Bayesian Matrix Factorization and Applications

Oumayma Dalhoumi

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

June 2022

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Oumayma Dalhoumi**

Entitled: **Bayesian Matrix Factorization and Applications**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

Dr. Mohammad Manan ______________________ Chair and Examiner

Dr. Chadi Assi ______________________ Examiner

Dr. Manar Amayri ______________________ Supervisor

Dr. Nizar Bouguila ______________________ Supervisor

Approved by ______________________

Abdessamad Ben Hamza, Chair
Department of Concordia Institute for Information Systems Engineering

______________ 2022

____

Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Bayesian Matrix Factorization and Applications

Oumayma Dalhoumi

Nonnegative matrix factorization (NMF) reduces the observed nonnegative matrix into a product of two nonnegative matrices. Nonnegativity entails two major implications: non-negative components and purely additive combination. These characteristics made this method useful in a wide range of applications. In this thesis, we propose two novel Bayesian nonnegative matrix factorization techniques.

First, we propose a model dedicated to semi-bounded data where each entry of the observed matrix is supposed to follow an Inverted Beta distribution. Latent variables of the factorized parameter matrices follow a Gamma prior. Variational Bayesian inference and lower bound approximation for the objective function are used to find an analytically tractable solution for the model. An online extension of the algorithm is also proposed for more scalability. Both models are evaluated on five different applications.

Second, we propose a Bayesian NMF that can be specifically useful for non intrusive load monitoring (NILM). NILM can be formulated as a source separation problem where the aggregated signal is expressed as linear combination of basis vectors in a matrix factorization framework. The model achieves superior performance by imposing sparsity on the activation matrix using Dirichlet priors. To estimate the parameters of the model, variational Bayesian inference is used. A novel optimization approach is proposed to find an analytically tractable solution for the model. We evaluate the model with three data sets: REDD, AMPds and IRISE, and with multiple experimental setups. The proposed model provides interpretability, flexibility and high performance.

Acknowledgments

I would like to extend my most sincere appreciation and gratitude to my supervisor, Professor Nizar Bouguila, who encouraged and supported me genuinely on my entire journey in this Master's program. As a knowledgeable, respectful and genius supervisor, he always led and motivated me with endless patience. His guidance and bits of advice in the personal, academic and professional scale will remain engraved in my thoughts forever.

I also want to express my sincere gratitude to my co-supervisor, Dr. Manar Amayri who provided me with endless encouragement, very constructive comments and valuable knowledge. I am fortunate to have had her in this journey. Thank you for the inspiration you provide.

I owe my most massive thanks to all my lab mates; Fatma, Rim, Hafsa, Ahmed Rebei, Omar Graja, Ornela, Nuha, Omar. B, Ons, Jason, Huda, Narges, Kamal, Pantea, Basim, Hussain and Meeta for the great time I spent with them during this adventure.

Last but not least I am deeply grateful to the best support system I can ever ask for: my family and friends. I owe the deepest gratitude to my parents who supported me to achieve and accomplish all my dreams and ambitions and I feel very blessed to be their daughter. To my husband who has always believed in me when I least believed in myself. To my siblings Amal and Hamza who guided and inspired me unconditionally. The fun and inspiring people around me my friends: Ahmed.M , Wissal, Anas, Mohamed, Imen, Ons, Amen, Rahma, Amine, Oussema and Oussema. A special thanks to the women who keep inspiring me, my grandmothers, my aunts, my cousins and my best friends.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The world has always been changing thanks to major breakthroughs in different domains thus the discovery of electricity, the steam engine, transistors, etc. These discoveries reshaped humans' lifestyle and yield to a sequence of scientific revolutions that created an all-time record growth. Since the end of the 20th century, we have been living in the middle of what is called the digital revolution. In fact, the immergence of new technologies allowed the collection of enormous amounts of data. In 2006, British mathematician Clive Humby claimed "Data is the new oil.". The statement has been proven right as data have become a tradable wealth in both its row or processed formats. Data are basically the fuel of the digital economy. The only convenient is that this wealth is continuously growing thanks to billions of everyday online connections among people, businesses, devices, and processes. The digital economy is continuously empowered by hyperconnectivity which means growing interconnectedness of people, organisations, and machines that results from the Internet, mobile technology and the internet of things (IoT).

Data have always been valuable. However, the breakthrough in computational and storage resources allowed the efficient usage of this data for information retrieval and decision-making processes. Advanced data analysis, data mining, pattern recognition and machine learning techniques are now the motor behind major decisions across the world. This includes political decisions, trading strategies, climate change, but also what video will show up on my youtube's home page and how much bananas will be found in the neighborhood store today.

With the emergence and the growth of this new economy, keep coming new challenges and limits. In fact, regardless of the application and the domain, modern data driven solutions have to be designed to suit certain constraints. These can be categorized into user related concerns, application specific restrictions, and implementation challenges. Firstly, users are more and more sensitive to their privacy. Data does usually contain sensitive information that the user is not comfortable revealing or sharing across platforms. Federated learning is a technique that allows a decentralized training of an algorithm. This approach is attracting more interest to preserve data privacy and to keep user's data on his/her own device. Secondly, from application perspective, one of the most common issues with machine learning algorithms is the lack of interpretability. Actually, for various applications, the explainability of classification or regression model could be a core reason behind this task and even more important than the results themselves. In business for example, understanding why your model is predicting a spike or a fall in the demand forecasting model can be crucial to decide on marketing strategies or pricing model. For energy disaggregation, the model has to provide a clear vision on how different devices impact the overall consumption. This information is used by both household and energy companies to optimize consumption and decide on efficient appliance usage. Last but not least, the collection of enormous amounts of data in different formats (image, text, audios, videos, etc.) requires appropriate tools to combine these information into significant numerical values that can be interpreted and used by a machine. Additionally, despite the improvement in computational and storage resources, the increase in data volume is always more significant. Scalable solutions have to be proposed to address the volume of data. This brings another challenge related to the choice of the adapted data solution. In fact, the structure and distribution of data are continuously changing even within the same business case. Highly parametric models can therefore be inefficient as the chosen hyper-parameters wouldn't be appropriate when data change. Usually, data scientists have a scheduled hyper-parameter tuning after certain data updates. Yet, this solution is not optimal. Generative models and Bayesian learning are more robust. With the appropriate hypothesis on the prior and posterior distributions, they can regularly auto-adjust to data changes. In this thesis, we are aiming to build machine learning solutions that are adapted to current data science challenges. We will propose methods that are scalable, can be implemented in an online framework with stream data, and can easily be adapted into a distributed context.

The increasing interest in data driven technologies have created the need for an effective way of data representation by appropriate dimensionality reduction techniques. Generally speaking, two basic properties are supposed to be satisfied: first, the dimension of the original data should be reduced; second, the principal components, hidden concepts, prominent features, or latent variables of the data, depending on the application context, should be identified efficaciously. In most of the cases, data can be presented in the format of a data matrix (or tensor) that we will be denoting as $X \in R_{N*M}$. From a mathematical stand point, dimensionality reduction techniques can be expressed as a transformation $\Phi : R_{N*M} \rightarrow R_{N*K}$, where in most cases $K << M$. This transformation can be either linear [11] or non-linear [76, 85]. Examples of linear methods include principal component analysis (PCA), singular value decomposition (SVD) [87], independent component analysis (ICA) [78], etc. Non linear models include kernelized PCA [98] and deep autoencoders [86], etc. Different methods can be differentiated based on the statistical properties of the constraints imposed on either the transformation or the resulting matrix. Yet, most of the methods listed here don't take into account structure and properties of the input data. In various real life applications, data are semi-bounded and represented as positive vectors by nature, for instance pixel intensities, amplitude spectra, occurrence counts, purchases data, user scores, stock market values, energy consumption, etc. Hence, for the sake of interpretability of the results, optimal processing of positive vectors may call for processing under nonnegativity constraints. Extending the nonnegativity constraint in dimensionality reduction models to the latent matrices helps to induce sparsity and leads to part-based decompositions. Therefore, the importance of Nonnegative Matrix Factorization (NMF) [62, 45] which adds to the properties mentioned above the non-negativity constraint. Those characteristics made NMF stand out of other methods and be useful in a wide range of applications beyond its mathematical exploration.

From evaluation perspective, evaluating machine learning methods highly depends on the nature of the application and its usage. A wide range of metrics have been discussed and used in literature [102, 13, 31]. In this work, we were able to assess the performance of our models using application specific methods by running different experiments. We provide specific analysis for each of the generated results and inspect them from an overall as well as a detailed perspective. Data cleaning and preprocessing have been performed for each of the applications.

1.2 Mathematical formulation of NMF

From mathematical stand point, NMF reduces an observed nonnegative matrix into a product of two nonnegative matrices called excitation and basis matrices **U** and **V**, respectively. The problem has been traditionally formulated as finding U and V such that: $X = UV$ where, $U > 0$ and $V > 0$. The idea of NMF has initially been introduced as Positive Matrix Facorization by Paatero and Tapper [62, 61]. The existance of such factorization was proven by the theory of Completely Positive (CP) factorization [83]. The initial models however suffered from convergence and computational issues. In fact, resolving NMF and formulating it as a convex optimization problem are still open research topics. As of the time of writing this thesis, NMF is still an ill-posed problem with non unique solutions [1]. Various models have therefore been proposed to improve the convergence and the performance of NMF and to also take into account other constraints such as sparcity, orthogonality, etc. [62, 45, 73, 93, 77, 83, 53, 29]. Various ways have been proposed to resolve NMF. The conventional NMF algorithms seek to maximize the similarity between X and the product of latent matrices UV by defining and optimizing a similarity measure $D(X\|\|UV)$. Examples of these functions are Frobenius norm and I-divergence [46]. Given the NP-hardness of the problem, the most common optimization approaches apply iterative multiplicative updates similar to expectation maximization algorithms and are inspired from the SED-MU and GKLDMU proposed in [46]. Convergence rate of these algorithms is yet to be improved. Optimization techniques are usually used to address this issue.

NMF has also been treated from a statistical point of view. A propablistic NMF model has initially been suggested by Mnih et al. in [56] where the similarity is determined based on a prior knowledge about the probability distribution of the noise. This model paved the way to Bayesian non-negative matrix factorization (BNMF) and the first approach was proposed by Salakhutdinov et al. in [69]. Generative models can properly reflect the statistical structure of the signal and the disclosed components. In the recent years, more attention has been brought to generative modeling in various AI models [68, 57]. Specifically for NMF, Bayesian learning allowed for more flexibility in modeling, adding further constraints and generating more robust solutions. In fact, unless regularization parameters are tuned carefully, discriminative models are prone to overfitting because they find a single point estimate of the parameters. Bayesian inference can usually resolve this issue and provide robust solutions [75]. A fully Bayesian treatment of a probabilistic matrix factorization was presented in [69, 17] where the model was trained using Markov Chain Monte Carlo (MCMC) methods. Variational inference is a scalable alternative to MCMC for Bayesian posterior inference. It has been proposed and tested in various domains [88, 5]. Variational inference was used to infer latent variables for Bayesian NMF in [63, 53, 29]. Unlike conventional Bayesian NMF, work in [53, 29] applied matrix factorization on the model parameters. This allows more flexibility and enables imposing further constraints on the factorization model.

1.3 Applications of NMF

Beyond its mathematical formulation, the philosophy behind NMF is to propose a feasible model for learning object parts. Parts-based representation is considered among the fundamentals of certain computational theories of recognition problems: perception of the whole is based on perception of its parts. Basically, nonnegativity entails two major implications: non-negative components and purely additive combination.

Firstly, as explained above negative data in both observations and latent components are irrelevant in various real life applications as the corresponding data are semi-bounded and represented by positive vectors by nature. Extracted positive embeddings do commonly correspond with semantic and meaningful interpretations. Secondly, objects of interest are most naturally characterized by the inventory of its parts, and the exclusively additive combination means that they can be reassembled by adding required parts together. Thanks to it's simple yet powerful characteristics, NMF achieved high performance in wide range of real life applications: natural language processing (NLP), collaborative filtering (CF), graph embedding, non intrusive load monitoring (NILM), pattern recognition, etc.

1.3.1 Collaborative filtering

Collaborative filtering (CF) is widely used in recommender systems. Recommender systems are a type of information filtering systems that involve predicting user responses. CF is based on the idea that people who liked an item on the past are very likely to agree on it in the future. Matrix factorization resolves CF by factorizing a utility function X where rows and columns represent users and items respectively. The inferred latent vectors correspond then to users and items embeddings and they represent their hidden characteristics. The rating of user i for item j can be reconstructed by multiplying the latent vectors that correspond to that user item combination. Collaborative filtering has traditionally been resolved by dot products and NMF represented the state of the art for a long time. With the development of deep learning approaches, He and al. proposed neural collaborative filtering (NCF) in [26]. Despite the wide success of deep collaborative filtering, researchers from google have demonstrated that with the proper parametrization, an NMF still outperfoms NCF [65]. In fact, despite their theoretical capabilities, NCF are costly, more complex to use in production environment and are prone to bias.

1.3.2 Natural language processing

In NLP, NMF was widely used for document clustering and topic modeling. Authors in [30] surveyed the usage of NMF methods for documents clustering. The usage of different algorithms showed superior performance compared to state of the art methods like spectral clustering in both accuracy and latent semantic topic identification. NMF was also applied for topic modeling. The returned basis matrix represents prominent topics contained in a document corpus [33, 3]. Authors in [8] also showed the efficiency and the potential of Poisson NMF on topic modeling.

1.3.3 Non intrusive load monitoring

Load disaggregation is the task of extracting single appliances' power consumption out of the aggregated power data using one single energy meter. Interest in understanding appliance level consumption has risen with current energy and climate challenges. As power consumption is positive, NMF provides the logic to recover individual components' power consumption through the learned latent basis. Works in NMF for energy aggregation showed efficiency and high performance [72, 64, 55]. Most of the proposed methods are however discriminative and lack generalization capabilities.

1.3.4 Parts based decomposition

In object recognition and computer vision applications, the learned basis images are localized rather than holistic and they can be mapped to actual parts of the original images [89, 59, 10]. Authors in [32, 21] have shown that setting a high sparseness value for the basis images results in a local representation of an imput image.

1.3.5 Graph embeddings

Graph embedding learning aims to automatically learn low-dimensional node representations. Among the applications of graph embedding is link prediction. It can be defined as: given a set of biomedical entities and their known interactions, we aim to predict other potential interactions between entities [52]. Matrix factorization techniques resolve this problem by factorizing the link matrix to learn low-dimensional representations in a latent space. A binary classification is then performed to decide if a link exists between the biological elements.

1.4 Contributions

As discussed in the previous sections, NMF is still an ill-defined problem. Yet, it is very efficient for a wide range of applications and is highly competitive even compared to high performing deep learning approaches. The main goal of this thesis is to propose new non negative matrix factorization models that are capable to address modern machine learning challenges. The contributions are listed as follows:

- **Bayesian Matrix Factorization for Semi-bounded Data**

  A novel Bayesian nonnegative matrix factorization technique dedicated to semi-bounded data where each entry of the observed matrix is supposed to follow an Inverted Beta distribution. The model has two parameter matrices with the same size as the observation matrix which we factorize into a product of excitation and basis matrices. Entries of the corresponding basis and excitation matrices follow a Gamma prior. To estimate the parameters of the model, variational Bayesian inference is used. A lower bound approximation for the objective function is used to find an analytically tractable solution for the model. An online extension of the algorithm is also proposed for more scalability and to adapt to streaming data. The model is evaluated on five different applications: parts-based decomposition, collaborative filtering, market basket analysis, transactions prediction and items classification, topic mining and graph embedding on biomedical networks.

This work has been published in IEEE Transactions on Neural Networks and Learning Systems [12]

- **Bayesian Non-negative Matrix Factorization for Non-Intrusive Load Monitoring**

  A Bayesian non-negative matrix factorization approach is proposed. We assume a generative model where each matrix element follows an exponential distribution. Exponential distribution (exp) with support $(0, \infty)$ can be used to model non-negative real variables. The matrix is modeled in a way to impose a sparsity constraint on the excitation matrix A which is guaranteed through the sum to k. To model this constraint on a Bayesian space, we assume that the coefficients of A_K, weights for device k, follow a Dirichlet distribution of parameters that are subject to our optimization problem. The matrix factorization is applied on the model parameter instead of directly applying it on the observed matrix. A Dirichlet prior is associated to the matrix A. We propose a novel approximation method using mean field variational inference to learn the model and estimate the parameters. The proposed model is evaluated with different applications and tested against different baselines with multiple datasets: REDD dataset, AMPds and IRISE dataset. The proposed model shows high performance against various supervised learning approaches. It is high performing for low frequency setup. Moreover, the proposed learning process is low dependent on other observations and therefore can be easily adapted in a federated learning framework.

1.5 Thesis Overview

- In chapter 1, we introduce non negative matrix factorization. We discuss different mathematical formulations of NMF as well as various applications. We present the current state of the art and its limitations. We also discuss the motivations behind our work.

- In chapter 2, we propose a novel Bayesian NMF model. We use variational learning and provide a novel lower bound approximation for the objective function to find an analytically tractable solution for the model. An online extension of the model is afterward proposed. The efficiency of the model has been evaluated by comparing it to state of the art models on five different applications.

- In chapter 3, we present a Bayesian NMF approach dedicated to non intrusive load monitoring. We explain the optimization model. The performance of our model is tested against several approaches. We used different datasets and metrics for evaluation to take into account different circumstances.

- In conclusion, we briefly summarize our contributions.

