigma}\}$ to be

$$\|\{\theta, \boldsymbol{\sigma}\}\| = \sqrt{\sum_{m\in[M], j\in[J]} \left\| d_j^m \right\|_2^2 + \sum_{i\in[L]} \|\gamma^i\|_2^2 + \sum_{j\in[J]} \sigma_j^2}. \tag{4.55}$$

The negative log-likelihood can be written as

$$-\log p\left(Y|\theta, \boldsymbol{\sigma}\right) \doteq \sum_{i\in[L], j\in[J]} \left(y_j^i\right)^T \left(\Sigma_{y,j}^i\right)^{-1} y_j^i + \log\left|\Sigma_{y,j}^i\right| \tag{4.56}$$

where $\doteq$ refers to dropping terms which do not depend on θ or $\boldsymbol{\sigma}$.

Next, we establish several results about $\Sigma_{y,j}^i$, which is defined in (4.11). Let $\left(\Gamma^i\right)^{0.5}$ be a diagonal matrix whose $[m,m]$'th entry is given by $\left(\gamma^i[m]\right)^{0.5}$. Then, $D_j\Gamma^i D_j^T$ is the Gramian matrix of $\left(\Gamma^i\right)^{0.5} D_j^T$. Since Gramian matrices are positive semi-definite (PSD), $D_j\Gamma D_j^T$ must be PSD. Since $\sigma_j^2 \mathrm{I}$ is PSD, $\Sigma_{y,j}^i$ is PSD. Finally, since $\Sigma_{y,j}^i$ is PSD, then $\left(\Sigma_{y,j}^i\right)^{-1}$ is also PSD. Therefore,

$$\left(y_j^i\right)^T \left(\Sigma_{y,j}^i\right)^{-1} y_j^i \geq 0 \tag{4.57}$$

in general.

Turning to the second term in (4.56), we can re-write it as

$$\log|\Sigma_{y,j}^i| = \sum_{n\in[N_j]} \log\left(\sigma_j^2 + \lambda_j^i[n]\right) \tag{4.58}$$

where $\lambda_j^i[n] \geq 0$ denotes the n'th eigenvalue of $D_j\Gamma^i D_j^T$. Combining (4.60) with (4.56), we have that

$$-\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right) \geq \sum_{j\in[J], i\in[L], n\in[N_j]} \log\left(\sigma_j^2 + \lambda_j^i[n]\right). \tag{4.59}$$

Note that $\sigma_j^2 + \lambda_j^i[n] > 0$ for all j,i,n, such that the right hand side of (4.59) tends to ∞ if $\sigma_{j^*}^2 + \lambda_{j^*}^{i^*}[n^*]$ tends to ∞ for some j^*, i^*, n^*.

In order for $\|\{\theta, \boldsymbol{\sigma}\}\| \to \infty$, one or more of the terms under the square root in (4.55) must also approach infinity. Starting with $\boldsymbol{\sigma}$, we observe that in order for $\sum_{j\in[J]} \sigma_j^2 \to \infty$, there must be at least one j^* such that $\sigma_{j^*} \to \infty$. Since $\lambda_j^i[n] \geq 0$ for all j,i,n, at least one of the terms in the right hand side of (4.59) must tend to ∞, leading to the result

$$\lim_{\sum_{j\in[J]} \sigma_j^2 \to \infty} -\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right) = \infty.$$

Turning to γ, we observe that in order for $\sum_{i\in[L]} \left\|\boldsymbol{\gamma}^i\right\|_2^2 \to \infty$, there must exist at least one i^* and m^* such that $\boldsymbol{\gamma}^{i^*}[m^*] \to \infty$. We also know that

$$\begin{aligned}
\sum_{j\in[J], n\in[N_j]} \lambda_j^i[n] &= \sum_{j\in[J]} trace\left(D_j \Gamma^i D_j^T\right) && (4.60)\\
&= \sum_{j\in[J], n\in[N_j], m\in[M]} \gamma^i[m] \left(d_j^m[n]\right)^2 && (4.61)\\
&= \sum_{j\in[J], m\in[M]} \gamma^i[m] \left\|d_j^m\right\|_2^2 && (4.62)\\
&\geq \gamma^{i^*}[m^*] \left\|d_{j^*}^{m^*}\right\|_2^2 && (4.63)\\
&=^a \infty. && (4.64)
\end{aligned}$$

where step (a) follows from the assumption that for each m, there exists a j^* such that $\left\|d_{j^*}^m\right\|_2^2 > 0$. Since N_j, J are finite, (4.64) implies that there must exist j', n^* such that $\lambda_{j'}^i[n^*] \to \infty$. In addition, since $\sigma_j^2 > 0$ for all j, $\lambda_{j'}^i[n^*] \to \infty$ implies that at least one of the terms in the right hand side of (4.59) tends to ∞, leading to the result

$$\lim_{\sum_{i\in[L]} \|\gamma^i\|_2^2 \to \infty} -\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right) = \infty$$

Finally, in order for $\sum_{j\in[J],m\in[M]} \left\| d_j^m \right\|_2^2 \to \infty$, there must exist at least one j^* and m^* such that $\left\| d_{j^*}^{m^*} \right\|_2^2 \to \infty$. We can apply the same argument as in (4.60)-(4.64) to conclude that $\left\| d_{j^*}^{m^*} \right\|_2^2 \to \infty$ implies that there exists j', n^* such that $\lambda_{j'}^i[n^*] \to \infty$, where in this case step (a) in (4.64) follows from the assumption that at least one of $\left\{\gamma^j[m]\right\}_{i\in[L]}$ is non-zero for all m. As a result, we have:

$$\lim_{\sum_{j\in[J],m\in[M]} \left\| d_j^m \right\|_2^2 \to \infty} -\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right) = \infty.$$