Chapter 2

Bayesian Matrix Factorization for Semi-bounded Data

2.1 Introduction

With the increasing amount of available raw data, due to the development in sensor and computer technology, rises the need of dimensionality reduction techniques. The observed data are usually organized in a form of matrix or tensor. Thus, from an algebraic perspective, dimensionality reduction can be interpreted as decomposing this matrix into the product of two matrices. To accomplish this task, several methods have been proposed such as SVD [20], PCA [92], ICA [82], etc... In general these methods have to satisfy two properties: reducing the dimensionality of the original data, and preserving the hidden concepts and latent variables of the data. Yet, they usually don't take into account structure and properties of the data. In various real life applications, data are semi-bounded and represented as positive vectors by nature, for instance pixel intensities, amplitude spectra, occurrence counts, purchases data, user scores, stock market values, etc. Hence, for the sake of interpretability of the results, optimal processing of positive vectors may call for processing under nonnegativity constraints. Extending the nonnegativity constraint in dimensionality reduction models to the factor matrices helps to induce sparsity and leads to part-based decompositions. Therefore, the importance of Nonnegative Matrix Factorization (NMF) [62, 45] which adds to the properties mentioned above the non-negativity constraint.

NMF reduces the observed nonnegative matrix into a product of two nonnegative matrices called excitation and basis matrices that we'll refer to as **U** and **V**, respectively. This property has shown a great utility in several applications such as visual features learning [101], face recognition [51], source separation [60], collaborative filtering and recommender systems [43], document clustering and topic mining [95, 74], etc. Observed data can usually be modeled in the form of a matrix $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M)$, where $\boldsymbol{x}_j$ is a column vector of size N. When requiring both the dimensionality and the factorization rank of **X** to increase, the optimization problem of NMF is NP-hard [84]. Therefore, only a local minimum is achievable in a reasonable computational time. Classically, NMF is approached by maximizing the similarity between **X** and the product **UV** by performing alternating minimization of a suitable cost function. Several algorithms have been proposed to tackle NMF [90, 49]. Authors in [46] used Frobenius norm and I-divergence as optimization functions. They proposed two algorithms by minimizing l_2-norm and Kullback-Leibler (KL) divergence. Unless regularization parameters are tuned carefully, classical models are prone to over-fitting because they find a single point estimate of the parameters. Bayesian inference can usually resolve this issue and provide robust solutions [75]. A fully Bayesian treatment of a probabilistic matrix factorization was presented in [69, 73, 17] where the model was trained using Markov Chain Monte Carlo (MCMC) methods. Variational inference is a scalable alternative to MCMC for Bayesian posterior inference. It has been proposed and tested in various domains [88, 5]. Variational inference was used to infer latent variables for Bayesian NMF in [63, 53, 29]. Unlike conventional Bayesian NMF, work in [53, 29] applied matrix factorization on the model parameters. This allows more flexibility and enables imposing further constraints on the factorization model.

Most of NMF applications involve large datasets where scalability is an issue. In general, data continuously arrive in streams or batches. Online learning is a well known solution for the problems mentioned above. Online approach processes the data, one at a time, or in mini-batches. This is particularly important in the context of image and video processing and news text mining. The standard variational Bayes formulation is adapted to the online setting by stochastic coordinate ascent. The online learning gives two advantages: only a limited amount of data needs to be stored at a time in memory, independently of the size of the original dataset; by processing the data in a random sequence, we gain robustness to local optima and maintain convergence guarantees.

In this chapter, a Bayesian nonnegative matrix factorization approach is proposed. We assume a generative model where each matrix element follows an Inverted Beta distribution. Inverted Beta (IB) with support $(0, \infty)$ can be used to model nonnegative real variables [44]. IB has a flexible shape and can be symmetric or highly skewed. The model has two parameter matrices. The matrix factorization is applied on the model parameters instead of directly applying it on the observed matrix. Each parameter matrix is factorized into a product of excitation and basis matrices. Correlation between the parameter matrices is modeled by setting the excitation matrix to be the same for both matrices. A Gamma prior is associated to each entry, thus the naming IBG-NMF. Variational inference with a lower bound approximation is proposed to learn the model and estimate the parameters. Due to the properties of the Gamma distribution, the sparseness constraint can be achieved by imposing a low shape parameter on either the excitation or the basis matrix. An online extension of the model is proposed to allow for more scalability and to adapt to streaming data. The proposed models are evaluated with different applications and tested against different baselines with multiple datasets: parts-based decomposition, collaborative filtering, market basket analysis, transactions prediction and items classification, topic mining and graph embedding on biomedical networks. Our application for transactions prediction and items classification allows a novel interpretation for market basket analysis. It allows using the NMF to model market baskets in more efficient way than the traditional association rules. This approach can also be used for e-commerce recommender systems to recommend items based on those already selected in a basket and not only the user's historical purchases.

The rest of this chapter is organized as follows: a general introduction to nonnegative matrix factorization with a brief litterature review is given in Section 2. In Section 3, the generative model and model specifications are introduced. In sections 4 and 5, we describe the IBG-NMF and online IBG-NMF solutions and algorithms. Experiments, results and comparisons are presented in section 6. Finally, conclusion is drawn in Section 7.

2.2 Nonnegative Matrix Factorization

In this chapter, we refer to matrices by upper case bold letters as for **X**. A vector of **X** is denoted as lower case bold italic letters $\boldsymbol{x}_i$ and elements of **X** are x_{ij}. All other scalars are denoted by roman non-bold letters. Model parameters are denoted by greek letters.

Observed data can usually be modeled in the form of a matrix $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_M)$, where $\boldsymbol{x}_j$ is a column vector of size N. NMF, introduced in [45] is a dimentionality reduction technique that reduces a nonnegative matrix **X** into a factor of a nonnegative basis matrix **U** and a nonnegative excitation matrix **V** such that:

$$\mathbf{X} \approx \mathbf{UV}, \tag{2.1}$$

where: $\mathbf{X} \in \mathbb{R}_{+}^{N \times M}$, $\mathbf{U} \in \mathbb{R}_{+}^{N \times K}$ and $\mathbf{V} \in \mathbb{R}_{+}^{K \times M}$. The columns of X are a linear combination of all the columns of **U**, with weighting coefficients from the columns of **V**, $\boldsymbol{x}_j \approx \sum_{i=1}^{K} \boldsymbol{u}_i V_{ij}$. We usually have $K << M, N$. This generally results in part-based representations.

The existence of NMF solution was proved via the theory of completely positive factorization in [83]. Based on the conventional NMF, many extensions [93, 77, 48] have been studied, and several constraints can be added on the NMF to enhance the reconstruction performance, such as the orthogonal constraint [50], the sparse constraint [21], and the discriminant constraint [36]. NMF can be treated in the conventional way as well in a probabilistic way where we try to estimate the parameters of the underlying model instead of estimating the basis and excitation matrices directly. Due to the NP-hardness of the problem, a unique solution is not achievable in a reasonable computational time [90], and only local minima can therefore be achieved. A Bayesian estimation, can be used instead. The authors in [69, 73, 17] presented a fully Bayesian treatment of probabilistic matrix factorization trained using MCMC methods that showed significant accuracy levels on Netflix dataset. Variational inference is a scalable alternative to MCMC for Bayesian posterior inference. It has been proposed and tested in various domains [88, 5]. Variational inference was used to infer latent variables for Bayesian NMF in [63, 53, 29]. Authors in [100] have reformulated the classical matrix factorization problem to a hierarchical Bayesian generative model and utilized variational inference to infer model parameters. This approach showed robustness and effectiveness with different applications. Unlike conventional Bayesian NMF, [53, 29], applied matrix factorization on the model's parameters. This allows for more flexibility and enables imposing further constraints on the factorization model.

In order to scale up larger datasets and to take into account stream data we propose an online extension to our algorithm. Online approaches process the data, one at a time, or in mini-batches. For the online schemes of NMF, Guan et al. [23] proposed an approach using robust stochastic approximation. Lefevre et al. [47] proposed an online updated algorithm with Itakura–Saito divergence to estimate NMF solutions. For NMF, Gu et al. proposed a fast two-stage algorithm for non-negative matrix factorization in streaming data [22]. In this chapter, the standard variational Bayes formulation is adapted to online settings by stochastic coordinate ascent. In section 2.3, we propose a Bayesian framework for nonnegative matrix factorization.

2.3 Model Specification

Data generated by real life applications, such as sales data, collaborative filtering data, representations of texts and image data, etc, are semi bounded by nature. Therefore, for the sake of interpretability of the results, optimal processing of this kind of data may call for processing under nonnegativity constraints. In this section, we propose a novel model for Bayesian matrix factorization for semi-bounded data.

2.3.1 Generative Model

We assume that each nonnegative data point x_{nm} is generated from an IB distribution, with parameters u_{nm} and v_{nm}. For an observation matrix **X** we have two parameter matrices **U** and **V** all of size $(N \times M)$. Similar to [53, 29], we jointly factorize each parameter instead of the observation matrix as:

$$\begin{gathered}\mathbf{U} \approx \mathbf{AH} \\ \mathbf{V} \approx \mathbf{BH} \qquad (2.2) \\ \mathbf{A}, \mathbf{B} \in \mathbb{R}_{+}^{(N,K)}, H \in \mathbb{R}_{+}^{(K,M)}\end{gathered}$$

Given the nonnegativity property of **A**, **B** and **H**, we can assign a Gamma prior to each entry. With the above, we can create the following generative model:

$$\begin{aligned}
a_{n,k} &\sim Gamma(a_{nm}|\mu_{nk}, \alpha_{nk}) \\
b_{n,k} &\sim Gamma(b_{nk}|\nu_{nk}, \beta_{nk}) \\
h_{k,m} &\sim Gamma(h_{km}|\rho_{km}, \zeta_{km}) \\
x_{n,m} &\sim iBeta(x_{nm}| \sum_k a_{nk}h_{km}, \sum_k b_{nk}a_{km})
\end{aligned} \tag{2.3}$$

where $Gamma(x|k, \theta)$ is the Gamma density function with parameters k and θ, and $iBeta(x|\alpha, \beta)$ is the IB density function with parameters α and β. We have :

$$Gamma(x|k, \theta) = \frac{\theta^k}{\Gamma(k)} x^{k-1} e^{-\theta x}, k, \theta > 0 \tag{2.4}$$

$$iBeta(x|\alpha, \beta) = \frac{x^{\alpha-1}(1+x)^{-\alpha-\beta}}{\mathbb{B}(\alpha, \beta)} \tag{2.5}$$

where $\Gamma(.)$ is the Gamma function and $\mathbb{B}(., .)$ is the Beta function defined as $\mathbb{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. IB is flexible and allows different shapes (i.e.. symmetric, skewed to the right, skewed to the left). The distribution can be a good representation for sparse, heavy tailed or symmetric semi bounded data based on the choice of the model parameters a and b.

We define the latent variable $\mathbf{Z}$ such that: $\mathbf{Z} = \{\mathbf{A}, \mathbf{B}, \mathbf{H}\}$. The posterior distribution is then $p(\mathbf{Z}|\mathbf{X})$. $\mathbf{Z}$ is divided into disjoint groups $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{H}$ and a Gamma prior has been assigned to each entry by definition of the generative model in (2.3). We end up with the following equations:

$$\begin{aligned}
p(\mathbf{Z}|\mathbf{X}) &= \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{X})} \\
p(\mathbf{Z}) &= p(\mathbf{A})p(\mathbf{B})p(\mathbf{H}) \\
p(\mathbf{A}) = \prod_{n,k} p(a_{nk}), p(\mathbf{B}) &= \prod_{n,k} p(b_{nk}), p(\mathbf{H}) = \prod_{k,m} p(h_{km}) \\
p(\mathbf{X}|\mathbf{Z}) &= \prod_{n,m} \frac{1}{\mathbb{B}(\sum_k a_{nk}h_{km}, \sum_k b_{nk}h_{km})} \\
&\times (x_{nm})^{\sum_k a_{nk}h_{km}-1} \\
&\times (1+x_{nm})^{-\sum_k a_{nk}h_{km} - \sum_k b_{nk}h_{km}}
\end{aligned} \tag{2.6}$$

2.3.2 Matrix Decomposition

Depending on the application, there might be a need to have the following form $\mathbf{X} = \mathbf{UV}$. Based on the model proposed in Eq (2.3) we have:

$$\bar{x}_{nm} = \frac{\sum_k a_{nk} h_{km}}{\sum_k b_{nk} h_{km} - 1} \tag{2.7}$$

$$\bar{\mathbf{X}} = \mathbf{AH} \oslash (\mathbf{BH} - 1) \tag{2.8}$$

where $\oslash$ is the element-wise division.

If we impose the sparseness constraint on the excitation matrix $\mathbf{H}$ (column sparseness equal to 1), we can decompose the matrix as in (2.9). Details of this decomposition can be found in Appendix B.

$$\bar{\mathbf{X}} = \mathbf{AH} \oslash (\mathbf{BH} - 1) = (\mathbf{A} \oslash (\mathbf{B} - 1))\mathbf{H} \tag{2.9}$$

The sparseness constraint can be achieved by choosing low values of the shape parameter. In fact, the excitation matrices A and $\mathbf{B}$ and the basis matrix $\mathbf{H}$ follow a Gamma prior. The Gamma distribution has two parameters: shape parameter k and a scale parameter θ. The expectation value is $\frac{k}{\theta}$. For a given mean value, a small shape parameter forces the variable to have a very high probability around zero. Thus, the sparseness constraint can be achieved. After normalizing the matrices, the condition for (2.9) can be satisfied.

2.4 Batch Variational Bayesian Inference

The main goal of Bayesian analysis is to infer the posterior distribution given the prior distribution and the data likelihood. However, the exact Bayesian inference of the generative model stated in (2.3) is not analytically tractable. So, we shall use variational inference to learn our model. Our problem is equivalent to maximizing an objective function $\mathcal{L}(q) = E_{q(z)}[\ln(\frac{p(\mathbf{Z},\mathbf{X})}{q(\mathbf{Z})})]$. This leads to an optimal solution $q^*(a_{nk})$ expressed in (A.2) calculation details can be found in Appendix A, where $\bar{x}$ denotes the expected value of x and $E_{\backslash q(a_{nk})}[X] = \int \prod_{(i,j)\neq(n,k)} Xq(a_{ij})da_{ij}$:

$$\begin{aligned}
\ln q^*(a_{nk}) &= E_{\backslash q(a^*_{nk})}[\ln p(\mathbf{X}, \mathbf{Z})] + const \\
&= \sum_m E_{\backslash q(a^*_{nk})}[-\ln \mathbb{B}(\sum_k a_{nk}h_{km}, \sum_k b_{nk}h_{km})] \\
&+ \sum_m \bar{h}_{km} \ln x_{nm}a_{nk} - \sum_m \bar{h}_{km} \ln(1 + x_{nm})a_{nk} \\
&+ (\mu_0 - 1)\ln(a_{nk}) - \alpha_0 a_{nk} + const
\end{aligned} \tag{2.10}$$

One way to find an analytically tractable solution is by identification, where we separate elements of (A.2) that only include a_{nk} from elements that only include $\ln a_{nk}$. However, due to the integral expression in the Γ function, the expectation of $\ln \mathbb{B}(.)$ is not analytically tractable. Thus, an analytically tractable solution cannot be obtained directly. According to the extended factorized approximation [29, 37, 35], we can find a lower bound to the objective function $\mathcal{L}(q)$. Maximizing this lower bound is asymptotically equivalent to maximizing the objective function $\mathcal{L}(q)$.

In this chapter we use Extended Factorization Approximation (EFA) to derive an analytically tractable solution to the Bayesian Estimation of IBG-NMF. We find an auxiliary function $E_{q(\mathbf{Z})}[\ln \widetilde{p}]$, such that:

$$E_{q(\mathbf{Z})}[\ln p(\mathbf{X}, \mathbf{Z})] \geq E_{q(\mathbf{Z})}[\ln \widetilde{p}(\mathbf{X}, \mathbf{Z})] \tag{2.11}$$

We can write:

$$E_{q(\mathbf{Z})}\Big[\ln(p(\mathbf{X}, \mathbf{Z}))\Big] = E_{q(\mathbf{Z})}\Big[\sum_{n,m}(f_{nm} + r_{nm})\Big] \tag{2.12}$$

With:

$$
\begin{aligned}
E_{q(\mathbf{Z})}[f_{nm}] &= E_{q(\mathbf{Z})}[-\ln \mathbb{B}(\sum_k a_{nk}h_{km}, \sum_k b_{nk}h_{km})] \\
E_{q(Z)}[r_{nm}] &= E_{q(Z)}\Big[\sum_k (a_{nk}h_{km} - 1)\ln x_{nm} \\
&\quad - \sum_k (a_{nk}h_{km} + b_{nk}h_{km})\ln(1 + x_{nm}) \\
&\quad + \sum_k \big[(\mu - 1)\ln a_{nk} + (\nu - 1)\ln b_{nk} \\
&\quad + (\rho - 1)\ln h_{km} - \alpha a_{nk} - \beta b_{nk} - \zeta h_{km}\big]\Big] \\
&\quad + const
\end{aligned}
\tag{2.13}
$$

Therefore, a lower bound of the objective function is:

$$
E_{q(\mathbf{Z})}[\ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}] \geq E_{q(\mathbf{Z})}[\ln \widetilde{p}(\mathbf{X}, \mathbf{Z})] - E_{q(\mathbf{Z})}[\ln q(\mathbf{Z})] \tag{2.14}
$$

Our goal is to find an analytically tractable solution for the objective function via the lower bound. The term r_{nm} is analytically tractable. However, we cannot resolve f_{nm}. In the next section we try to find a lower bound for f_{nm}.