4.10.5 Proof of Corollary 6

This proof follows closely to the first part of the proof of (Theorem 1, [86]). Let

$$\mathscr{S}_0 = \left\{\theta : -\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right) \leq -\log p\left(\boldsymbol{Y}|\theta^0,\boldsymbol{\sigma}\right)\right\}, \tag{4.65}$$

where θ^0 denotes the initial value of θ. Theorem 5 established that $-\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right)$ is coercive. In addition, assume, for now, that $-\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right)$ is a continuous function of θ. Under these conditions, $\mathscr{S}_0$ is a compact set (Theorem 1.2, [102]). In addition, we have that

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1},\boldsymbol{\sigma}\right) \leq -\log p\left(\boldsymbol{Y}|\theta^{t},\boldsymbol{\sigma}\right)$$

because θ is updated using EM, which guarantees monotonicity of the log-likelihood [26]. Therefore, the sequence $\{-\log p\left(\boldsymbol{Y}|\theta^t,\boldsymbol{\sigma}\right)\}_{t=1}^{\infty}$ is a monotonically decreasing sequence. This monotonicity property guarantees that $\{\theta^t\}_{t=1}^{\infty} \subseteq \mathscr{S}_0$. Since $\mathscr{S}_0$ is compact, $\{\theta^t\}_{t=1}^{\infty}$ admits at least one limit point.

What remains is to show that $-\log p\left(\boldsymbol{Y}|\theta,\boldsymbol{\sigma}\right)$ is continuous. The continuity of the negative log-likelihood follows directly from the fact that both the determinant and matrix inverse

functions are continuous[18].

4.10.6 Proof of Theorem 6

From Corollary 6, we know that $\{\theta^t\}_{t=1}^{\infty}$ admits a limit point. What remains to be shown is that all limit points are stationary and that $\{\log p(\boldsymbol{Y}|\theta,\boldsymbol{\sigma})\}_{t=1}^{\infty}$ converges monotonically to $\log p(\boldsymbol{Y}|\theta^*,\boldsymbol{\sigma})$ for stationary point θ^*. These follow directly from [Theorem 2, [26]] if it can be shown that $Q(\theta,\theta^t)$ is continuous in both θ and θ^t, where $Q(\theta,\theta^t)$ is given by

$$Q(\theta,\theta^t) = \left\langle \log p\left(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{D},\{\gamma^i\}_{i=1}^{L}\right)\right\rangle \tag{4.66}$$

$$= \sum_{j\in[J]} \left\langle \log p\left(Y_j|X_j,D_j\right) + \log p\left(X_j|\{\gamma^i\}_{i=1}^{L}\right)\right\rangle \tag{4.67}$$

$$= \sum_{j\in[J]} Q_j\left(\theta,\theta^t\right). \tag{4.68}$$

We will proceed by showing that $Q_j(\theta,\theta^t)$ is continuous in both θ and θ^t for all j.

First, consider the dependence of $Q_j(\theta,\theta^t)$ on $\gamma^i[m]$, which is given by

$$-\frac{\log\gamma^i[m]}{2} - \left(\frac{\Sigma_{x,j}^i[m,m] + \left(\mu_j{}^i[m]\right)^2}{2\gamma^i[m]}\right) \tag{4.69}$$

where $\Sigma_{x,j}^i$ and $\mu_j{}^i$ are given by (4.13) and (4.14), respectively, and depend on θ^t. It suffices to show that (4.69) is continuous on the open interval $(0,\infty]$. Since both $\log(\cdot)$ and $(\cdot)^{-1}$ are continuous functions on the interval $(0,\infty]$, it follows that (4.69) is continuous in $\gamma^i[m]$. The dependence of $Q_j(\theta,\theta^t)$ on D_j is given by

$$\sum_{i\in[L]} \left(y_j^i\right)^T D_j\mu_j{}^i - \frac{\operatorname{tr}\left(\left(D_j^T D_j\left(\Sigma_{x,j}^i + \mu_j{}^i\left(\mu_j^i\right)^T\right)\right)\right)}{2} \tag{4.70}$$

[18] See [Theorem 5.19 [134]] and [Theorem 5.20 [134]] for continuity of the matrix determinant and inverse functions, respectively.

which is continuous in D_j. Next, we turn to the task of showing that $Q_j(\theta, \theta^t)$ is continuous in θ^t, which reduces to showing that $\Sigma^i_{x,j}$ is continuous in both D^t_j and γ^t. Let B be the matrix being inverted in (4.13):

$$B = \sigma_j^2 \mathrm{I} + D^t_j \Gamma^{t,i} \left(D^t_j\right)^T. \tag{4.71}$$

The task then reduces to showing that $(B)^{-1}$ is continuous in D^t_j and $\gamma^{t,i}$. We first show that $(B)^{-1}$ exists. Using the assumption that D^t_j is full rank, B is full rank over $\gamma \in (0, \infty]^M$ since

$$\begin{aligned} rank(B) &\geq rank\left(D^t_j \Gamma^{t,i} \left(D^t_j\right)^T\right) && (4.72)\\ &= rank\left(D^t_j \left(\Gamma^{t,i}\right)^{\frac{1}{2}} \left(D^t_j \left(\Gamma^{t,i}\right)^{\frac{1}{2}}\right)^T\right) && (4.73)\\ &= rank\left(D^t_j \left(\Gamma^{t,i}\right)^{\frac{1}{2}}\right) && (4.74)\\ &= N_j && (4.75) \end{aligned}$$

where $\left(\Gamma^{t,i}\right)^{\frac{1}{2}}$ is a diagonal matrix with the m'th diagonal entry given by $\sqrt{\gamma^{t,i}[m]}$. Therefore, B is full rank and admits an inverse.

Since B is continuous in D^t_j and γ^t, what remains to be shown is that $(B)^{-1}$ is continuous in B, which has previously been shown in [135]. Therefore, $Q_j(\theta, \theta^t)$ is continuous in θ^t.

4.10.7 Proof of Theorem 7

The guarantee given in Theorem 7 follows directly from [Proposition 6, [136]] if we can show that $Q(\theta, \theta^t)$ has a unique maximizer with respect to θ. Consider the optimization of $Q(\theta, \theta^t)$ with respect to $\boldsymbol{D}$. This optimization problem can be rewritten as

$$\arg\max_{\boldsymbol{D}} \sum_{j \in [J]} \sum_{i \in [L]} -\left(y^i_j\right)^T D_j \mu^i_j + D_j \left(\Sigma^i_{x,j} + \mu^i_j \left(\mu^i_j\right)^T\right). \tag{4.76}$$

Since the terms being summed over j in (4.76) are independent of each other, (4.76) is equivalent to solving

$$\arg\max_{D_j} \sum_{i\in[L]} -\left(y_j^i\right)^T D_j \mu_j^i + D_j\left(\Sigma_{x,j}^i + \mu_j^i\left(\mu_j^i\right)^T\right) \tag{4.77}$$

for all j.

Since (4.77) is an unconstrained optimization problem, its maxima must occur at points where the gradient of the objective function vanishes. Taking the gradient of the objective function in (4.77) with respect to D_j and setting the result to zero, we get

$$\underbrace{D_j\left(U_j U_j^T + \sum_{i\in[L]} \Sigma_{x,j}^i\right)}_{B} = Y_j U_j^t \tag{4.78}$$

If B is invertible, all of the stationary points of the objective function in (4.77) have the form $Y_j U_j (B)^{-1}$. Since Y_j, U_j are fixed and $(B)^{-1}$ is unique given U_j and $\left\{\Sigma_{x,j}^i\right\}_{i=1}^L$, we conclude that the objective function in (4.77) has exactly one, unique stationary point. In order to show that B is invertible, we observe that $\Sigma_{x,j}^i$ is positive semi-definite for all i and U_j is full rank by assumption. Since the sum of a positive semi-definite and positive definite matrix is positive definite, it follows that B is invertible.

We now turn to the optimization of $Q(\theta,\theta^t)$ with respect to $\gamma^j[m]$ (since $Q(\theta,\theta^t)$ is separable in the elements of γ^j). This optimization problem can be rewritten as

$$\begin{aligned}
&\arg\max_{\gamma^j[m]\geq 0} \sum_{j\in[J]} -\frac{\log\gamma^j[m]}{2} - \left(\frac{\Sigma_{x,j}^i[m,m] + \left(\mu_j{}^i[m]\right)^2}{2\gamma^j[m]}\right) \\
&= \arg\min_{\gamma^j[m]\geq 0} \log\gamma^j[m] + \frac{\rho}{\gamma^j[m]}
\end{aligned} \tag{4.79}$$

where

$$\rho = \frac{1}{J} \sum_{j \in [J]} \Sigma^i_{x,j}[m,m] + \left(\mu_j{}^i[m]\right)^2 \tag{4.80}$$

Note that we explicitly state the constraint on $\gamma^j[m]$ in (4.79). For $\rho = 0$, (4.79) reduces to

$$\arg\min_{\gamma^j[m] \geq 0} \log \gamma^j[m] = 0. \tag{4.81}$$

For $\rho > 0$, we can show that $\log \gamma^j[m] + \frac{\rho}{\gamma^j[m]} \geq \log \rho + 1$:

$$\begin{aligned}
\log \rho + 1 - \log \gamma^j[m] - \frac{\rho}{\gamma^j[m]} &= \log \frac{\rho}{\gamma^j[m]} + 1 - \frac{\rho}{\gamma^j[m]} \\
&\leq^a \frac{\rho}{\gamma^j[m]} - 1 + 1 - \frac{\rho}{\gamma^j[m]} \\
&= 0 \\
&\downarrow \\
\log \rho + 1 &\leq \log \gamma^j[m] + \frac{\rho}{\gamma^j[m]}
\end{aligned}$$

where step (a) follows from the identity $\log x \leq x - 1$ for $x > 0$ [137]. Equality in step (a) is achieved only for $\gamma^j[m] = \rho$. To see this, we observe that $\log x$ is a strictly concave function that is tangent to the function $x - 1$ at $x = 1$. The strict concavity of $\log x$ implies that it can only be tangent to the function $x - 1$ at a single point. Therefore, the objective in (4.79) is lowerbounded by $\log \rho + 1$, with the lowerbound achieved at $\gamma^j[m] = \rho$. Putting this result together with the case when $\rho = 0$, we see that (4.79) admits a single, unique optimizer given by ρ in (4.80).

4.10.8 Proof of Corollary 7

This proof follows closely to the first part of the proof of (Theorem 1, [86]). Let

$$\mathscr{S}_0 = \left\{ \{\theta, \boldsymbol{\sigma}\} : -\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right) \leq -\log p\left(\boldsymbol{Y}|\theta^0, \boldsymbol{\sigma^0}\right) \right\}, \tag{4.82}$$

where θ^0 and $\boldsymbol{\sigma^0}$ denote the initial values of θ and $\boldsymbol{\sigma}$, respectively. Theorem 5 established that $-\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right)$ is coercive. In addition, assume, for now, that $-\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right)$ is a continuous function of $\{\theta, \boldsymbol{\sigma}\}$. Under these conditions, $\mathscr{S}_0$ is a compact set (Theorem 1.2, [102]). In addition, we have that

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^t\right) \leq -\log p\left(\boldsymbol{Y}|\theta^t, \boldsymbol{\sigma}^t\right)$$

because θ is updated using EM, which guarantees monotonicity of the log-likelihood [26]. Likewise, we have that

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^{t+1}\right) \leq -\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^t\right)$$

by construction of the update rule in (4.20). Therefore, the sequence $\left\{-\log p\left(\boldsymbol{Y}|\theta^t, \boldsymbol{\sigma^t}\right)\right\}_{t=1}^{\infty}$ is a monotonically decreasing sequence, i.e.

$$-\log p\left(\boldsymbol{Y}|\theta^{t+1}, \boldsymbol{\sigma}^{t+1}\right) \leq -\log p\left(\boldsymbol{Y}|\theta^t, \boldsymbol{\sigma}^t\right). \tag{4.83}$$

This monotonicity property guarantees that $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty} \subseteq \mathscr{S}_0$. Since $\mathscr{S}_0$ is compact, $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty}$ admits at least one limit point.