2.4.1 A Lower Bound Approximation

In [53], the authors showed that $-\ln \mathbb{B}(\sum_k x_k, \sum_k y_k)$ is convex relative to $\ln x$ for arbitrary y, if and only if $\sum_k y_k > 1$. With this relative convexity, and by restricting that $\sum_k a_{nk}h_{km}$ and $\sum_k b_{nk}h_{km}$ are greater than 1, we can use Jensen inequality and first order Taylor decomposition to find a lower bound for the expectation of the objective function.

The first-order expansion of $E_{q(\boldsymbol{x},\boldsymbol{y})}[f_{nm}]$ function with respect to $\ln \boldsymbol{x}$ around $\ln \bar{\mathbf{x}}$:

$$
\begin{aligned}
E_{q(\boldsymbol{x},\boldsymbol{y})}[f_{nm}] \geq E_{q(\boldsymbol{x},\boldsymbol{y})}\Big[&- \ln \mathbb{B}(\sum_k \bar{x}_k, \sum_k y_k)\Big] \\
&+ E_{q(\boldsymbol{x},\boldsymbol{y})}\Big[\Big(\psi(\sum_k(\bar{x}_k + y_k)) - \psi(\sum_k \bar{x}_k)\Big) \\
&\times \sum_k \bar{x}_k(\ln x_k - \ln \bar{x}_k)\Big]
\end{aligned} \tag{2.15}
$$

where ψ is the Digamma function defined as $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$.

As $-\ln(\mathbb{B}(x,y))$ in (2.15) is also relative convex to $\ln y$ for any x, we can further lower bound the LIB function as:

$$
\begin{aligned}
E_{q(\boldsymbol{x},\boldsymbol{y})}[F_{nm}] \geq E_{q(\boldsymbol{x},\boldsymbol{y})}\Big[&- \ln \mathbb{B}(\sum_k \bar{x}_k, \sum_k \bar{y}_k)\Big] \\
&+ E_{q(\boldsymbol{x},\boldsymbol{y})}\left[\Big(\psi(\sum_k(\bar{x}_k + \bar{y}_k) - \psi(\sum_k \bar{y}_k))\Big) \times \sum_k \bar{y}_k(\ln y_k - \ln \bar{y}_k)\right] \\
&+ E_{q(\boldsymbol{x},\boldsymbol{y})}\left[\Big(\psi(\sum_k(\bar{x}_k + y_k) - \psi(\sum_k \bar{x}_k))\Big) \times \sum_k \bar{x}_k(\ln x_k - \ln \bar{x}_k)\right]
\end{aligned} \tag{2.16}
$$

In the equation above, the first term (2.16) is a constant. The second term (**??**) depends only on the variable $\boldsymbol{y}$. However, the third term depends on both $\boldsymbol{x}$ and $\boldsymbol{y}$ which are not mutually independent and their expectations cannot be carried out separately. However, as x_i and y_j are independent for $i \neq j$, (**??**) term can be written as follows:

$$\begin{aligned}
&E_{q(\boldsymbol{x},\boldsymbol{y})}\Bigg[\Big[\psi\big(\sum_k(\bar{x}_k+y_k)\big)-\psi\big(\sum_k\bar{x}_k\big)\Big] \\
&\qquad\qquad\times\sum_k\bar{x}_k(\ln x_k-\ln\bar{x}_k)\Bigg] \\
&\geq E_{q(\boldsymbol{x},\boldsymbol{y})}\Big[\psi\big(\sum_k(\bar{x}_k+y_k)\big)-\psi\big(\sum_k\bar{x}_k\big)\Big] \\
&\qquad\qquad\times E_{q(x,y)}\Big[\sum_k\bar{x}_k(\ln x_k-\ln\bar{x}_k)\Big] \\
&=E_{q(\boldsymbol{y})}\Big[\psi\big(\sum_k(\bar{x}_k+y_k)\big)-\psi\big(\sum_k\bar{x}_k\big)\Big] \\
&\qquad\qquad\times\sum_k E_{q(x_k)}\Big[\bar{x}_k(\ln x_k-\ln\bar{x}_k)\Big]
\end{aligned}\tag{2.17}$$

Since $\psi(x)$ and $\ln(x)$ are concave functions in x, we can get the following inequalities by applying Jensen's inequality:

$$\begin{aligned}
E_{q(x)}[\ln(x)] &\leq \ln(E_{q(x)}[x])=\ln(\bar{x}) \\
E_{q(x)}[\psi(x)] &\leq \psi(E_{q(x)}[x])=\psi(\bar{x})
\end{aligned}\tag{2.18}$$

Equation (2.17) can be rewritten as:

$$\begin{aligned}
&E_{q(\boldsymbol{x},\boldsymbol{y})}\Bigg[\Big[\psi\big(\sum_k(\bar{x}_k+y_k)\big)-\psi\big(\sum_k\bar{x}_k\big)\Big] \\
&\times\sum_k\bar{x}_k(\ln x_k-\ln\bar{x}_k)\Bigg] \\
&=\Big[\psi\big(\sum_k(\bar{x}_k+y_k)\big)-\psi\big(\sum_k\bar{x}_k\big)\Big] \\
&\times\sum_k E_{q(x_k)}\Big[\bar{x}_k(\ln x_k-\ln\bar{x}_k)\Big]
\end{aligned}\tag{2.19}$$

Finally, the expectation F_{nm} can be lower-bounded as:

$$
\begin{aligned}
E_{q(\boldsymbol{x},\boldsymbol{y})}[F_{nm}] \geq E_{q(\boldsymbol{x},\boldsymbol{y})}\Big[& -\ln \mathcal{B}(\sum_k \bar{x}_k, \sum_k \bar{y_k})\Big] \\
& + \Bigg[\Big[\psi\big(\sum_k(\bar{x}_k + \bar{y_k})\big) - \psi(\sum_k \bar{y_k})\Big] \\
& \times \sum_k \bar{y}_k(E_{q(y_k)}[\ln y_k] - \ln \bar{y}_k)\Bigg] \\
& + \Bigg[\Big[\psi\big(\sum_k(\bar{x}_k + \bar{y_k})\big) - \psi(\sum_k \bar{x}_k)\Big] \\
& \times \sum_k \Big[\bar{x}_k(E_{q(x_k)}[\ln x_k] - \ln \bar{x}_k)\Big]\Bigg]
\end{aligned} \tag{2.20}
$$

Therefore, by substituting x_k and y_k in (2.20) by $\sum_k a_{nk}h_{km}$ and $\sum_k b_{nk}h_{km}$, we can write the approximation as:

$$
\begin{aligned}
& E_{q(\mathbf{Z})}[f_{nm}] \geq \Big[- \ln \mathcal{B}(\sum_k \bar{a}_{nk}\bar{h}_{km}, \sum_k \bar{b}_{nk}\bar{h}_{km})\Big] \\
& + \Big[\psi(\sum_k(\bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km})) - \psi(\sum_k \bar{a}_{nk}\bar{h}_{km})\Big] \\
& \times \sum_k \bar{a}_{nk}\bar{h}_{km}\Big[E_{q(a_{nk})q(h_{nk})}\big[\ln(\bar{a}_{nk}\bar{h}_{km}) - \ln(\bar{a}_{nk}\bar{h}_{km})\big]\Big] \\
& + \Big[\psi(\sum_k(\bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km})) - \psi(\sum_k \bar{b}_{nk}\bar{h}_{km})\Big] \\
& \times \sum_k \bar{b}_{nk}\bar{h}_{km}\Big[E_{q(b_{nk})q(h_{nk})}\big[\ln(\bar{b}_{nk}\bar{h}_{km}) - \ln(\bar{b}_{nk}\bar{h}_{km})\big]\Big] \\
& = E_{q(Z)}[\widetilde{f}_{nm}]
\end{aligned} \tag{2.21}
$$

2.4.2 Optimal Estimation via the EFA

Based on (2.12) and (2.21), a lower bound for $E_{q(\mathbf{Z})}\Big[\ln(p(\mathbf{X}, \mathbf{Z}))\Big]$ can be written as:

$$
E_{q(\mathbf{Z})}[\ln \widetilde{p}(\mathbf{X}, \mathbf{Z})] = E_{q(\mathbf{Z})}\Big[\sum_{n,m}(\widetilde{f}_{nm} + r_{nm})\Big] \tag{2.22}
$$

From (2.14), the objective function that we want to maximize can be lower bounded as:

$$E_{q(Z)}\Big[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\Big] \geq E_{q(\mathbf{Z})}\Big[\sum_{n,m}(\widetilde{f}_{nm}+r_{nm})\Big] - E_{q(\mathbf{Z})}[q(\mathbf{Z})] \tag{2.23}$$

Based on the principles of the variational inference framework, we can find the optimal $q^*(a_{nk})$, $q^*(b_{nk})$ and $q^*(h_{km})$, and the optimal updates are:

$$\ln q^*(a_{nk}) = E_{\backslash(q(a_{nk}))}\Big[\sum_{m}(\widetilde{f}_{nm}+r_{nm})\Big] + const \tag{2.24}$$

$$\ln q^*(b_{nk}) = E_{\backslash(q(b_{nk}))}\Big[\sum_{m}(\widetilde{f}_{nm}+r_{nm})\Big] + const \tag{2.25}$$

$$\ln q^*(h_{km}) = E_{\backslash(q(h_{km}))}\Big[\sum_{n}(\widetilde{f}_{nm}+r_{nm})\Big] + const \tag{2.26}$$

$\widetilde{\mathbf{F}}$ and $\mathbf{R}$ are replaced by their expressions in (2.24), (2.25) and (2.26). To resolve $q^*(a_{nk})$ we can skip all the terms that do not contain a_{nk}. The final expression of $q^*(a_{nk})$:

$$\begin{aligned}\ln(q^*(a_{nk})) = &\Bigg\{\sum_{m}\Big[\psi(\sum_{k}\bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) \\ &- \psi(\sum_{k}\bar{a}_{nk}\bar{h}_{km})\Big]\bar{a}_{nk}\bar{h}_{km} + \mu_0 - 1\Bigg\}\ln a_{nk} \\ &- \Bigg\{\sum_{m}(\bar{h}_{km}\ln(1+x_{nm}) \\ &- \bar{h}_{km}\ln x_{nm}) + \alpha\Bigg\}a_{nk} + const\end{aligned} \tag{2.27}$$

Equation (2.27) has the logarithmic format of the Gamma density function. It is possible to separate terms with $\ln a_{nk}$ and terms with a_{nk} in (2.27). The shape and the scale parameters μ and

α can be computed as:

$$\begin{aligned} \mu &= \mu_0 + \sum_m \big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) \\ &\quad - \psi(\sum_k \bar{a}_{nk}\bar{h}_{km})\big]\bar{a}_{nk}\bar{h}_{km} \\ \alpha &= \alpha_0 + \sum_m (\bar{h}_{km}\ln(1+x_{nm}) - \bar{h}_{km}\ln x_{nm}) \end{aligned} \tag{2.28}$$

We can verify that the shape and scale parameters of the Gamma distribution are positive. In fact, $\psi(.)$ and $\ln()$ are increasing functions. Then, terms $\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) - \psi(\sum_k \bar{a}_{nk}\bar{h}_{km})$ and $\ln(1+\mathbf{X}) - \ln(\mathbf{X})$ are always positive. $\ln q^*(b_{nk})$ and $\ln q * (h_{km})$ can be obtained similarly as follows:

$$\begin{aligned} \ln(q^*(b_{nk})) = \Big\{ &\sum_m \big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) \\ &- \psi(\sum_k \bar{b}_{nk}\bar{h}_{km})\big]\bar{b}_{nk}\bar{h}_{km} + \nu_0 - 1\Big\} \ln b_{nk} \\ &- \Big(\sum_m (\bar{h}_{km}\ln(1+x_{nm}) + \beta_0\Big) b_{nk} + const \end{aligned} \tag{2.29}$$

$$\begin{aligned} \ln(q^*(h_{km})) = &\left\{ \sum_n \Big[\big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) - \psi(\sum_k \bar{a}_{nk}\bar{h}_{km})\big]\bar{h}_{km}\bar{a}_{nk} \right. \\ &+ \big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) - \psi(\sum_k \bar{b}_{nk}\bar{h}_{km})\big]\bar{h}_{km}\bar{b}_{nk}\Big] \\ &\left. + \rho_0 - 1 \right\} \ln h_{km} \\ &- \left\{ \sum_n (\bar{a}_{nk} + \bar{b}_{nk}) \ln(1+x_{nm}) \right. \\ &\left. - \sum_n \bar{a}_{nk}\ln(x_{nm}) + \zeta_0 \right\} h_{km} + const \end{aligned} \tag{2.30}$$

2.4.3 IBG-NMF Algorithm

In order to simplify the equations we can organize the parameters in (2.28) in a matrix format. The six parameter matrices μ, α, ν, β, ρ, and ζ can be written as [1]:

$$
\begin{aligned}
\mu &= \mu_0 + \left[\psi(\bar{\mathbf{A}}\bar{\mathbf{H}} + \bar{\mathbf{B}}\bar{\mathbf{H}}) - \psi(\bar{\mathbf{A}}\bar{\mathbf{H}})\right]\bar{\mathbf{H}}^T \odot \bar{\mathbf{A}} \\
\alpha &= \alpha_0 + \left[\ln(1 + \mathbf{X}) - \ln(\mathbf{X})\right]\bar{\mathbf{H}}^T \\
\nu &= \nu_0 + \left[\psi(\bar{\mathbf{A}}\bar{\mathbf{H}} + \bar{\mathbf{B}}\bar{\mathbf{H}}) - \psi(\bar{\mathbf{B}}\bar{\mathbf{H}})\right]\bar{\mathbf{H}}^T \odot \bar{\mathbf{B}} \\
\beta &= \beta_0 + \ln(1 + \mathbf{X})\bar{\mathbf{H}}^T \qquad (2.31) \\
\rho &= \rho_0 + \Big\{\bar{\mathbf{A}}^T\left[\psi(\bar{\mathbf{A}}\bar{\mathbf{H}} + \bar{\mathbf{B}}\bar{\mathbf{H}}) - \psi(\bar{\mathbf{A}}\bar{\mathbf{H}})\right] \\
&\quad + \bar{\mathbf{B}}^T\left[\psi(\bar{\mathbf{A}}\bar{\mathbf{H}} + \bar{\mathbf{B}}\bar{\mathbf{H}}) - \psi(\bar{\mathbf{B}}\bar{\mathbf{H}})\right]\Big\} \odot \bar{H} \\
\zeta &= \zeta_0 + (\bar{\mathbf{A}}^T + \bar{\mathbf{B}}^T)\ln(1 + \mathbf{X}) - \bar{\mathbf{A}}^T \ln \mathbf{X}
\end{aligned}
$$

The values of the parameter matrices are given by iteratively updating the parameters and the expectations of A, B and H accordingly. The expectation of A, B and H are given by [2]:

$$
\bar{\mathbf{A}} = \mu \oslash \alpha, \bar{\mathbf{B}} = \nu \oslash \beta, \bar{\mathbf{H}} = \rho \oslash \zeta \qquad (2.32)
$$

The corresponding algorithm is summarized as follows:

Algorithm 1 IBG-NMF

```
Input: Observation matrix $\mathbf{X}$, number of basis $K$, initialization of $\alpha_0$, $\beta_0$, $\rho_0$, $\mu_0$, $\nu_0$ and $\zeta_0$.
Output: Hyperparameters $\alpha$, $\beta$, $\rho$, $\mu$, $\nu$ and $\zeta$.
  Generate $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$ and $\bar{\mathbf{H}}$ according to (2.32)
  while $iterations < maxIter$ do
    Update parameters $\alpha_0$, $\beta_0$, $\rho_0$, $\mu_0$, $\nu_0$ and $\zeta_0$ according to (2.31)
    Update expectations $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$ and $\bar{\mathbf{H}}$ according to (2.32)
    if convergent precision is satisfied then
      break;
    end if
  end while=0
```

An empirical study of the convergence of algorithm 1 showed that the proposed IBG-NMF

[1] $\odot$ denotes element-wise multiplication.

[2] $\oslash$ denotes element-wise division.