What remains is to show that $-\log p\left(\boldsymbol{Y}|\theta, \boldsymbol{\sigma}\right)$ is continuous. The continuity of the negative log-likelihood follows directly from the fact that both the determinant and matrix inverse

functions are continuous[19].

4.10.9 Proof of Theorem 8

We will show that any point in $\Omega_{\sigma_j^1,j}$ must be in $\Omega_{\sigma_j^2}$. Let the operator $\Theta_l : \mathbb{R}^{M\times M} \rightarrow \mathbb{R}^{M-N_j\times M-N_j}$ be defined such that $\Theta_l(\Gamma)$ extracts the top left $M-N_j \times M-N_j$ submatrix of Γ. Let the operator $\Theta_h : \mathbb{R}^{M\times M} \rightarrow \mathbb{R}^{N_j\times N_j}$ be defined such that $\Theta_h(\Gamma)$ extracts the bottom right $N \times N$ submatrix of Γ. Using these operators, we can express any element of $\Omega_{\sigma_j,j}$ as

$$\Sigma_{y_j} = \left(\sigma^2 \mathrm{I} + \Theta_h(\Gamma)\right) + \check{D}_j \Theta_l(\Gamma) \check{D}_j^T . \tag{4.84}$$

Let $\Sigma_{y,j}^1(D_j^1, \gamma^1) \in \Omega_{\sigma_j^1,j}$, where $\Sigma_{y,j}(\cdot,\cdot)$ denotes the dependence of $\Sigma_{y,j}$ on D_j and γ. We can then show that $\Sigma_{y,j}^1(D_j^1, \gamma^1) = \Sigma_{y,j}^2(D_j^2, \gamma^2) \in \Omega_2$ for

$$\gamma^2[m] = \begin{cases} \gamma^1[m] & \text{if } m \leq M - N_j \\ \gamma^1[m] + \left(\sigma_j^1\right)^2 - \left(\sigma_j^2\right)^2 & \text{else} \end{cases}$$

and $\check{D}_j^2 = \check{D}_j^1$. Such a choice of $\Sigma_{y,j}^2(D_j^2, \gamma^2)$ is always possible because $\sigma_j^1 > \sigma_j^2$. The converse is not true for arbitrary choices of D_j^1 and γ^1, leading to the relation $\Omega_{\sigma_j^1} \subseteq \Omega_{\sigma_j^2}$.

4.10.10 Proof of Theorem 9

We begin by studying the shape of

$$\log p\left(Y_j | \theta^t, \sigma_j\right). \tag{4.85}$$

[19] See [Theorem 5.19 [134]] and [Theorem 5.20 [134]] for continuity of the matrix determinant and inverse functions, respectively.

The log-likelihood in (4.85) depends on σ_j through the covariance matrices $\left\{\Sigma^i_{y,j}\right\}_{i=1}^L$ shown in (4.11). If we parametrize $p\left(y^i_j;\Sigma^i_{y,j}\right)$ by the precision matrix $\Lambda^i_j = \left(\Sigma^i_{y,j}\right)^{-1}$, then it can be shown that $\log p\left(y^i_j;\Lambda^i_j\right)$ is a strictly concave function of Λ^i_j. Since the log-likelihood of Y_j is a sum of such functions, $\log p\left(Y_j|\left\{\Lambda^i_j\right\}_{i=1}^L\right)$ is itself a strictly concave function. Therefore, $\log p\left(Y_j|\{\Lambda^i_j\}_{i\in[L]}\right)$ admits a single local maximum $\left\{\Lambda^{i,*}_j\right\}_{i=1}^L$, which is also its global maximum. Since the mapping from Λ^i_j to $\Sigma^i_{y,j}$ is one-to-one, we conclude that $\log p\left(Y_j|\left\{\Sigma^i_{y,j}\right\}_{i=1}^L\right)$ also admits a single local maximum, which is also a global maximum. In other words, the function $\log p\left(Y_j|\left\{\Sigma^i_{y,j}\right\}_{i=1}^L\right)$ is a strictly quasiconcave function [97]. Consider maximizing the log-likelihood over the convex set $\Sigma^i_{y,j} \in \{\sigma^2 \mathrm{I} + D_j \Gamma^i D_j^T : \sigma_j > 0\}$. Quasiconcave functions admit a single local maximum, which is also the global maximum, over convex sets [97]. We conclude that $\log p\left(Y_j|\left\{\Sigma^i_{y,j}\right\}_{i=1}^L\right)$ admits a single maximum with respect to σ_j.

Suppose that the condition

$$\sigma_j^t = \sigma_j^{t-1} \tag{4.86}$$

is satisfied at iteration t (and not before), meaning that

$$\log p\left(Y_j|\theta^t, \alpha\sigma_j^{t-1}\right) < \log p\left(Y_j|\theta^t, \sigma_j^{t-1}\right). \tag{4.87}$$

Because α_σ can be arbitrarily close to 1 (4.87) implies that there exists a neighborhood

$$\left[\sigma_j^{t-1} - \varepsilon_1, \sigma_j^{t-1}\right],\ \varepsilon_1 > 0,$$

over which (4.85) is increasing. We now claim that there must also exist a neighborhood $\left[\sigma_j^{t-1}, \sigma_j^{t-1} + \varepsilon_2\right], \varepsilon_2 > 0$, for which (4.85) is decreasing. Suppose that the converse is true and

that there exists a σ_j^* such that

$$\log p\left(Y_j|\theta^t,\sigma_j^{t-1}\right) < \log p\left(Y_j|\theta^t,\sigma_j^*\right),\sigma_j^* > \sigma_j^{t-1}. \tag{4.88}$$

Remember that $\theta^t = \left\{\left\{\gamma^{t,i}\right\}_{i=1}^L, \boldsymbol{D^t}\right\}$ where $D_j^t \in \Psi_j$. The inequality in (4.88) means that there exists $\theta^* = \left\{\left\{\gamma^{*,i}\right\}_{i=1}^L, \boldsymbol{D^t}\right\}$ with

$$\gamma^{*,i}[m] = \begin{cases} \gamma^{t,i}[m] & \text{if } m \leq M - N_j \\ \gamma^{t,i}[m] + \left(\sigma_j^*\right)^2 - \left(\sigma_j^{t-1}\right)^2 & \text{else} \end{cases}. \tag{4.89}$$

such that

$$\log p\left(Y_j|\theta^t,\sigma_j^{t-1}\right) < \log p\left(Y_j|\theta^*,\sigma_j^{t-1}\right). \tag{4.90}$$

But (4.90) must be a contradiction since we have assumed that

$$\theta^t = \arg\max_{\theta} \log p\left(Y_j|\theta,\sigma_j^{t-1}\right). \tag{4.91}$$

The condition that $\sigma_j^0 \geq \arg\max_{\sigma_j} \max_{\theta} \log p\left(Y_j|\theta,\sigma_j\right)$ ensures that the preceding argument holds at iteration $t = 1$.

To conclude, we have shown that if (4.86) is satisfied, σ_j^{t-1} is a local maximum. Since we have already shown that the objective in (4.85) only admits a single maximum, this completes the proof.

4.10.11 Proof of Theorem 10

Since Corollary 7 established that $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty}$ admits at least one limit point, what remains is to show that all limit points are stationary. Let $\{\theta^*, \boldsymbol{\sigma^*}\}$ denote any limit point of $\{\theta^t, \boldsymbol{\sigma^t}\}_{t=1}^{\infty}$.

For any $\boldsymbol{\sigma}$, we know that a limit point $\bar{\theta}$ is stationary from Theorem 6. Therefore, $\bar{\theta} = \theta^*$ must be a stationary point.

For any θ, we know that a limit point $\bar{\boldsymbol{\sigma}}$ must be the global maximizer of the log-likelihood from Theorem 9. Since a global maximizer must be stationary, we conclude that $\bar{\boldsymbol{\sigma}}$ must be stationary. Therefore, $\bar{\boldsymbol{\sigma}} = \boldsymbol{\sigma}^*$ must be a stationary point.

4.10.12 Proof of Thm. 11

This proof is an extension of the proof shown in [127]. Under the assumptions of Theorem 11, we can focus exclusively on recovering $\boldsymbol{D}$ because, given $\{\boldsymbol{Y}, \boldsymbol{D}\}$, the sparse codes $\boldsymbol{X}$ are unique [128]. To prove that $\boldsymbol{D}$ is unique, we will show how $\boldsymbol{D}$ can be recovered by construction.

The construction of $\hat{\boldsymbol{D}}$ proceeds in three steps:

1. Divide the columns of Y_j into $R = \binom{M}{s}$ sets $\{G_j^1, \cdots, G_j^R\}$ for all j, where

$$G_j^k = \left\{ i : y_j^i \in span\left(D_j[:, \Upsilon^k] \right) \right\}$$

   and Υ^k denotes the k'th subset of size s of $[M]$.

2. Detect pairs $(G_j^{k_1}, G_j^{k_2})$ such that $|G_j^{k_1} \cap G_j^{k_2}| = 1$ for all j.

3. For each j, find the atom common to Υ^{k_1} and Υ^{k_2}. This atom is necessarily one of the atoms of D_j [127]. Repeat for all pairs (k_1, k_2).

We begin by describing how the data is clustered. Starting with modality j^*, we begin by testing every group of $s+1$ data points from Y_{j^*}. The rank of this group of points will be s if and only if the points lie in the subspace spanned by a set of s columns from D_{j^*} [127]. Once $\left\{ G_{j^*}^k \right\}_{k=1}^R$ has been established, the remaining points in Y_{j^*} which have not been assigned to a set are combined with one of the groups $G_{j^*}^k$ based on the fact that the rank of the subspace spanned by the columns indexed by $G_{j^*}^k$ and an additional column, y_j^i, from Y_{j^*} is s if an only

if $y_j^i \in span\left(D_j[:,\Upsilon_k]\right)$. Finally, due to the nature of the data generation process, we know that $G_j^k = \{i : P[j,i] = 1 \text{ and } i \in G_{j^*}^k\}, \forall j$. Note that the construction of $G_{j^*}^k$ requires $s+1$ data points from modality j^*, but we get G_j^k directly from $G_{j^*}^k$.

Next, we describe the process by which we detect pairs $(G_j^{k_1}, G_j^{k_2})$ such that $|\Upsilon^{k_1} \cap \Upsilon^{k_2}| = 1$. This can be done for each modality j independently. Namely, for each pair $(G_j^{k_1}, G_j^{k_2})$, we test the rank of the subspace spanned by the columns of Y_j indexed by $G_j^{k_1} \cup G_j^{k_2}$. The rank of this subspace will be $2s-1$ if and only if the intersection of $span(G_j^{k_1})$ and $span(G_j^{k_2})$ has dimension 1. This process is guaranteed to produce every atom of D_j at least once [127].

Finally, we describe how to form $\hat{D}_j$. Given a pair $(G_j^{k_1}, G_j^{k_2})$ such that $|\Upsilon^{k_1} \cap \Upsilon^{k_2}| = 1$, we extract any s points from $G_j^{k_1}$ and any s points from $G_j^{k_2}$ and concatenate them into two matrices B^{k_1} and B^{k_2}, respectively. There exist vectors v^1 and v^2 such that both $B^{k_1}v^1$ and $B^{k_2}v^2$ are parallel to the atom of D_j of interest. We can now set up a system of equations given by

$$\underbrace{\begin{bmatrix} B^{k_1} & -B^{k_2} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} v^1 \\ v^2 \end{bmatrix}}_{v} = 0 \tag{4.92}$$

and find v, which is guaranteed to exist since $rank(B) = 2s-1$ from the detection step. We can then extract v^1 from v and find one of the columns of D_j (up to scaling) using $B^{k_1}v^1$. This process can be repeated to find all M atoms of D_j (up to scale) [127].