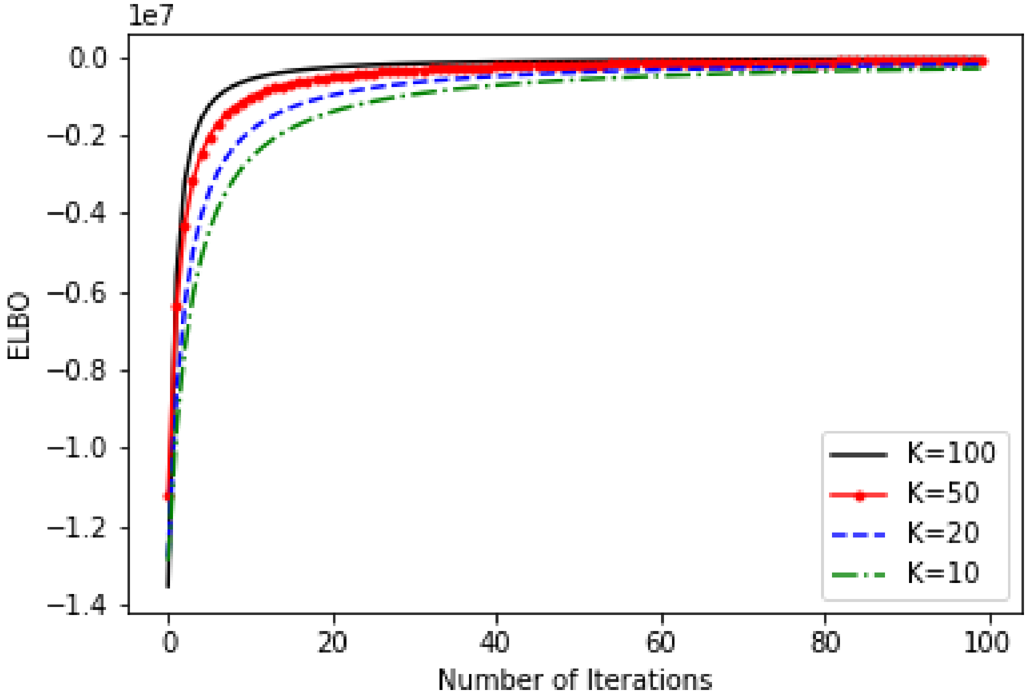


Figure 2.1: Illustration of the convergence of the IBG-NMF algorithm.

algorithm could always converge after about 60 to 80 rounds of iterations. The objective function is numerically calculated by generating samples from the posterior distributions. An example of the convergence rate for Olivetti faces database is presented in figure 2.1. For this example, we downsample observations to size 32×32. We set K to $10, 20, 50$ and 100. The algorithm could always converge after about 60 to 80 iterations. The objective function is numerically calculated by generating samples from the posterior distributions.

2.5 Online Variational Inference

The model proposed above introduces an innovative robust approach for matrix factorization with semi-bounded data. However, most of NMF applications involve large dataset where scalability is an issue. In many cases, data continuously arrive in streams or batches. Online learning is a well known solution for the problems mentioned above. The standard variational Bayes formulation is adapted to the online setting by stochastic coordinate ascent. The advantages of this online learning are: only a limited amount of data needs to be stored at a time in memory, independently of the size of the original dataset; by processing the data in a random sequence, we gain robustness to local optima and maintain convergence guarantees.

The batch variational inference approach for learning IBG-NMF model is extended to online settings by adopting the framework. The observation matrix X can be partitioned into smaller subsets $S^{(t)}$, called mini-batches which change for each iteration t. For a fixed amount of data T at each iteration, the current lower bound for the observed data at iteration t is given by:

$$\mathcal{L}^{(t)}(q) = \frac{NM}{T} \sum_{(n,m)\in S^{(t)}} \Big\{ E_{\,q(\mathbf{Z})}[p(x_{nm}|\mathbf{Z})] - E_{\,q(\mathbf{Z})}[q(\mathbf{Z})] \Big\} \tag{2.33}$$

Assume that we have already observed $(t-1)$ batches $S^{(1)}, ..., S^{(t-1)}$, for a new observation $S^{(t)}$, we can maximize the current lower bound $\mathcal{L}^{(t)}(q)$ with regard to $q^t(\mathbf{Z})$ for each variational factor while keeping other values to their $(t-1)$ values. Therefore, the model hyperparameters $\Theta^{(t)} = \{\alpha^{(t)}, \mu^{(t)}, \beta^{(t)}, \nu^{(t)}, \zeta^{(t)}, \rho^{(t)}\}$ computed at iteration t can be calculated as:

$$\begin{aligned} \alpha^{(t)} &= \alpha^{t-1} + r^{(t)}\Delta\alpha^{(t)}, \mu^{(t)} = \mu^{t-1} + r^{(t)}\Delta\mu^{(t)} \\ \beta^{(t)} &= \beta^{t-1} + r^{(t)}\Delta\beta^{(t)}, \nu^{(t)} = \nu^{t-1} + r^{(t)}\Delta\nu^{(t)} \\ \zeta^{(t)} &= \zeta^{t-1} + r^{(t)}\Delta\zeta^{(t)}, \rho^{(t)} = \rho^{t-1} + r^{(t)}\Delta\rho^{(t)} \end{aligned} \tag{2.34}$$

where r_t is the learning rate which is used to reduce the earlier inaccurate estimation effects that contributed to the lower bound and accelerate the convergence of the learning process. In this work, we adopt a learning rate function such that $r_t = (\eta_0 + t) - a$, subject to the constraints $a \in (0.5, 1]$ and $\eta_0 \geq 0$. In 2.34, $\Delta\Theta^{(t)} = \{\Delta\alpha^{(t)}, \Delta\mu^{(t)}, \Delta\beta^{(t)}, \Delta\nu^{(t)}, \Delta\zeta^{(t)}, \Delta\rho^{(t)}\}$ are the natural gradients of the corresponding hyperparameters. The natural gradient of a parameter is obtained by multiplying the gradient by the inverse of Riemannian metric, which cancels the coefficient matrix for the posterior parameter distribution. It is defined as :

$$
\begin{aligned}
\Delta\mu^{(t)} =& \mu_0 - \mu^{(t-1)} + \frac{MN}{T} \sum_{m \in S(t)} \Big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) - \psi(\sum_k \bar{a}_{nk}\bar{h}_{km})\Big]\bar{a}_{nk}\bar{h}_{km} \\
\Delta\alpha^{(t)} =& \alpha_0 - \alpha^{(t-1)} + \frac{MN}{T} \sum_{m \in S(t)} \Big[\ln(1 + x_{nm}) - \ln x_{nm}\Big]\bar{h}_{km}
\end{aligned}
\tag{2.35}
$$

$$
\begin{aligned}
\Delta\nu^{(t)} =& \nu_0 - \nu^{(t-1)} + \frac{MN}{T} \sum_{m \in S(t)} \Big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) - \psi(\sum_k \bar{b}_{nk}\bar{h}_{km})\Big]\bar{b}_{nk}\bar{h}_{km} \\
\Delta\beta^{(t)} =& \beta_0 - \beta^{(t-1)} + \frac{MN}{T} \sum_{m \in S(t)} \Big[\bar{h}_{km} \ln(1 + x_{nm})\Big]
\end{aligned}
\tag{2.36}
$$

$$
\begin{aligned}
\Delta\rho^{(t)} =& \rho_0 - \rho^{(t-1)} + \frac{MN}{T} \sum_{n \in S(t)} \Big[\big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) - \psi(\sum_k \bar{a}_{nk}\bar{h}_{km})\big]\bar{h}_{km}\bar{a}_{nk} \\
&+ \big[\psi(\sum_k \bar{a}_{nk}\bar{h}_{km} + \bar{b}_{nk}\bar{h}_{km}) - \psi(\sum_k \bar{b}_{nk}\bar{h}_{km})\big]\bar{h}_{km}\bar{b}_{nk}\Big] \\
\Delta\zeta^{(t)} =& \zeta_0 - \zeta^{(t-1)} + \frac{MN}{T} \sum_{n \in S(t)} \Big[(\bar{a}_{nk} + \bar{b}_{nk}) \ln(1 + x_{nm}) - \bar{a}_{nk} \ln(x_{nm})\Big]
\end{aligned}
\tag{2.37}
$$

where $\Theta_i^{(t)}$ corresponds to the optimal values of the hyperparameter Θ_i while optimizing $\mathcal{L}^{(t)}(Q)$ with regard to $Q^{(t)}$. The online nonnegative matrix factorization approach is summerized in algorithm 2.

Algorithm 2 online IBG-NMF

```
Input: Observation matrix X, number of basis K, initialization of $\alpha_0$, $\beta_0$, $\rho_0$, $\mu_0$, $\nu_0$ and $\zeta_0$.
Output: Hyperparameters $\alpha$, $\beta$, $\rho$, $\mu$, $\nu$ and $\zeta$.
  Generate $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$ and $\bar{\mathbf{H}}$ according to (2.32)
  while Receiving new stream do
    Calculate natural gradients $\Delta\alpha^{(t)}$, $\Delta\mu^{(t)}$, $\Delta\beta^{(t)}$, $\Delta\nu^{(t)}$, $\Delta\zeta^{(t)}$, $\Delta\rho^{(t)}$ according to (3.22)
    Update variational parameters $\alpha_0$, $\beta_0$, $\rho_0$, $\mu_0$, $\nu_0$ and $\zeta_0$ according to (2.31)
    Update expectations $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$ and $\bar{\mathbf{H}}$ according to (2.32)
  end while=0
```

2.6 Experimental Results

2.6.1 Experiments setup

In this section, we present the results of the IBG-NMF models on five different applications. First, we use our models for parts based representation with ORl and CBCL datasets. This experiment highlights both models' capabilities for parts based decomposition and how they can be used while imposing sparsity constraint. Secondly, the model is used for collaborative filtering application on two data sets: Netflix data and MovieLens data. Then, the IBG-NMF model is used for topic modeling on Reuters dataset. We also propose the usage of NMF for a retail application. Finally, we use our model for graph embedding on biomedical link prediction.

In each experiment, the two proposed models are compared against baseline results from state of the art factorization models namely NMF, PMF [56], BPMF [69] and BNMF [73] when applicable. Additional baseline models are used for comparison for specific experiments. Particularly, for collaborative filtering, we also run the experiment on BG-NMF model that takes into consideration the fact that the data is bounded. For the topic modeling experiments we add results obtained from LDA [6] and hierarchical Dirichlet process (HDP) [81]. Finally, for the graph embedding we compare our models' performance against the models proposed in [96]. Experiments are run with a 10 folds cross validation. That is, 10 random splits each with 90% training data and 10% test data where the 10 test splits do not overlap. The average performance metric is reported. We test on different values of the latent space dimension $K \in \{10, 20, 50, 100, 200, 500, 800\}$. K values with highest results for each algorithm is reported.

2.6.2 Parts based representation

Extending the non-negativity constraint in matrix factorization models induces sparsity and leads to part-based decomposition. We apply parts based decomposition on ORL dataset [70] and CRCL dataset.

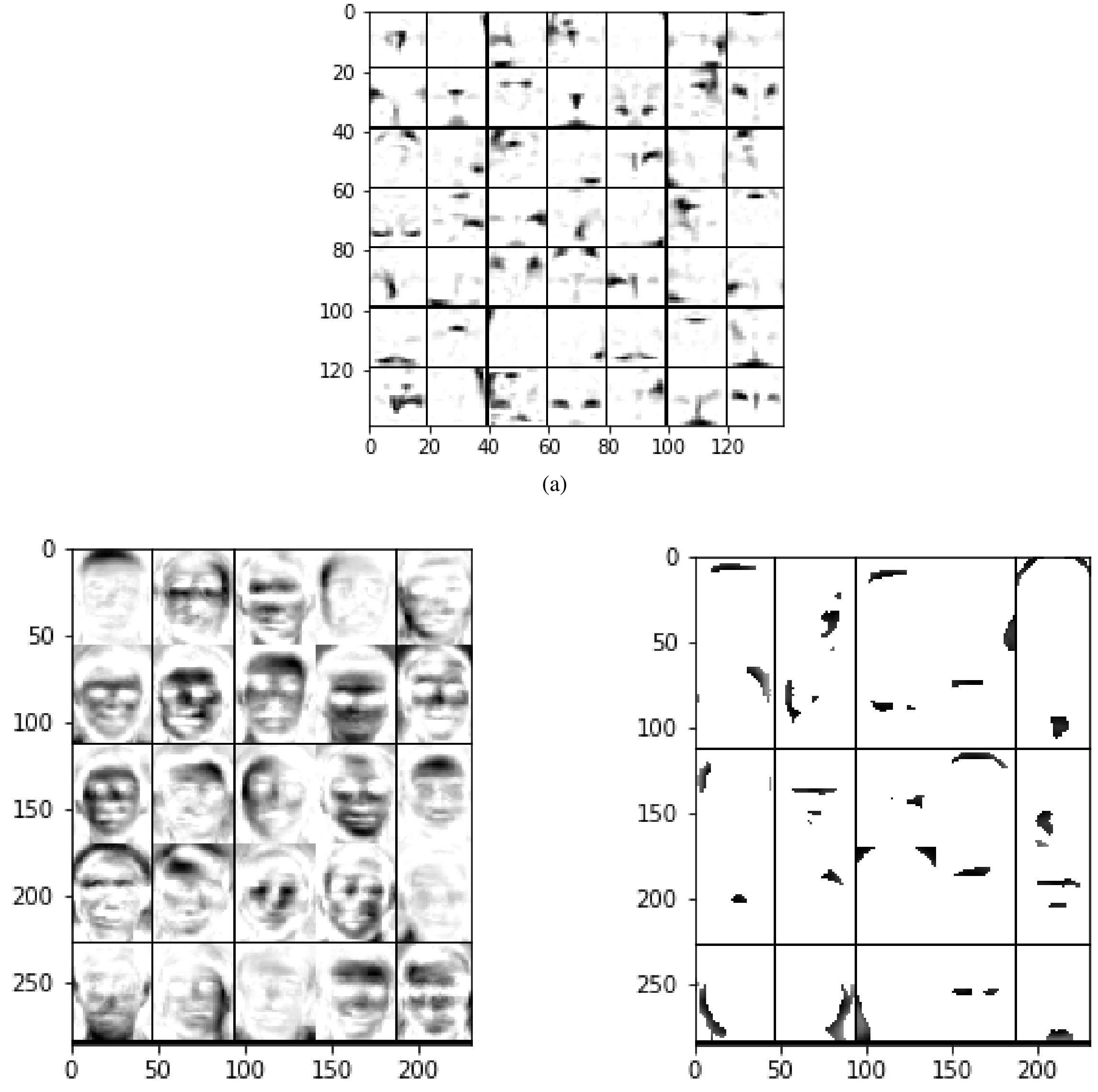
ORL is a dataset of 400 face images of size 112×92. It contains 40 distinct persons' images with 10 examples each. Images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The matrix factorization is performed on reduced face images by constructing a matrix of shape 2576 (pixels) x 400 (faces). The number of basis is 25. The CBCL dataset consists of 2429 face images of size 19x 19. Each face image is preprocessed. For this data, we choose a basis equal to 49.

Results are reported in figures 3.4 and 2.3. With regular shape and scale parameters (no sprseness), a parts-based representation can be found on face images from CBCL database. However, with the same parameters applied on the ORL dataset, in which images are not as well aligned, a global decomposition emerges. Authors in [32, 21] have shown that setting a high sparseness value for the basis images results in a local representation. The sparseness constraint in our model can be imposed by setting a low shape parameter for the matrix **H**. We set the shape parameter to 0.0001. Imposing sparseness constraint on the basis matrix allows obtaining parts based representation. Both IBG-NMF and online IBG-NMF give very close results.

2.6.3 Collaborative Filtering

Recommendation systems play a major role in the internet industry. They help users find new products and are used by companies to find potential buyers/customers. Collaborative filtering is a technique used in recommendation systems based on historical reviews. During the Netflix competition, matrix factorization techniques showed a good performance to predict the reaction of some users based on their historical data. The Netflix dataset along with the Movielens dataset are used here to evaluate our model. We use mean square error (MSE) as evaluation metric.

The Netflix dataset was initially published during the 2008 competition, and has since then attracted several researchers. The original dataset includes $17,770$ movies and $480,189$ users. Users' ratings are on a five star (integral) scale from 1 to 5. A subset of the original dataset with $10,164,642$ ratings of $46,584$ users over $2,249$ movies was used. The subset is selected in a way that we are only keeping the 90% percentile of the users with highest number of ratings. MovieLens 1M dataset [24] contains $1,000,209$ anonymous ratings of approximately $3,706$ movies made by $6,040$ MovieLens users who joined MovieLens in 2000. The main goal here is to reconstruct the original matrix in order to predict non observed ratings. The number of basis K is considered as a hyperparameter. Elements from the test set are excluded from the training phase by setting the mask matrix elements to zero. Results on the test set are reported in table 2.1. We show that our model performs the best

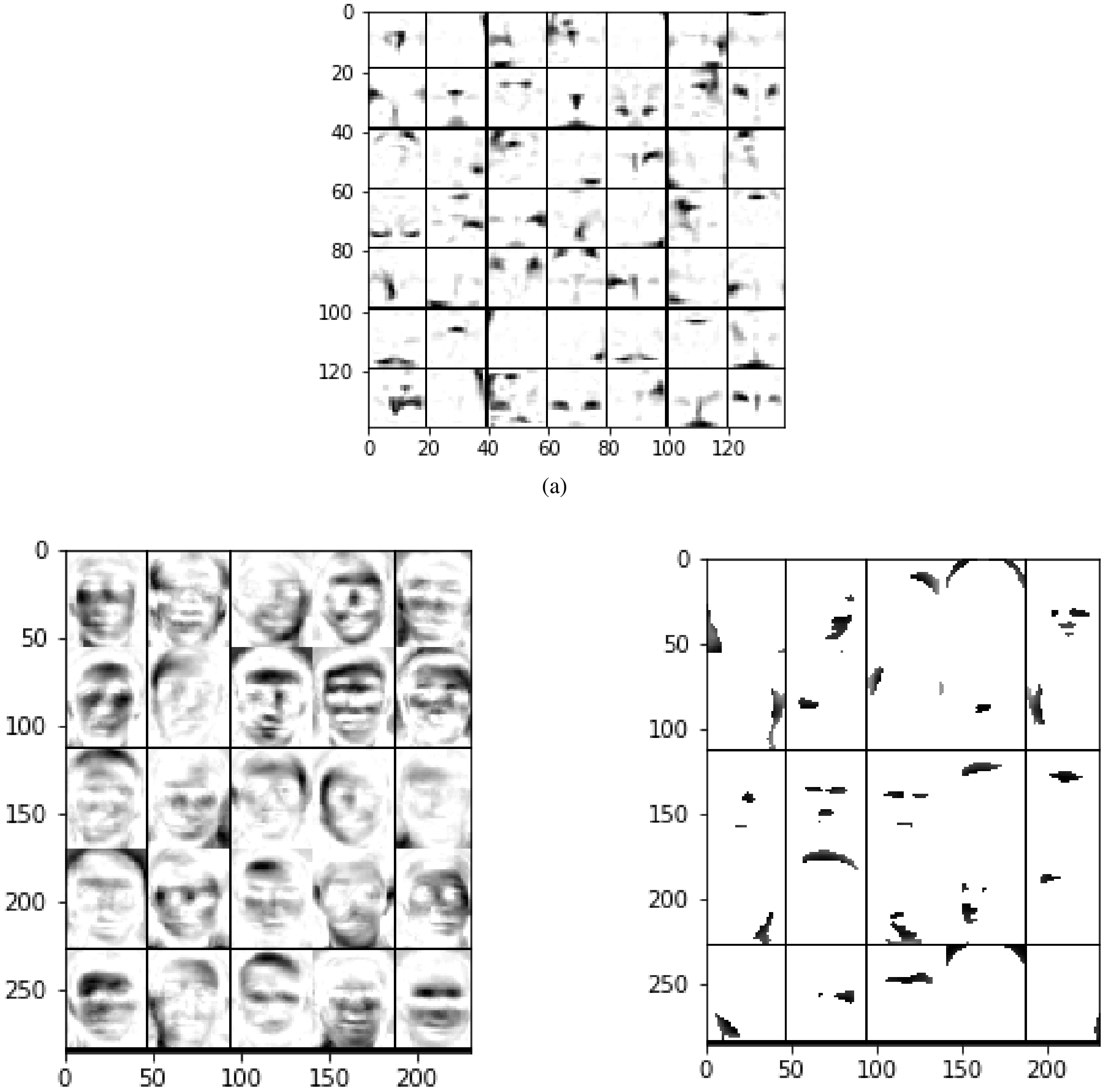


(a)

(b)

(c)

Figure 2.2: Results from IBG-NMF (a) parts-based decomposition on CBCL dataset without constarints. (b) parts based decomposition without constraints on ORL dataset. (c) parts-based decomposition on ORL dataset with sparsity constraint.



(a)

(b)

(c)

Figure 2.3: Results from IBG-NMF (a) parts-based decomposition on CBCL dataset without constarints. (b) parts based decomposition without constraints on ORL dataset. (c) parts-based decomposition on ORL dataset with sparsity constraint.

Table 2.1: MSE Score for Collaborative Filtering

	Netflix Dataset				
Model	50	100	200	500	800
online-IBG-NMF	0.8597	0.8502	0.8208	0.7692	**0.7446**
IBG-NMF	0.8686	0.7971	0.7872	0.7783	**0.7587**
BG-NMF	0.7758	0.8413	0.8985	0.9029	0.8753
NMF	0.954	0.9728	0.9819	0.9889	0.9126
PMF	0.8597	0.8727	0.8382	0.8278	0.8197
BPMF	0.9234	0.8527	0.8507	0.8218	0.8021
BNMF	0.8784	0.8434	0.8252	0.7962	0.7820
	1M Movielens Dataset				
	50	100	200	500	800
online-IBG-NMF	0.9500	0.9207	0.8789	0.8411	**0.8381**
IBG-NMF	0.9606	0.9398	0.8982	0.8524	**0.8430**
BG-NMF	0.8602	0.8964	0.9027	0.8993	0.9012
NMF	0.9110	0.9598	0.9689	0.9769	0.9912
PMF	0.9921	0.9685	0.9617	0.9317	0.9117
BPMF	0.9770	0.9787	0.9602	0.9280	0.9017
BNMF	1.1646	1.0292	0.9475	0.9303	0.9162

with high values of K. The lowest MSE value is achieved by online IBG-NMF then IBG-NMF for both datasets for a value of $K = 800$, Netflix data's score is 0.7446 and Movielens score is 0.8381. Whereas, the benchmark model BG-NMF achives its best score for a value ok $K = 50$. Netflix data's score is 0.7759 and Movielens score is 0.8602. NMF's results are very poor compared to the other algorithms. BNMF has the closest values to our model. However, zhen taking into account the convergene rate and computational cost of both models, variational inference based models are less expensive than MCMC based models which gives further advantages to the proposed models.

2.6.4 Topic Modeling

Topic modeling is used to uncover the latent aspects (topics) from text data. With the increasing amount of text data available on digital format, navigation should be aided by data mining tools that allow to quickly locate data based on some information within it. We use Reuters public dataset to evaluate the performance of the proposed models against two popular approaches that have shown great success on this application: regular NMF, BNMF, HDP and LDA. We select texts from the top 20 predefined topics. After removing stop words and rare words, TF-IDF transformation was

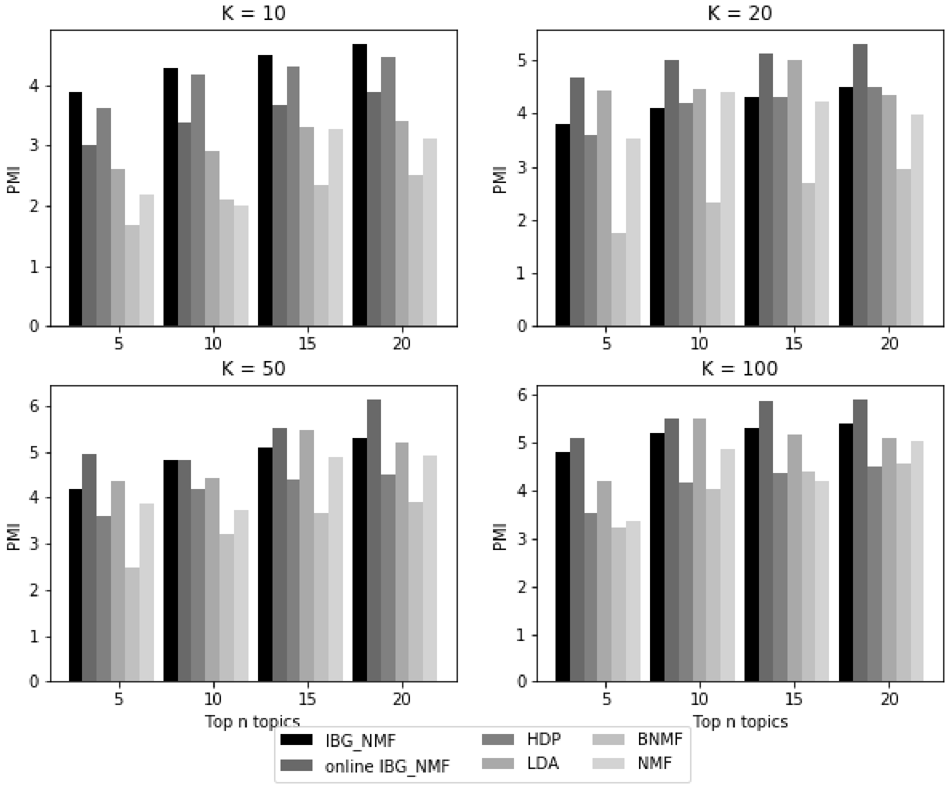


Figure 2.4: PMI score for different values of K.

used to model the text data. The final data is represented in a $(8609, 3000)$ matrix where the rows represent documents and columns represent dictionary words.

For the model evaluation, several automatic metrics have been proposed to evaluate topic quality such as perplexity , topic coherence, and Pointwise Mutual Information (PMI) scores. Kaplan et al. [38] have shown that the coherence based on PMI gave the largest correlation with human ratings. PMI for a given topic t is calculated as:

$$PMI(t|\boldsymbol{v}^{(t)}) = \frac{1}{M(M-2)} \sum_{m=2}^{M} \sum_{l=1}^{m-1} \left(\frac{p(v_m^{(t)}, v_l^{(t)}) + 1}{p(v_m^{(t)})p(v_l^{(t)})}\right) \tag{2.38}$$

where $\boldsymbol{v}^{(t)} = (v_1^{(t)}, v_2^{(t)}, ..., v_M^{(t)})$ is a list of the M most probable words in topic t, and $v_m^{(t)}$ and $v_l^{(t)}$ represent the m^{th} and l^{th} words of the specific topic t, respectively. $p(v_m^{(t)}, v_l^{(t)})$ denotes the probability that words $v_m^{(t)}$ and $v_l^{(t)}$ appear in the same document, while $p(v_m^{(t)})$ or $p(v_l^{(t)})$ mark the probability that the m_{th} or l_{th} term occurs in the document corpus. A smoothing count of 1 is included to avoid taking the logarithm of zero. Generally speaking, the more the co-occurrence under the same topic it owns, the larger value the PMI score will be, and the better performance the mined topic is. With respect to the K topics, the final average PMI score is then computed as the topic model's performance in topic-quality.

Results reported in Figure 2.4 show that IBG-NMF and online IBG-NMF outperform classical methods for different values of K and different numbers of selected words. Increasing the number of words per topic increases the coherence within the topic. In general, online IBG-NMF outperforms full batch IBG-NMF because of the gradient capacity to avoid local minimum and overfitting. Online approach is also adapted to the nature of this application. In fact news data continuously arrives in streams (hourly, daily or weekly).

2.6.5 Transactions Prediction and items classification

Data mining approaches are being widely used in the retail industry for demand forecasting, improving customer experience, and understanding market behaviour. Transaction data have a rich hidden information that can help improve the retail business. Yet, due to its complexity, exploring it might be challenging. These data can be represented in the form of a matrix, where rows represent transactions and columns represent products. The matrix entries take positive values. NMF can be used in this case for two different applications:

Products clustering :

Clusters are given by the excitation matrix **H**. This approach can be seen as a solution for market basket analysis. Rows of **H** can be later used as product features allowing to model positive interaction between different products also known as halo effect. Results can also be used for shelves and aisle optimization inside stores. Items within the same group should be presented in same aisle or section within a store. The weights measure the importance of each product. The model performance is measured by the PMI scores defined in (2.38). This metric is usually used to evaluate topic coherence in topic mining. From statistical perspective, it measures the point-wise mutual information. In other words, it evaluates how likely two items existing in the same column of the excitation matrix are likely to be observed in the same row of the observed matrix. In the retail context, in the case where we are creating baskets, we need to measure how likely it is, for items within the same basket/group/category, to be purchased together. The basis matrix can then be used to model the positive interaction between items in a feature based forecasting model.

Transaction items prediction :

This can be used for recommendation systems in e-commerce platforms. In fact, the model predicts when a customer added certain items to his basket, what other items he is likely to be interested in. Mean Absolute Percentage Error (MAPE) is used to evaluate this model:

$$MAPE = \frac{1}{NM} \sum_{i,j} \frac{|r_{ij} - \hat{r_{ij}}|}{r_{ij}} \tag{2.39}$$

We use a transaction dataset that contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail [9]. Data cleaning and preprocessing included removing returns (negative items) and outliers (remove $99th$ percentile). After data cleaning, the final dataset includes $20,116$ transactions and $3,935$ products. The algorithm is tested against the classical Bayesian NMF [73] refered to as BNMF, regular NMF. Results are reported in tables 2.2 and 2.3. The proposed model outperformed the benchmark models for both applications. For the basket analysis, we see that the highest PMI score is found for IBG-NMF with $K = 10$. For other models the optimal K is higher. This shows the capacity of our model to preserve the hidden concepts and latent variables of the data while reducing it to a smaller space. An example of three baskets with top 3 items is shown in table 2.4. For example, items from the first basket are all related to the christmas theme. The third basket is related to antique decoration. For the transaction prediction, IBG-NMF achieved 0.5821 MAPE with $K = 100$ compared to 0.6725 for the BNMF.

2.6.6 Graph embedding on biomedical networks

Graph embedding learning aims to automatically learn low-dimensional node representations. Authors in [96] have evaluated graph embedding methods on biomedical networks. We extend

Table 2.2: PMI Score for Basket Analysis

Model	PMI			
	10	20	50	100
OIBG_NMF	**5.6728**	5.3210	4.4779	4.6198
IBNMF	**5.6026**	5.2024	4.7768	4.5481
BPMF	3.6074	4.1523	4.2462	4.3128
BNMF	4.2574	4.4671	4.5219	4.6579
NMF	4.1975	4.3712	4.5734	4.7400

Table 2.3: MAPE Score for Transaction Prediction

Model	MAPE			
	10	20	50	100
OIBG_NMF	0.7754	0.6902	0.6102	**0.5821**
IBNMF	0.7833	0.6972	0.6044	**0.5940**
BPMF	4.0304	3.8457	3.9353	4.0311
BNMF	0.7238	0.7104	0.6953	0.6725
NMF	0.8959	0.8734	0.8384	0.7679

Table 2.4: Baskets Examples with top 3 items

Basket 1	Part Cone Chrismas Decoration	T-Light Glass Flued Antique	Spaceboy Gift Wrap
Basket 2	Doormat New England	Chest of Drawers Gingham Heart	Doormat Spotty Home Sweet Home
Basket 3	Antique Heart Shelf Unit	Love Seat Antique White Metal	SET/4 White Retro Storage Cubes

their study to evaluate our model against traditional matrix factorization based methods used for link prediction in three important biomedical link prediction tasks: drug-disease association (DDA) prediction, drug–drug interaction (DDI) prediction and protein–protein interaction (PPI) prediction. Link prediction can be defined as: given a set of biomedical entities and their known interactions, we aim to predict other potential interactions between entities [52]. Matrix factorization techniques resolve this problem by factorizing the link matrix to learn low-dimensional representations in a latent space. A binary classification is then performed to decide if a link exists between the biological elements. Yue et al. [96] provided a python package that includes 4 datasets that we use here for evaluation. First for DDA, two datasets are used: Comparative Toxicogenomics Database (CTD) [15] with $92,813$ edges between $12,765$ nodes ($9,580$ chemicals and $3,185$ diseases), and National Drug File Reference Terminology (NDF-RT) in UMLS [7] with $13,545$ nodes ($12,337$ drugs and $1,208$ diseases) and $56,515$ edges. As for DDI graph, data is collected from DrugBank [91] with $242,027$ DDIs between $2,191$ drugs. For PPI graph, Homo sapiens PPIs are extracted from STRING database [80] with $359,776$ interactions among $15,131$ proteins.

IBG-NMF and online IBG-NMF are trained against top three matrix factorization preforming methods HOPE, GraRep and SVD and we also use it on NMF. We use BioNEV [3] to train HOPE and GraRep. We use accuracy metric for model evaluation. Results are reported in table 2.5. The proposed models provided competitive results. With the four datasets, online IBG-NMF provided provided the most effective results. Beside classical matrix factorization algorithms, online IBG-NMF is also able to beat HOPE and GraRep algorithms which are designed to capture the high-order proximity of graphs.

[3] https://github.com/xiangyue9607/BioNEV

Table 2.5: Accuracy for link prediction on the four compiled biomedical datasets

Model	Link Prediction			
	CTD DDA	NDFRT DDA	DrugBank DDI	STRING PPI
OIBG-NMF	0.890	0.951	0.868	0.798
IBG-NMF	0.871	0.942	0.863	0.787
HOPE	0.885	0.928	0.844	0.763
GraRep	0.889	0.938	0.850	0.780
SVD	0.853	0.700	0.837	0.794
NMF	0.781	0.652	0.801	0.698

Chapter 3

Bayesian Non-negative Matrix Factorization for Non-Intrusive Load Monitoring

3.1 Introduction

Non-Intrusive load monitoring (NILM) is defined as the task of extracting single appliances' power consumption out of the aggregated power data using one single energy meter. The problem was first described by Hart in 1982 [97], while more attention has been brought to these techniques in the last few years with the increase of energy demand and the rapid advancements in data driven technologies. To answer to the high demand, energy management and sustainable buildings have become the primary focus of urban planners. Energy management requires monitoring and controlling of electrical utilities for optimizing energy utilization and thereby reducing consumption. Providing details about appliance level power consumption would help consumers understand their usage, optimize their consumption, and detect malfunctioning or inefficient appliances [14, 41]. This set of information can also be used by energy service providers and planners to allocate resources, identify their customers' needs, etc.

The emersion of load monitoring techniques is mostly driven by the advancements in internet of things (IoT), smart meters, and smart grids. However, cost and implementation constraints would prevent including smart meters for every device to collect device level data. There comes the importance of NILM since the estimation of individual electrical loads is based only on the aggregated signal. If a load curve X monitored at a power meter is the sum of three loads consuming respectively d_1, d_2 and d_3, then the task is to determine the state of d_1, d_2 and d_3 individually with the only knowledge of X.

Various techniques have been proposed to resolve the NILM problem. These techniques can be categorized into supervised vs. unsupervised algorithms [66]. Authors in [2] provided a detailed review of most recent NILM approaches and their challenges. Unlike supervised approaches, unsupervised models don't rely on a prior knowledge of the individual appliances' consumption. The user have to set up a set of rules based on domain knowledge and appliances signature to identify appliances. The major drawback of these techniques, is that they can miss appliances with similar signature. Especially that no robust set of features that can effectively describe the appliances have been identified. Such case is very common among low energy consuming appliances for example iron and hear dryer. Two commonly used unsupervised learning techniques are hidden markov models and mixture models. Different variants of Hidden Markov Models (HMM) have been proposed as an unsupervised learning model. The most common one is factorial HMM (FHMM) that generalizes the HMM state representation by letting the state be represented by a collection of state variables [19]. Mixture models and HMM are usually inefficient when the number of disaggregated appliances increases.