The major difference between the proof given here and the one in [127] for the unimodal dictionary learning problem is that we require $s+1$ data points per every s dimensional subspace for *one* of the modalities and only s data points per subspace for the rest of the modalities. The reason that we can use *less* samples stems from the special structure of the multimodal data generation process. In the end, we require $s+1$ data points from modality j^* to complete the data clustering step and s data points from each of the $J-1$ data modalities to complete the extraction step.

Chapter 4, in full, is a reprint of material in Igor Fedorov and Bhaskar D. Rao, "Multimodal Sparse Bayesian Dictionary Learning," submitted to the IEEE Transactions on Signal Processing. I was the primary author and B. D. Rao supervised the research.

Chapter 5

Conclusion and Future Work

This thesis has presented three algorithms which employ scale mixtures to achieve structured sparse learning. In the following, we review the main contributions delineated in this thesis and suggest avenues for future work.

1. Chapter 2 presented a novel approach for modeling signals exhibiting a stationary sparsity profile but corrupted by both stationary and non-stationary noise. We showed how scale mixture priors can be used to model both the signal and noise and reported signal recovery results that lend evidence to the superiority of our proposed approach in comparison with existing methods. We applied our algorithm to the task of face recognition in the context of a multiple observation system where each observation is heavily occluded. The results show that our algorithm achieves a significant classification accuracy gain over existing algorithms from the literature.

    - Future work in this line of research could focus on various applications of the model and inference scheme we proposed. We have conducted preliminary work in applying our algorithm to person re-identification, especially in the context of airport camera systems [138].

2. Chapter 3 focused on the problem of non-negative least squares and non-negative matrix factorization. We developed a broad, novel class of priors to model sparse, non-negative signals. We then defined a unified MAP inference framework and equipped it with efficient, provably convergent multiplicative update rules. The significance of this work is two-fold. First, we showed that our framework encompasses a large class of possible algorithms, some of which exist in the literature already. Our approach makes it easy to formulate novel approaches by simpling choosing a prior from the rectified power exponential scale mixture family. Second, we showed that our framework provably converges to the set of stationary points.

    - We believe there are at least two significant avenues of future research that could stem from this work. First, it is still unclear which priors from the proposed family of signal priors are better in various application areas. Second, we did not address the fundamental question of how well the proposed approaches are able to recover the dictionary which generated the data and what modifications need to be made to yield improved dictionary recovery performance.

3. Chapter 4 presented our work on multimodal dictionary learning. We presented an algorithm called MSBDL, which improves upon existing methods in a number of ways. First, MSBDL is unique in that it is capable of modeling dictionaries of varying cardinality across modalities. Second, we equipped MSBDL with an automatic hyperparameter tuning strategy, making it easy for practitioners to deploy MSBDL on novel datasets without the hassle of a large hyperparameter search. Third, we devised scalable inference strategies which make MSBDL applicable in the context of large datasets. Fourth, we extended MSBDL to cover supervised datasets and showed that it is capable of learning classifiers for multiple datasets simultaneously. We then studied the properties of MSBDL through the lenses of theory and experimentation. Our analysis revealed that MSBDL is a convergent

algorithm and can be proven to converge to a stationary point under certain assumptions. We showed that the success of our hyperparameter tuning strategy can partially be explained by the fact that it successively increases the size of the search space, in some sense eliminating the possibility of getting stuck in poor local optima. We also presented an argument in favor of multimodal dictionary in general, showing that the minimum number of samples needed to guarantee dictionary recovery is strictly smaller when learning dictionaries jointly as opposed to independently. Our experimental results showed that MSBDL dramatically outperforms competing algorithms in dictionary recovery, it retains its favorable dictionary recovery qualities even when scalable inference is performed, and it is able to capture complex relationships in both synthetic and real-world data.

- We believe that there are at least two significant future research directions that can build on the MSBDL algorithm described in Chapter 4. The first issue is that the linear forward model (i.e. that y_j is a linear function of x_j) is rather restrictive. Ultimately, we hope that our work on linear models can be used to guide researchers in extending MSBDL to more complex models, such as neural networks. The cost of a richer forward model is that closed-form inference is no longer possible and one must resort to approximate inference, such as variational inference (VI) [139]. Of course, VI itself can be restrictive in that it constrains the family of approximating distributions, but recent work has shown that this class of approximating distributions can be significantly extended [140, 141]. The second issue with MSBDL is that, although we have made progress toward addressing the scalability of MSBDL, there is much room for even higher reduction of computational and memory complexity. VI can be exploited for this purpose as well, where recent work on amortized [142], semi-amortized [143], and stochastic VI has shown that VI can be a highly efficient and robust density estimator.

Bibliography

[1] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic resonance in medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[2] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 83–91, 2008.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[4] M. Elad, *Sparse and Redundant Representations*. Springer New York, 2010.

[5] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[7] R. Giri and B. Rao, "Type I and Type II Bayesian Methods for Sparse Signal Recovery Using Scale Mixtures," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3418–3428, July 2016.

[8] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[9] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*. IEEE, 2008, pp. 3869–3872.

[10] M. Aharon, M. Elad, and A. Bruckstein, "k -svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[11] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.

[12] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[13] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11, pp. 2651–2664, 2013.

[14] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2691–2698.

[15] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[16] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[18] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural computation*, vol. 15, no. 2, pp. 349–396, 2003.

[19] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Strong sub-and super-gaussianity," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 303–310.

[20] J. Palmer, K. Kreutz-Delgado, B. D. Rao, and D. P. Wipf, "Variational em algorithms for non-gaussian latent variable models," in *Advances in neural information processing systems*, 2005, pp. 1059–1066.

[21] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[23] ——, "Iteratively reweighted least squares for linear regression when errors are normal/independent distributed," 1980.