Supervised learning approaches on the other hand can be divided into two categories: pattern recognition approaches and optimization approaches. Pattern recognition approaches include common classification models such as SVM [39], K-nearest neighbour (KNN) [28], tree based approaches [94] and deep learning techniques such as recurrent neural networks (RNN) [40] and convolutional neural networks (CNN) [25, 99]. Shallow machine leaning models are usually sensitive to the preprocessing step. Careful feature extraction and selection models have to be performed in order to obtain high performing results [71]. Most of the current work on energy disaggregation is based on deep learning given the huge success these methods have proved. Various deep learning approaches have been proposed, yet most of the proposed techniques treat the problem as a single task learning. They learn separate networks for each appliance and the signal y_k of appliance k given the aggregate signal instead of simultaneously estimating all individual signals [34]. Despite the high performance of these models, this modeling brings major computational challenges. Besides, deep learning architectures have a high dimensional hyperparamter space. The construction of optimal networks requires an excessive preprocessing.

The second category of supervised learning techniques is the optimization approach. It treats NILM as an optimization problem where we seek to find the optimal set of appliances that compose the aggregate signal. Among the most common optimization approaches is matrix factorization (MF) [62]. Matrix factorization models individual appliances' signatures as a basis and aims to reconstruct the aggregate signal as weighted sum over this basis. For the rest of this work, we will refer to the individual power consumption of appliance k as D_k, the corresponding weights as A_k [27]. Depending on whether the signal D is predefined (fixed) or obtained through optimization, MF could be treated as either in a supervised or unsupervised manner. Figueiredo et al. proposed a non-negative tensor factorization technique including additional information from the appliance dependencies [18] that demonstrated the superiority of the proposed MF model over benchmark methods. Compared to other models discussed above, matrix factorization is less sensitive to feature engineering or appliances' signature.

The major challenge with matrix factorization models is the modeling of a problem specific set of constraints. For the case of energy disaggregation, both total consumption and disaggregated signals represent power consumption which is positive. The weights also have to be positive. This specific type of matrix factorization is called non-negative matrix factorization (NMF) [45]. Sparsity constraint in the activation function is a common approach for signal decomposition [79]. In fact, the sparsity constraint guaranties an overcomplete representation of the data. In other words, there exists more basis functions than the dimensionality of the data in the dictionary (i.e., D). Usually, this constraint is achieved by adding an L_1 norm to the loss function. However, this approach is not suitable for energy disaggregation. Authors in [64] have proposed a novel approach to impose the sparsity of the weights matrix A using a sum to k constraint. Unless regularization parameters are tuned carefully, discriminative models are prone to over-fitting because they find a single point estimate of the parameters. Bayesian inference can usually resolve this issue and provide robust solutions [75]. A fully Bayesian treatment of a probabilistic matrix factorization was presented in [69, 17] where the model was trained using Markov Chain Monte Carlo (MCMC) methods. Variational inference is a scalable alternative to MCMC for Bayesian posterior inference. It has been proposed and tested in various domains [88, 5]. Variational inference was used to infer latent variables for Bayesian NMF in [63]. Unlike conventional Bayesian NMF, work in [29] applied matrix factorization on the model parameters. This allows more flexibility and enables imposing further constraints on the factorization model.

In this chapter, a Bayesian non-negative matrix factorization approach is proposed. We assume a generative model where each matrix element follows an exponential distribution. Exponential distribution (exp) with support $(0, \infty)$ can be used to model non-negative real variables. The matrix is modeled in a way to impose a sparsity constraint on the excitation matrix A which is guaranteed through the sum to k. We refer to our model as Bayesian sum to k non-nengative matrix factorization (BS2k-NMF). To model this constraint on a bayesian space, we assume that the coefficients of A_K, weights for device k, follow a Dirichlet distribution of parameters that are subject to our optimization problem. The matrix factorization is applied on the model parameters instead of directly applying it on the observed matrix [12, 29]. A Dirichlet prior is associated to the matrix A. We propose a novel approximation method using mean field variational inference to learn the model and estimate the parameters. The proposed model is evaluated with different applications and tested against different baselines with multiple datasets: REDD dataset, AMPds and IRISE dataset. The proposed model shows high performance against various supervised learning approaches. It performs well for low frequency setup. Moreover, the learning process is low dependent on observations from other houses and therefore can be easily adapted in a federated learning framework.

The rest of this chapter is organized as follows: the generative model and model specifications are introduced in section 2. Experiments, results and comparisons are presented in section 3. Finally, discussion and conclusion are drawn in Section 4.

3.2 Proposed Model

3.2.1 Problem Statement

Given a matrix X, the conventional matrix factorization problem seeks to approximate a matrix X with a product of two matrices A and D s.t.:

$$X \approx DA \tag{3.1}$$

D and A are called basis and activation matrices respectively. Further constraints can be applied on the problem above depending on the nature of the data and the properties of the application. For instance, data generated by real life applications, such as sales data, collaborative filtering data, representations of texts and image data, etc, are semi bounded by nature. Therefore, for the sake of interpretability of the results, optimal processing of this kind of data may call for processing under nonnegativity constraints. This constraint is interpreted as, given a nonnegative matrix X, the values of D and A have to be non negative as well. This model is known as nonnegative matrix factorization (NMF) and it has been widely used for different applications. Notably, NMF has shown great succes in source separation problems where several signals have been mixed together into a mixture signal and the objective is to recover the original component signals from the mixture signal. The nonnegativity constraints allows in this example to obtain comprehensible signal.

Load disaggregation is a special case of source separation problems. The input is the mixed energy consumption for all devices for time window D. So, for an M dimensional signal and for a time window D, we obtain an observation matrix $X \in R_{MD}$. Suppose that the consumption is coming from K devices. A columns of the basis matrix D correspond to the energy consumption profile of an individual component at a day d, and rows of activation matrix A represent the corresponding base for D. We consider two constraints for this problem: nonnegativity constraint to take into account the nonnegativity nature of the data, and the sum-to-k constraint for activation coefficients that imposes the “grouping” effect where the basis vectors from the same individual component form a “group”. The group D_k for device k represents the collected signals of device k over a T_k time windows. The coefficients of elements of D_k for a given time window d are

$A_{kd} = A_{1_k d}, ..., A_{T_k d}$. To preserve the grouping constraint during the factorization process we have $\sum_{t_k=1}^{T_k} A_{t_k d} = 1$. The aggregated signal X is computed as: $X = \sum_k \sum_{t_k=1}^{T_k} D_{mt_k} A_{t_k d}$.

3.2.2 Generative Model

To model the first constraint, we assume that each positive observation X_{md} is generated from an exponential distribution of mean $\frac{1}{\lambda} = DA = \sum_k \sum_{t_k=1}^{T_k} D_{mt_k} A_{t_k d}$.

Given the second constraint, the elements $A_{t_k d}$ of the activation matrix for each device k (i.e., A_k) are the weights of that device being represented via some bases of the signature matrix (i.e., D_k). Therefore, we enforce the summation of all the weights for each device to be equal to one, so that we can be confident that each device's signal is represented by a linear combination of the bases corresponding to that specific device. Correspondingly, summation of the elements of each column of matrix A is equal to k (e.g., number of devices at home). This can be modeled by assuming that the vector $A_{kd} = (a_{1_k d}, ...a_{n_k d},)$ corresponding to the activation of device k follows a Dirichlet distribution.

Therefore the matrix X is drawn according to the following generative model:

$$\begin{aligned} A_{kd} &\sim Dir(A_{kd}|\alpha_{kd}) \\ X_{md} &\sim Exp(X_{md}|(DA)_{md}) \end{aligned} \tag{3.2}$$

where:

$A_{kd} \in R_{T_k \times 1}$, $X \in R_{M \times D}$, $\sum_{t_k=1}^{T_k} A_{t_k d} = 1$, $A_{t_k d} > 0$ and $\alpha_{t_k d} > 0$

$Dir(A_{kd}|\alpha_{kd})$ is the Dirichlet density with parameter vector α_{kd} defined as:

$$Dir(A_{kd}|\alpha_{kd}) = \frac{\Gamma(\sum_{t_k=1}^{T_k} \alpha_{t_k d})}{\prod_{t_k=1}^{T_k} \Gamma(\alpha_{t_k d})} \prod_{t_k=1}^{T_k} A_{t_k d}^{\alpha_{t_k d} - 1} \tag{3.3}$$

and $Exp(X_{md}|(DA)_{mk})$ is the Exponential density with parameter vector $(DA)_{mk}$ defined as:

$$Exp(X_{md}|(DA)_{md}) = \frac{1}{\sum_k \sum_{t_k=1}^{T_k} D_{mt_k} A_{t_k d}} e^{\overline{\frac{-X_{md}}{\sum_k \sum_{t_k=1}^{T_k} D_{mt_k} A_{t_k d}}}} \tag{3.4}$$

3.2.3 Variational inference

Given the prior distribution, the inference to the posterior distribution is the central computational problem for analyzing data in the Bayesian analysis, which is also important in our BS2K-NMF model. Given an observed matrix X and a training data D, we want to compute the posterior distribution $p(A|X, D)$. Exact solution is intractable. We appeal to variational inference. The idea of variational inference [37, 4, 35], is to approximate the true posterior $p(A|X, D)$ by $q(A)$. In conjugate models this permits easy coordinate ascent updates using variational distributions of the same families as the prior distributions. However, for some specific applications, variational inference could be applied to non-conjugate models. The model proposed in this chapter is such a model. We give the activation vectors a variational distribution from the same family as its prior distribution:

$$q(A_{kd}) = \frac{\Gamma(\sum_{t_k=1}^{T_k} \alpha_{t_k d})}{\prod_{t_k=1}^{T_k} \Gamma(\alpha_{t_k d})} \prod_{t_k=1}^{T_k} A_{t_k d}^{\alpha_{t_k d}-1} \tag{3.5}$$

In the following parts we estimate the parameters that minimize the divergence between $q(A)$ and $p(A|X)$ which is measured by the KL divergence $KL(q(A)||p(A|X))$:

$$\begin{aligned} KL\big(q(A)||p(A|X)\big) &= E_q\big[\ln \frac{q(A)}{p(A|X)}\big] \\ &= \ln p(X) - E_q\big[\ln \frac{p(A,X)}{q(A)}\big] \end{aligned} \tag{3.6}$$

From (A.1), minimizing the KL divergence is equivalent to maximizing an Evidence Lower Bound (ELBO) that we denote as $\mathcal{L}$ and is equal to $E_q\big[\ln \frac{p(A,D,X)}{q(A)}\big]$.

$$\begin{aligned} \mathcal{L} &= E_q\big[\ln \frac{p(A,D,X)}{q(A)}\big] \\ &= E_q\big[\ln p(X|A,D)\big] + E_q\big[\ln p(A|\alpha)\big] - E_q\big[\ln q(A)\big] \end{aligned} \tag{3.7}$$

The likelihood in (3.7) is equal to:

$$E_q\big[\ln p(X|A,D)\big] = \sum_{m,d} E_q\left[\frac{-X_{md}}{\sum_k \sum_{t_k=1}^{T_k} D_{mt_k}A_{t_kd}}\right] - E_q\left[\ln \sum_k \sum_{t_k=1}^{T_k} D_{mt_k}A_{t_kd}\right] \tag{3.8}$$

The values of the expectations above are intractable. Similar to [29], they can be lower bounded by applying Jensen's inequalities and Taylor expansion. First, the function $-x^{-1}$ is concave. According to Jensen's inequality, for any vector ϕ such that $\phi_l \geq 0$ and $\sum_l \phi_l = 1$ we have:

$$-\frac{1}{\sum_l x_l} = -\frac{1}{\sum_l \phi_l \frac{x_l}{\phi_l}} \geq -\sum_l \phi_l \frac{1}{\frac{x_l}{\phi_l}} = -\sum_l \phi_l^2 \frac{1}{x_l} \tag{3.9}$$

Therefore, a lower bound of the first expectation in (3.8) is:

$$E_q\left[\frac{-X_{md}}{\sum_k \sum_{t_k=1}^{T_k} D_{mt_k}A_{t_kd}}\right] \geq \sum_k \sum_{t_k=1}^{T_k} \phi_{t_kmd}^2 E_q\left[\frac{-X_{md}}{D_{mt_k}A_{t_kd}}\right] \tag{3.10}$$

Given the convexity of $-\ln x$, we can bound the second expectation in (3.8) using a first order Taylor approximation about an arbitrary point ω_{md}:

$$-\ E_q\left[\ln \sum_k \sum_{t_k=1}^{T_k} D_{mt_k}A_{t_kd}\right] \ \geq\ -\ln(\omega_{md})\ +\ 1\ -\ \frac{1}{\omega}\sum_k \sum_{t_k=1}^{T_k} E_q\left[D_{mt_k}A_{t_kd}\right] \tag{3.11}$$

Finally the likelihood can be bounded as following:

$$E_q\big[\ln p(X|A,D)\big] \geq \sum_k \sum_{t_k=1}^{T_k} \phi_{t_kmd}^2 E_q\left[\frac{-X_{md}}{D_{mt_k}A_{t_kd}}\right] - \ln(\omega_{md}) + 1 - \frac{1}{\omega}\sum_k \sum_{t_k=1}^{T_k} E_q\left[D_{mt_k}A_{t_kd}\right] \tag{3.12}$$

In equations 3.10 and 3.11, we derived bounds on the intractable expectations in (3.8). After updating the variational distributions on each set of parameters A. We update ϕ and ω to re-tighten these bounds. Using Lagrange multipliers, we find that the optimal ϕ is:

$$\phi_{t_kmd} \propto E_q\left[\frac{1}{D_{mt_k}A_{t_kd}}\right]^{-1} \tag{3.13}$$

The bound in (3.11) is tightest when

$$\omega_{md} = \sum_k \sum_{t_k=1}^{T_k} E_q\left[D_{mt_k}A_{t_kd}\right] \tag{3.14}$$

Optimizing the variational distribution

The objective function $\mathcal{L}$ can be written as:

$$\mathcal{L} = \underbrace{E_q\big[\ln p(X|A,D)\big]}_{E1} + \underbrace{E_q\big[\ln p(A|\alpha)\big]}_{E2} - \underbrace{E_q\big[\ln q(A)\big]}_{E3} \tag{3.15}$$

$$E1 = \sum_k \sum_{t_k=1}^{T_k} \phi^2_{t_kmd} E_q\left[\frac{-X_{md}}{D_{mt_k}A_{t_kd}}\right] - \ln(\omega_{md}) + 1 - \frac{1}{\omega}\sum_k \sum_{t_k=1}^{T_k} E_q\left[D_{mt_k}A_{t_kd}\right] \tag{3.16}$$

$$E2 = \sum_k \sum_{t_k}^{T_k}(a_{t_kd} - 1)E_q\left[\ln A_{t_kd}\right] + cst \tag{3.17}$$

$$E3 = \sum_k \Big[\ln\Gamma(\sum_{t_k}^{T_k}\alpha_{t_kd}) - \sum_{t_k}^{T_k}\ln\Gamma(\alpha_{t_kd}) + \sum_{t_k}^{T_k}(\alpha_{t_kd} - 1)E_q\big[\ln A_{t_kd}\big]\Big] \tag{3.18}$$

We have:

$$E_q\big[A_{t_kd}\big] = \frac{\alpha_{t_kd}}{\sum_{p_k=1}^{T_k}\alpha_{p_kd}} \tag{3.19}$$

$$E_q\big[\frac{1}{A_{t_kd}}\big] = \frac{\sum_{p_k=1}^{T_k}\alpha_{p_kd} - 1}{\alpha_{t_kd} - 1} \tag{3.20}$$

$$E_q\big[\ln A_{t_kd}\big] = \Psi(\alpha_{t_kd}) - \Psi(\sum_{p_k=1}^{T_k}\alpha_{p_kd}) \tag{3.21}$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha_{t_k d}} &= \sum_m X_{md} \Big[\sum_{p_k} \big(\frac{\phi^2_{t_k md} \alpha_{p_k d}}{D_{mt_k}(\alpha_{t_k d} - 1)^2} - \frac{\phi^2_{p_k md}}{D_{mp_k}(\alpha_{t_k d} - 1)} \big) \Big] \\
&+ (a_{t_k d} - \alpha_{t_k d}) \Psi'(\alpha_{t_k d}) - \sum_{p_k = 1}^{T_k} (a_{p_k d} - \alpha_{p_k d}) \Psi(\alpha_{p_k d}) \\
&- \frac{1}{(a_{t_k d} - 1)^2}
\end{aligned}
\tag{3.22}
$$

Algorithm

The model parameters are updated using gradient descent. In this work, we adopt a learning rate function such that $\rho_t = (\eta_0 + t) - a$, subject to the constraints $a \in (0.5, 1]$ and $\eta_0 \geq 0$.

Algorithm 3 BS2K-NMF

```
Input: Observation Matrix $X$, Basis matrix $D$, number of basis $K$, initialization of $\alpha_0$.
Output: Hyperparameters $\alpha$.
  while $iteration < maxiter$ do
    Update learning rate $\rho^{(t)}$ according to $\rho_t = (\eta_0 + t) - a$
    Update variational parameters $\alpha_0$:
    $\alpha_{t_k d}(t) = \alpha_{t_k d}^{(t-1)} - \rho^{(t)} \times \frac{\partial \mathcal{L}}{\partial \alpha_{t_k d}}$
    Update expectations $\bar{A}$
  end while=0
```

3.3 Experiments and Results

We are evaluating the performance of the proposed model against state of the art models using three publicly available datasets: REDD dataset [42], AMPds dataset [54] and IRISE dataset [16]. The proposed model is used to isolate the contribution of each appliance to the total energy consumption. Each time, our model is compared against state of the art supervised algorithms that are commonly used for NILM. We will be using another group-based decomposition approach: Elastic-Net [103] and a time series based approaches: recurrent neural network (RNN). Detailed analysis at appliance level are also provided.

Both time duration (time period of the collected signal) and time intervals of data is important because we want to capture differences in behavior for several devices in different seasons and

Table 3.1: NILM Datasets Used for Model Evaluation.

Dataset	Sampling rate	Duration	Type	Location
Ampds	1 min	2 years	Residential	Canada
REDD	165 KHz	19 days	Residential	US
IRISE	10 mintes	1 year	Residential	France

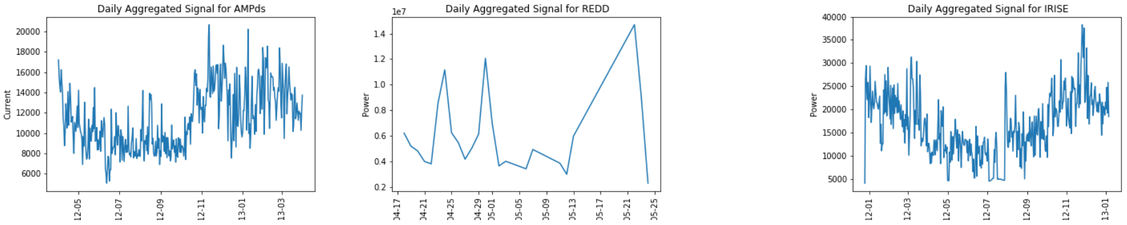
include them in our model at the training stage.

3.3.1 Datasets

To test our models, we are using three different datasets with different characteristics: frequency, sample duration and location. Studies on NILM approaches showed high sensitivity to data frequency and most of the state of the art models showed higher performance with high frequency data. The location of building in question could play a major role on the nature of the data. In fact, the country and city of the considered building could impact the seasonality of the power consumption and the behaviour of the consumer. Table 3.1 describes the three datasets we are using.

AMPd dataset is a public dataset for load disaggregation and eco-feedback research. It includes data between April 1, 2012 and March 31, 2013, including different types of power signal, current and voltage that includes 11 measurements at a sampling rate of one sample per minute for 21 sub-meters. The REDD dataset consists of aggregate and circuit-level power profiles of six US households. The sampling frequency is 3 seconds, which is higher than usual for conventional smart meters in residential applications. The third dataset, IRISE is collected under a European project called Residential Monitoring to Decrease Energy Use and Carbon Emissions in Europe (REMODECE). The data contains over a year of residential consumption of houses located in France. It has recordings of aggregated power for almost all electric appliances in the house at a sampling time of 10 minutes over a year.

We perform an analysis of the consumption patterns across datasets, houses and devices which will be useful for results discussion later. Examples of aggregated daily signal of each of the datasets is presented in figure 3.1. We notice a change of the overall power consumption through different seasons. In fact, for both AMPds and REDD datasets, the power consumption in the months of Mai to September are lower than the consumption from October to April. This supports our assumption



(a) a (b) b (c) c

Figure 3.1: Aggregated daily signal for each of the datasets (a) AMPds dataset (b) REDD dataset (c) IRISE dataset.

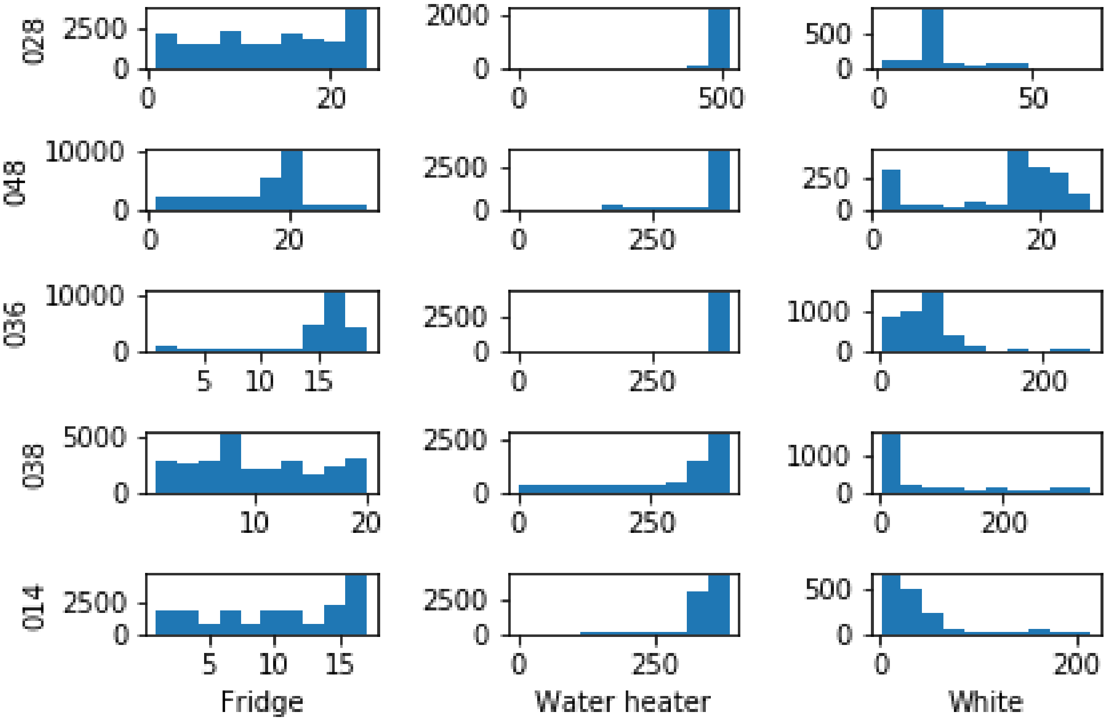


Figure 3.2: Density Charts of Power Consumption in Watt-Hour of Different Devices by House, IRISE dataset

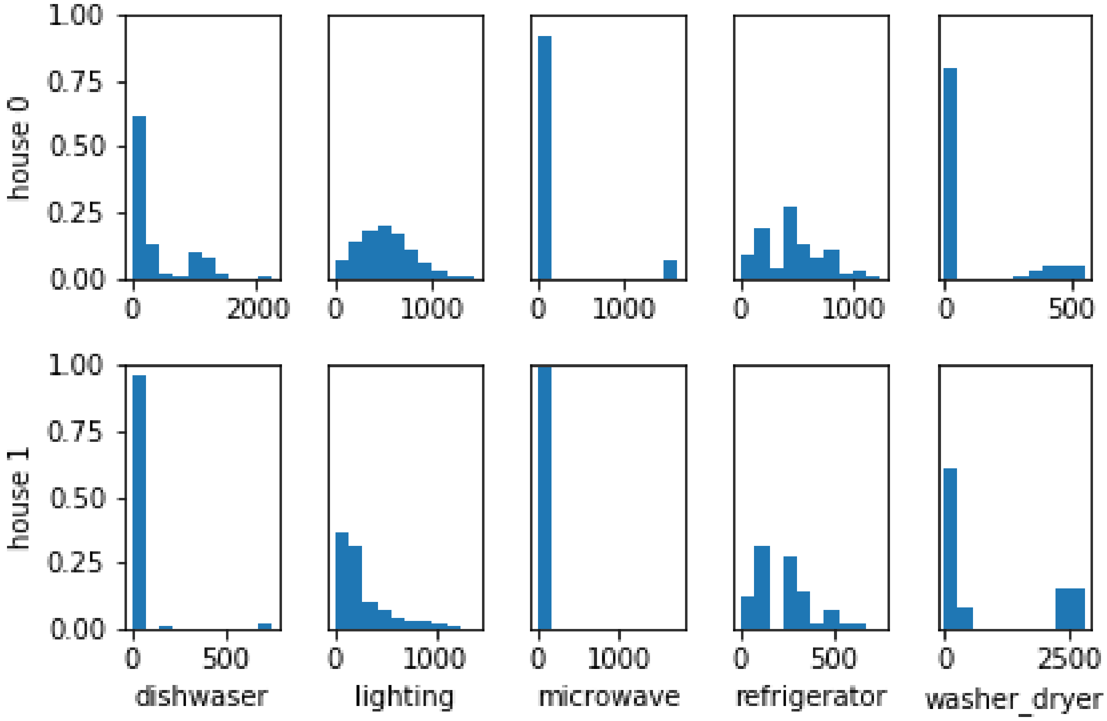


Figure 3.3: Density Charts of Power Consumption in Watt-Hour of Different Devices by House, REDD dataset

on the data consumption variation across seasons and therefore locations. In fact, some appliances' consumption such as heating and air conditioning are highly seasonal. This characteristic should be taken into account when choosing the profiles on which we perform the training, as well as the seen houses on which the model will be trained to later estimate the unseen ones. According to figures 3.2 and 3.3, for the same device types, data consumption levels might vary from house to another, we see that for the IRISE dataset, for white appliances, power consumption is around 20 Watt-Hour for houses 28 and 48, while it goes up to 200 Watt-Hour for houses 36, 38 and 14. Therefore, data normalization should be considered. Additionally, distributions shapes also change across houses and devices.

3.3.2 Evaluation Metrics

Measuring the performance of the NILM is an open research area. Different approaches were discussed in [58]. Evaluation metrics can be categorized based on the targeted performance. Overall accuracy metrics are used to compare between the observed aggregate power signal and the reconstructed signal after disaggregation. These include, but not restricted to, root mean square error (RMSE), disaggregation percentage (DE) and accuracy (Acc). Appliance-based metrics provide a detailed description of how effectively the disaggregated signal signatures are assigned to appliance signatures for example: percentage of contribution in energy consumption (PCEC) and accuracy. However, accuracy metric lacks the capacity to generalize to different appliances. In fact, for appliances that are off most of the time, a model that predicts zero values for all the duration, would have a high accuracy. However, this model would not be capturing the working hours of the appliance. F1-score is a better choice in this case. In this work, we are exploring different metrics to provide a robust analysis of our model. F1-score is reported for both overall and device level assessments. Overall RMSE is reported to compare the proposed model against state of the art models. Device level PCEC values are reported graphically through pie charts.

3.3.3 Results and Discussion

We run 3 different sets of experiments to evaluate the model.

Table 3.2: Overall Performance Comparison with Different NILM approaches, all devices included

		F1-Score	RMSE	DE
AMPds	Elastic Net	0.8613	51.62	12.63
	RNN	0.9215	49.85	10.98
	S2KNMF	**0.9425**	**49.22**	**10.82**
REDD	Elastic Net	0.8213	162.16	81.40
	RNN	0.9415	**160.98**	83.16
	S2KNMF	**0.9514**	161.15	**80.54**
IRISE	Elastic Net	0.8716	9.30	15.16
	RNN	0.9213	**7.44**	**12.81**
	S2KNMF	**0.9279**	8.15	13.73

Table 3.3: Device Level Performance of the AMPds dataset: Test set - All year

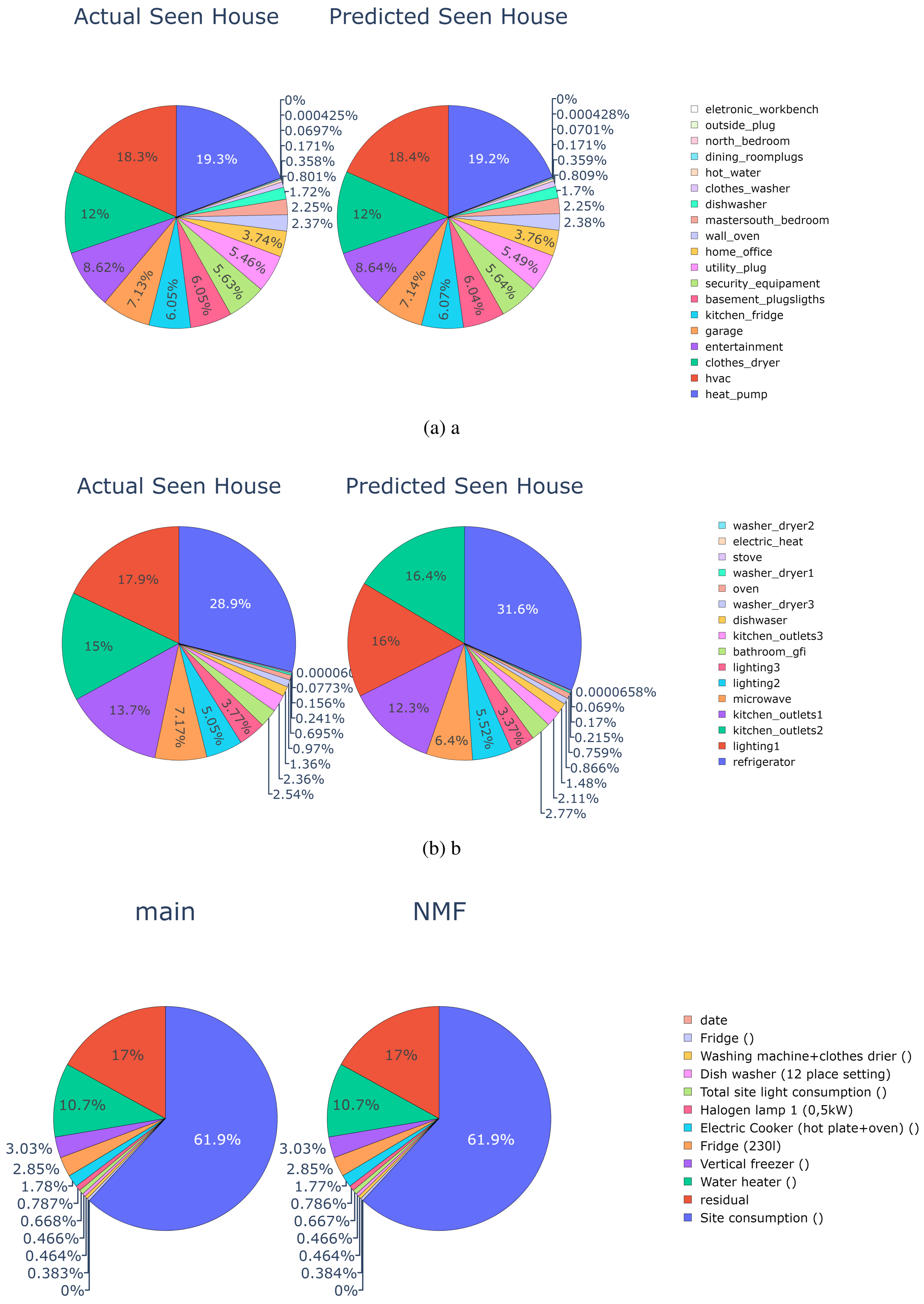
	Precision	Recall	F1-score	Accuracy
Washer	97.70	98.84	98.27	98.55
Dish washer	97.68	98.83	98.25	98.51
Hot water	90.50	95.13	92.76	95.04
Fridge	64.40	80.25	71.46	79.91
HVAC	58.25	38.27	27.18	36.34

Experiment 1 - All appliances within the same house

For each of the datasets, we consider houses individually by training and testing the model using all appliances within the same house. Each house's data is split into train and test sets. For robustness, a 5-fold cross validation is run. Results reported in table 3.2 describe the average of the overall performance of the model across different houses of each dataset. Our model outperformes benchmark techniques when it comes to detecting the states of the appliances reflected by F1-score. Additionally, examples of decomposition are reported in figure 3.4. We see that for decomposition, our model has a superior performance compared to ElasticNet. On average, RNN has better RMSE values. However, we should note that for each appliance, and each house, we had to run a separate network.

Experiment 2 - Selected appliances and unseen houses

For each of the datasets, we select a subset of appliances that are common between houses. We then seperate the buildings into seen and unseen houses. Each time the model is trained with a seen house and tested for the unseen ones (this is applicable only for REDD and IRISE datasets since



(a) a

(b) b

(c) c

Figure 3.4: Pie plots for the energy usage contribution of each device. Ground truth Vs S2K-BNMF (a) AMPds dataset (b) REDD dataset (c) IRISE dataset.

Table 3.4: Performance of the AMPds dataset: Test set, Summer

	Precision	Recall	F1-score	Accuracy
Washer	97.70	98.84	98.27	98.24
Dish washer	97.68	98.83	98.25	98.79
Hot water	90.50	95.13	92.76	95.88
Fridge	64.40	80.25	71.46	76.06
HVAC	66.26	46.03	49.11	48.5

Table 3.5: Performance of the REDD dataset: House 1 seen and House 2 unseen.

	Precision	Recall	F1-Score	Accuracy
Dishwasher	99.04	75.15	85.39	75.15
Lighting	93.26	96.40	96.67	96.40
Microwave	99.33	90.71	94.79	90.71
Washer-Dryer	98.07	91.49	94.63	91.49

AMPds has only one house). For this set of experiments, we consider the overall consumption as the sum of the three components.

For the REDD dataset, we select house1 as seen house and house 2 as unseen house. Four major appliances that are common within residential buildings are considered: dishwasher, lighting, microwave and the washer/dryer (we will refer to them as white appliances). Results are presented in table 3.5.