[24] J. A. Palmer, "Variational and scale mixture representations of non-gaussian densities for estimation in the bayesian linear model: Sparse coding, independent component analysis, and minimum entropy segmentation," 2006.

[25] I. Fedorov, A. Nalci, R. Giri, B. D. Rao, T. Q. Nguyen, and H. Garudadri, "A unified framework for sparse non-negative least squares using multiplicative updates and the non-negative matrix factorization problem," *Signal Processing*, vol. 146, pp. 79 – 91, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168418300033

[26] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[27] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153–2164, 2004.

[28] J. Palmer, B. D. Rao, and D. P. Wipf, "Perspectives on sparse bayesian learning," in *Advances in neural information processing systems*, 2004, pp. 249–256.

[29] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[30] I. Fedorov, B. D. Rao, and T. Q. Nguyen, "Multimodal sparse bayesian dictionary learning applied to multimodal data classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2237–2241.

[31] I. Fedorov and B. D. Rao, "Multimodal sparse bayesian dictionary learning," *arXiv preprint arXiv:1804.03740*, 2018.

[32] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, Jan 2016.

[33] I. Fedorov, R. Giri, B. D. Rao, and T. Q. Nguyen, "Robust bayesian method for simultaneous block sparse signal recovery with applications to face recognition," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 3872–3876.

[34] D. J. MacKay, "Hyperparameters: Optimize, or integrate out?" in *Maximum entropy and Bayesian methods*. Springer, 1996, pp. 43–59.

[35] ——, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[36] S. F. Gull, "Bayesian inductive inference and maximum entropy," in *Maximum-entropy and Bayesian methods in science and engineering*. Springer, 1988, pp. 53–74.

[37] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan 2006.

[38] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, June 2010.

[39] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 2790–2797.

[40] A. Y. Yang, S. Iyengar, P. Kuryloski, and R. Jafari, "Distributed segmentation and classification of human actions using a wearable motion sensor network," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, June 2008, pp. 1–8.

[41] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 1873–1879.

[42] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 625–632.

[43] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, Feb 2012.

[44] T. Li and Z. Zhang, "Robust face recognition via block sparse bayesian learning," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[45] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, July 2005.

[46] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. Springer, 2011, pp. 91–102.

[47] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 8, pp. 1629–1642, Aug 2015.

[48] S. Karanam, Y. Li, and R. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 33–40.

[49] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.

[50] D. P. Wipf and B. D. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3704–3716, 2007.

[51] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, Sept 2011.

[52] ——, "Extension of sbl algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2009–2015, April 2013.

[53] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 3830–3833.

[54] S. Bakin, "Adaptive regression and model selection in data mining problems," Ph.D. dissertation, 1999.

[55] J. Liu, S. Ji, J. Ye *et al.*, "Slep: Sparse learning with efficient projections," *Arizona State University*, vol. 6, p. 491, 2009.

[56] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, Jun 2001.

[57] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, "Text mining using non-negative matrix factorizations." in *SDM*, vol. 4, 2004, pp. 452–456.

[58] V. Monga and M. K. Mihçak, "Robust and secure image hashing via non-negative matrix factorizations," *Information Forensics and Security, IEEE Transactions on*, vol. 2, no. 3, pp. 376–390, 2007.

[59] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 857–869, 2005.

[60] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[61] P. Sajda, S. Du, T. R. Brown, R. Stoyanova, D. C. Shungu, X. Mao, and L. C. Parra, "Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 12, pp. 1453–1465, 2004.

[62] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1974, vol. 161.

[63] R. Bro and S. De Jong, "A fast non-negativity-constrained least squares algorithm," *Journal of chemometrics*, vol. 11, no. 5, pp. 393–401, 1997.

[64] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[65] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with 0-constraints," *Neurocomputing*, vol. 80, pp. 38–46, 2012.

[66] P. O. Hoyer, "Non-negative sparse coding," in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*. IEEE, 2002, pp. 557–565.

[67] ——, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[68] I. Tošić and P. Frossard, "Dictionary learning," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 27–38, 2011.

[69] M. J. Gangeh, A. K. Farahat, A. Ghodsi, and M. S. Kamel, "Supervised dictionary learning and sparse representation-a review," *arXiv preprint arXiv:1502.05928*, 2015.

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

[71] M. Aharon, M. Elad, and A. M. Bruckstein, "K-svd and its non-negative variant for dictionary design," in *Optics & Photonics 2005*. International Society for Optics and Photonics, 2005, pp. 591 411–591 411.

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

[73] C.-b. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[74] D. P. Bertsekas, "Nonlinear programming," 1999.

[75] G. Zhou, A. Cichocki, and S. Xie, "Fast nonnegative matrix/tensor factorization based on low-rank approximation," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2928–2940, 2012.

[76] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie, "Nonnegative matrix and tensor factorizations: An algorithmic perspective," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 54–65, 2014.

[77] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *Signal Processing Letters, IEEE*, vol. 22, no. 3, pp. 293–297, 2015.

[78] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 322–329.

[79] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition." in *Interspeech*. Citeseer, 2010, pp. 717–720.

[80] E. F. Gonzalez and Y. Zhang, "Accelerating the lee-seung algorithm for non-negative matrix factorization," *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, 2005.

[81] N. Gillis, "Sparse and unique nonnegative matrix factorization through data preprocessing," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3349–3386, 2012.

[82] D. Wipf and S. Nagarajan, "Iterative reweighted and methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[83] A. Lefevre, F. Bach, and C. Févotte, "Itakura-saito nonnegative matrix factorization with group sparsity," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 21–24.

[84] P. D. Grady and S. T. Rickard, "Compressive sampling of non-negative signals," in *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*. IEEE, 2008, pp. 133–138.

[85] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.

[86] R. Zhao and V. Y. Tan, "A unified convergence analysis of the multiplicative update algorithm for nonnegative matrix factorization," *arXiv preprint arXiv:1609.00951*, 2016.

[87] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102, 1974.

[88] A. Nalci, I. Fedorov, M. Al-Shoukairi, T. T. Liu, and B. D. Rao, "Rectified gaussian scale mixtures and the sparse non-negative least squares problem," *IEEE Transactions on Signal Processing (to appear)*, 2018.

[89] R. Giri and B. Rao, "Learning distributional parameters for adaptive bayesian sparse signal recovery," *IEEE Computational Intelligence Magazine Special Issue on Model Complexity, Regularization and Sparsity*, November 2016.

[90] K. Lange and J. S. Sinsheimer, "Normal/independent distributions and their applications in robust regression," *Journal of Computational and Graphical Statistics*, vol. 2, no. 2, pp. 175–198, 1993.

[91] R. Schachtner, G. Poeppel, A. Tomé, and E. Lang, "A bayesian approach to the lee–seung update rules for nmf," *Pattern Recognition Letters*, vol. 45, pp. 251–256, 2014.

[92] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A novel hierarchical bayesian approach for sparse semisupervised hyperspectral unmixing," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 585–599, 2012.

[93] F. Chen and Y. Zhang, "Sparse hyperspectral unmixing based on constrained lp-l 2 optimization," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1142–1146, 2013.

[94] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 141–145.

[95] W. Dong, X. Li, L. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 457–464.

[96] S. Jiang and Y. Gu, "Block-sparsity-induced adaptive filter for multi-clustering system identification," *arXiv preprint arXiv:1410.5024*, 2014.

[97] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[98] J. M. Bioucas-Dias and M. A. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*. IEEE, 2010, pp. 1–4.

[99] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on pure and applied mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[100] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 760–770, March 2003.

[101] K. Kreutz-Delgado and B. D. Rao, "A general approach to sparse basis selection: Majorization, concavity, and affine scaling," *University of California, San Diego, Tech. Rep. UCSD-CIE-97-7-1*, 1997.

[102] J. V. Burke, *Undergraduate Nonlinear Continuous Optimization*.

[103] M. Cha, Y. Gwon, and H. Kung, "Multimodal sparse representation learning and applications," *arXiv preprint arXiv:1511.06238*, 2015.

[104] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, Jan 2014.

[105] J. C. Caicedo and F. A. González, "Multimodal fusion for image retrieval using matrix factorization," in *Proceedings of the 2nd ACM international conference on multimedia retrieval*. ACM, 2012, p. 56.

[106] Y. Ding and B. D. Rao, "Joint dictionary learning and recovery algorithms in a jointly sparse framework," in *2015 49th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2015, pp. 1482–1486.

[107] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[108] Y. Gwon, W. Campbell, K. Brady, D. Sturim, M. Cha, and H. Kung, "Multimodal sparse coding for event detection," *NIPS MMML*, 2015.

[109] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[110] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, 2011, pp. 2546–2554.

[111] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressive sensing," *arXiv preprint arXiv:0901.3403*, 2009.

[112] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[113] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable bayesian models for promoting sparsity," *Information Theory, IEEE Transactions on*, vol. 57, no. 9, pp. 6236–6255, 2011.

[114] A. Boisbunon, "The class of multivariate spherically symmetric distributions," 2012.

[115] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, 2006.

[116] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[117] D. J. MacKay, "Hyperparameters: optimize, or integrate out?" *Fundamental theories of physics*, vol. 62, pp. 43–60, 1996.

[118] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*. Springer, 1998, pp. 355–368.

[119] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.

[120] Y. Marnissi, Y. Zheng, E. Chouzenoux, and J.-C. Pesquet, "A variational bayesian approach for image restoration. application to image deblurring with poisson-gaussian noise," *IEEE Transactions on Computational Imaging*, 2017.

[121] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," 2010.

[122] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2777–2824, 2011.

[123] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, pp. 3468–3497, 2009.

[124] R. Jenatton, J. Mairal, G. Obozinski, and F. R. Bach, "Proximal methods for sparse hierarchical dictionary learning." in *ICML*, no. 2010. Citeseer, 2010, pp. 487–494.

[125] G. Zhang, T. D. Roberts, and N. Kingsbury, "Image deconvolution using tree-structured bayesian group sparse modeling," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4537–4541.

[126] G. Zhang and N. Kingsbury, "Fast l0-based image deconvolution with variational bayesian inference and majorization-minimization," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 1081–1084.

[127] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[128] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[129] L. Yang, J. Fang, and H. Li, "Sparse bayesian dictionary learning with a gaussian hierarchical model," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2564–2568.

[130] H. Gunawan, O. Neswan, and W. Setya-Budhi, "A formula for angles between subspaces of inner product spaces," *Contributions to Algebra and Geometry*, vol. 46, no. 2, pp. 311–320, 2005.

[131] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 223–232.

[132] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[133] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[134] J. R. Schott, *Matrix analysis for statistics*. John Wiley & Sons, 2016.

[135] G. Stewart, "On the continuity of the generalized inverse," *SIAM Journal on Applied Mathematics*, vol. 17, no. 1, pp. 33–45, 1969.

[136] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2049–2073, 2005.

[137] E. Love, "64.4 some logarithm inequalities," *The Mathematical Gazette*, vol. 64, no. 427, pp. 55–57, 1980.

[138] I. Fedorov, R. Giri, B. D. Rao, and T. Q. Nguyen, "Relevance subject machine: A novel person re-identification framework," *arXiv preprint arXiv:1703.10645*, 2017.

[139] M. B. Christopher, *PATTERN RECOGNITION AND MACHINE LEARNING*. Springer-Verlag New York, 2016.

[140] J. Marino, Y. Yue, and S. Mandt, "Iterative amortized inference," *arXiv preprint arXiv:1807.09356*, 2018.

[141] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville, "Neural autoregressive flows," *arXiv preprint arXiv:1804.00779*, 2018.

[142] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[143] M. Yin and M. Zhou, "Semi-implicit variational inference," *arXiv preprint arXiv:1805.11183*, 2018.