For the IRISE dataset, we train the model separately with three selected houses: house 28, house 48 and house 36. It is then tested with a test subset from these houses and two unseen houses: house 38 and house 14. For this dataset, we select 3 major appliances that are common within residential buildings: the fridge, the water heater and white appliances. Originally, the metrics are recorded every 10 minutes. However, given the fact that most modern unexpensive smart meters measure values at a very low frequency, we also consider resampled data with 30 minutes interval. We evaluate the model's capacity to capture the on and off status for each appliance using accuracy, F1-score, recall and precision, and the model's capacity to recover the original signal with RMSE score. Overall performance comparison is reported in table 3.6. Appliance level performance metrics are represented in table 3.7. Performance of the proposed model is very close to the state of the art models for the 10 minutes sampling. However, BS2KNMF has higher capacity to generalize to unseen houses and to handle lower frequencies. Unlike other NILM approaches, the performance of our model is not very sensitive to the frequency, it even improves for certain cases. In fact, RNN based models take into account the relationship with previous states. Therefore, when increasing the time interval between two consecutive states, the dependency is reduced. Performance for house 38 is lower compared to other buildings. This can be explained by the difference of the consumption profiles in house 38 as described in figure 3.2.

At appliance level, water heater consumption at house 14 is best captured with house 28 despite the apparent differences in the distributions presented in figure 3.2. The model was able to select the adequate data.

Table 3.6: Overall Performance Comparison with Different NILM approaches, IRISE Dataset

			F1-score	RMSE	F1-score	RMSE	F1-score	RMSE
			House 28		House 14		House 38	
House 28	10 minutes	ElasticNet	73.76	0.2270	72.73	0.2770	61.85	0.3121
		RNN	85.48	0.2406	73.36	0.4586	61.81	0.3794
		S2kNMF	84.91	0.0748	82.15	0.2668	62.52	0.2745
	30 minutes	ElasticNet	84.65	0.3653	68.24	0.2923	58.96	0.2925
		RNN	87.76	0.3160	80.59	0.733	51.57	0.18
		S2kNMF	94.23	0.2500	85.92	0.1289	54.70	0.1289
			House 36		House 14		House 38	
House 36	10 minutes	ElasticNet	80.44	0.2193	64.32	0.3165	67.48	0.3135
		RNN	92.60	0.2671	65.86	0.4423	69.67	0.2958
		S2kNMF	94.57	0.0764	66.27	0.2668	72.04	0.2745
	30 minutes	ElasticNet	78.25	0.3508	66.79	0.2804	56.13	0.2903
		RNN	91.50	0.3684	67.93	0.1412	64.07	0.1329
		S2kNMF	93.52	0.4190	69.49	0.1324	68.11	0.1324
			House 48		House 14		House 38	
House 48	10 minutes	ElasticNet	47.73	0.2343	71.64	0.2536	58.73	0.3033
		RNN	72.77	0.0769	82.54	0.1769	44.01	0.3814
		S2kNMF	71.25	0.1462	79.17	0.2668	78.17	0.2745
	30 minutes	ElasticNet	52.70	0.3500	69.05	0.3029	56.56	0.3029
		RNN	79.51	0.3468	83.25	0.1442	46.49	0.1397
		S2kNMF	80.81	0.3351	81.43	0.1312	55.25	0.1360

Table 3.7: Performance Metrics for Different Devices Across different Houses with 10 Minutes and 30 Minutes Frequencies, IRISE dataset

			Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
			House 38				House 14			
House 28	10 minutes	Fridge	72.23	84.98	78.09	84.98	74.07	86.06	79.62	86.06
		Water Heater	85.54	89.77	87.60	89.77	1	1	1	1
		White Appliances	92.11	95.97	94.00	95.97	93.89	96.90	95.37	96.90
	30 minutes	Fridge	69.10	83.13	75.74	83.13	86.03	74.85	64.08	74.85
		Water Heater	92.22	92.14	92.18	92.14	74.56	83.96	78.95	83.96
		White Appliances	91.01	95.43	93.20	95.43	93.44	96.66	95.02	96.66
			House 38				House 14			
House 48	10 minutes	Fridge	85.17	92.29	88.59	92.29	99.99	99.99	99.99	99.99
		Water Heater	93.61	96.75	95.15	96.75	91.25	95.52	93.33	95.52
		White Appliances	92.11	95.97	94.40	95.97	93.89	96.90	95.37	96.90
	30 minutes	Fridge	94.74	97.04	95.83	97.04	79.92	88.16	82.87	88.16
		Water Heater	80.82	84.97	81.23	84.97	83.95	86.60	85.21	86.60
		White Appliances	92.67	96.27	94.44	96.27	90.44	95.10	92.71	95.10
			House 38				House 14			
House 36	10 minutes	Fridge	77.46	88.01	82.40	88.01	83.89	91.59	87.57	91.59
		Water Heater	94.95	97.44	96.18	97.44	94.71	97.32	95.99	97.32
		White Appliances	94.26	97.09	95.65	97.09	97.20	98.59	97.89	98.59
	30 minutes	Fridge	87.03	93.29	90.05	93.29	98.88	99.43	99.16	99.44
		Water Heater	98.92	99.45	99.18	99.45	93.01	96.44	94.69	96.44
		White Appliances	92.20	96.00	94.06	96.03	96.89	98.39	97.63	98.39

Table 3.8: Appliance Level Performance Including Residual Signal from Unknown Appliances

	House 48	House 14	House 38
Dish Washer	99.17	98.73	98.86
Electrical Cooker	96.17	97.13	92.62
Fridge	86.18	84.87	78.35
Light	89.75	90.13	81.85
White Appliances	96.88	96.18	93.68
Residual	65.38	63.95	53.05

Experiment 3 - Unseen appliances and unseen houses

In real life situation, the number of appliances operating in a building varies constantly and new undetermined signals can always be introduced. Therefore, the hypothesis that the aggregated signal is equal to the sum of predetermined set of appliances cannot be generalized. Therefore, we extend experiment 2 by supposing the following: the aggregated energy consumption X_i is equal to the sum of energy of well defined devices $y_1, y_2, ..., y_k$ for which we know the signature and a signal y_{k+1}. y_{k+1} can be the aggregated signal of multiple unknown devices, or just one new device that was not seen before. The addition of a new device would be systematically captured by this signal. Also, this would allow for the generalization of the proposed approach across different buildings that share a limited number of devices. This experiment is tested with IRISE dataset. We consider 5 appliances: dish washer, electrical cooker, fridge, light and white appliances that are common between houses 14, 38 and 48. We set X as the total energy consumption coming from all appliances within the house. House 48 has 13 appliances, house 38 has 16 and house 14 has 20 appliances. House 48 is used for the training. D is formed by individual signals of each appliance plus an additional set of signals that we refer to as residuals equal to the difference between X and specific appliances' energy. Appliance level F1-score from this experiment is reported in table 3.8. The performance of the selected appliances did not deteriorate. Since the residual signal is coming from a different sets of appliances in the train and test sets, the F1-score is lower for this hypothetical appliance compared to the others. However we are still capable of reconstructing it with a confidence of up to 65%.

Chapter 4

Conclusion

In this thesis, we have developed two Bayesian non negative matrix factorization models. We started by introducing and explaining Nonnegative matrix factorization. In the first part of this work, a Bayesian nonnegative matrix factorization approach is proposed. To model the nonnegativity constraint, we assume data follows an inverted Beta distribution. IB parameters are assigned Gamma prior thus the naming IBG-NMF. The model parameters were approximated using variational inference. Due to the integral properties, an analytically tractable solution could not be directly obtained. Therefore, we use a lower bound approximation to obtain an analytically tractable solution. An extension of this model is proposed with online learning using stochastic gradient ascent and natural gradients. Online IBG-NMF is shown to be more robust and more scalable than batch IBG-NMF. Setting small values for shape parameters allowed a sparse representation of NMF. Both models demonstrated a success in multiple important applications such as: Parts-based decomposition, collaborative filtering, market basket analysis, transactions prediction and items classification, topic mining and graph embedding on biomedical networks. The proposed models outperform state of the art models such as PMF and BPMF as well as modern models like BG-NMF.

While MCMC methods are capable of producing exact samples from the target density, our models (based on variational inference) were able to outperform MCMC based models. This could be explained by the fact that the proposed model is more suitable for the nature of the data [5]. However, given the complexity of the model, and the size of the datasets used for different applications, the usage of MCMC inference to solve this model would be very costly. The online learning approach is also more adapted to the nature of various applications such as collaborative filtering, transaction prediction, market basket analysis and topic mining, where large amounts of data are continuously arriving and retraining the model each time is inefficient if not impossible. For future work, the inference method could be further improved by using variational inference with Normalizing Flows [67].

We have proposed a novel matrix factorization model for non intrusive load monitoring. The model presents a supervised learning approach that uses historical devices consumption for energy disaggregation. The sparsity of the excitation matrix was imposed through a Dirichlet distribution. To learn the model's parameters a novel optimization approach was proposed for a variational learning problem with non conjugate priors. The usage of Bayesian learning allowed for robust results. We tested the model with different datasets coming from various types of buildings and with diverse appliances and measuring methods. The approach gives very competitive results compared to state of the art models. Basically, for the seen houses with high frequency, our model's performance is very close to state of the art methods. In addition, the model is less complex and converges quicker than deep learning based models. However, our model outstands when generalizing to unseen houses, low frequency domains and can capture new and unknown appliances. This robustness is guaranteed by the Bayesian approach.

Basically, we were able to tackle different challenges relative to energy disaggregation. Firstly, most of the state of the art approaches are either weak or computationally expensive when the number of appliances increases. According to experiment 1, our model is still as efficient even with a large set of appliances. Secondly, common smart meters collect data every 30 minutes to 1 hour. The problem with common NILM approaches is that are optimized to work with high frequency data. We experimented with a low frequency data in experiment 2 and we see that our model is robust against the frequency variation. Another important challenge with energy disaggregation is data collection. It is hard to collect ground truth data for every building. Therefore, it is very important to mention that our approach can learn from one building and performs energy disaggregation on another one. The appliance space is dynamically changing, our model was able to adapt to these changes as described in experiment 3.

Overall, we have presented an approach that can be used to address different NILM challenges.

The model is robust and flexible. However, we have noticed that the reconstruction error is high. This is due to the usage of raw signals coming from historical data. For future work, we would be interested in preprocessing the signal using dictionary learning for temporal data. Our model does not rely on additional features and does not require a complicated preprocessing analysis. This could be considered as an advantage especially for non domain experts. However, when additional features are available, such as cycle of each appliance during sliding window, energy, time, appliances characteristics, etc, an advanced version of this model would take into account these features to improve the results.

Appendix A

Variational Inference

The idea of variational inference [37, 4, 35], is to approximate the true posterior $p(\mathbf{Z}|\mathbf{X})$ by $q(\mathbf{Z})$. In the following parts we estimate the parameters that minimize the divergence between $q(\mathbf{Z})$ and $p(\mathbf{Z}|\mathbf{X})$ which is measured by the KL divergence $KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}))$:

$$KL\big(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})\big) = E_{q(\mathbf{Z})}\big[\ln\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\big] = \ln p(\mathbf{X}) - E_{q(\mathbf{Z})}\big[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\big] \tag{A.1}$$

From (A.1), our problem is equivalent to maximizing an objective function $\mathcal{L}(q) = E_{q(z)}[\ln(\frac{p(\mathbf{Z},\mathbf{X})}{q(\mathbf{Z})})]$. If we consider that the variables **A**, **B** and **H** are independent, and matrix elements are independent, we can consider a_{nk} as the only variable and fix the remaining elements of **Z**. The optimal solution $q^*(a_{nk})$ can be obtained by (A.2), where $\bar{x}$ denotes the expected value of x and $E_{\backslash q(a_{nk})}[X] = \int\prod_{(i,j)\neq(n,k)} Xq(a_{ij})da_{ij}$:

$$\begin{aligned}\ln q^*(a_{nk}) &= E_{\backslash q(a^*_{nk})}[\ln p(\mathbf{X},\mathbf{Z})] + const\\ &= \sum_m E_{\backslash q(a^*_{nk})}[-\ln\beta(\sum_k a_{nk}h_{km}, \sum_k b_{nk}h_{km})]\\ &+ \sum_m \bar{h}_{km}\ln x_{nm}a_{nk} - \sum_m \bar{h}_{km}\ln(1+x_{nm})a_{nk}\\ &+ (\mu_0 - 1)\ln(a_{nk}) - \alpha_0 a_{nk} + const\end{aligned} \tag{A.2}$$

Appendix B

Sparseness constraint

Imposing the sparseness constraint on the columns of **H** means that for a given column $\boldsymbol{h}_j$ of **H**

$$\exists 1 < k_j < k, h_{k_j j} = 1, h_{ij} = 0 i \neq k_j \tag{B.1}$$

$$\begin{aligned}\bar{x}_{nm} &= \frac{\sum_k a_{nk} h_{km}}{\sum_k b_{nk} h_{km} - 1} \\ &= \frac{a_{nk_m} h_{k_m m}}{b_{nk_m} h_{k_m m} - 1} \\ &= \frac{a_{nk_m}}{b_{nk_m} - 1} h_{k_m m}\end{aligned} \tag{B.2}$$

Bibliography

[1] ALQUIER, P., AND GUEDJ, B. An oracle inequality for quasi-bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics 26*, 1 (2017), 55–67.

[2] ANGELIS, G.-F., TIMPLALEXIS, C., KRINIDIS, S., IOANNIDIS, D., AND TZOVARAS, D. Nilm applications: Literature review of learning approaches, recent developments and challenges. *Energy and Buildings* (2022), 111951.

[3] ARORA, S., GE, R., HALPERN, Y., MIMNO, D., MOITRA, A., SONTAG, D., WU, Y., AND ZHU, M. A practical algorithm for topic modeling with provable guarantees. In *International conference on machine learning* (2013), PMLR, pp. 280–288.

[4] BISHOP, C. M. *Pattern recognition and machine learning*. springer, 2006.

[5] BLEI, D. M., KUCUKELBIR, A., AND MCAULIFFE, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association 112*, 518 (2017), 859–877.

[6] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[7] BODENREIDER, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research 32*, suppl_1 (2004), D267–D270.

[8] CARBONETTO, P., SARKAR, A., WANG, Z., AND STEPHENS, M. Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv preprint arXiv:2105.13440* (2021).

[9] Chen, D., Sain, S. L., and Guo, K. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management 19*, 3 (2012), 197–208.

[10] Chen, W.-S., Ge, X., and Pan, B. A novel general kernel-based non-negative matrix factorisation approach for face recognition. *Connection Science 34*, 1 (2022), 785–810.

[11] Cunningham, J. P., and Ghahramani, Z. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research 16*, 1 (2015), 2859–2900.

[12] Dalhoumi, O., Bouguila, N., Amayri, M., and Fan, W. Bayesian matrix factorization for semibounded data. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[13] Dalianis, H. Evaluation metrics and evaluation. In *Clinical text mining*. Springer, 2018, pp. 45–53.

[14] Darby, S., et al. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays 486*, 2006 (2006), 26.

[15] Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. The comparative toxicogenomics database: update 2017. *Nucleic acids research 45*, D1 (2017), D972–D978.

[16] de Almeida, A., Fonseca, P., Bandeirinha, R., Fernandes, T., Araújo, R., and Urbano, N. Remodece: residential monitoring to decrease energy use and carbon emissions in europe. *Final report* (01 2008).

[17] Févotte, C., and Cemgil, A. T. Nonnegative matrix factorizations as probabilistic inference in composite models. In *2009 17th European Signal Processing Conference* (2009), IEEE, pp. 1913–1917.

[18] Figueiredo, M., Ribeiro, B., and de Almeida, A. Electrical signal source separation via nonnegative tensor factorization using on site measurements in a smart home. *IEEE Transactions on Instrumentation and Measurement 63*, 2 (2013), 364–373.

[19] Ghahramani, Z., and Jordan, M. Factorial hidden markov models. In *Advances in Neural Information Processing Systems* (1995), D. Touretzky, M. Mozer, and M. Hasselmo, Eds., vol. 8, MIT Press.

[20] Golub, G. H., and Reinsch, C. Singular value decomposition and least squares solutions. In *Linear Algebra*. Springer, Berlin, Heidelberg, 1971, pp. 134–151.

[21] Gong, M., Jiang, X., Li, H., and Tan, K. C. Multiobjective sparse non-negative matrix factorization. *IEEE transactions on cybernetics 49*, 8 (2018), 2941–2954.

[22] Gu, R., Du, Q., and Billinge, S. J. A fast two-stage algorithm for non-negative matrix factorization in streaming data. *arXiv preprint arXiv:2101.08431* (2021).

[23] Guan, N., Tao, D., Luo, Z., and Yuan, B. Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems 23*, 7 (2012), 1087–1099.

[24] Harper, F. M., and Konstan, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis) 5*, 4 (2015), 1–19.

[25] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[26] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web* (2017), pp. 173–182.

[27] Henriet, S., Fuentes, B., Şimşekli, U., and Richard, G. Matrix factorization for high frequency non intrusive load monitoring: Definitions and algorithms. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring* (2020), pp. 20–24.

[28] HIDIYANTO, F., AND HALIM, A. Knn methods with varied k, distance and training data to disaggregate nilm with similar load characteristic. In *Proceedings of the 3rd Asia Pacific Conference on Research in Industrial and Systems Engineering 2020* (2020), pp. 93–99.

[29] HOFFMAN, M. D., BLEI, D. M., AND COOK, P. R. Bayesian nonparametric matrix factorization for recorded music. In *ICML* (2010).

[30] HOSSEINI-ASL, E., AND ZURADA, J. M. Nonnegative matrix factorization for document clustering: A survey. In *International Conference on Artificial Intelligence and Soft Computing* (2014), Springer, pp. 726–737.

[31] HOSSIN, M., AND SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process 5*, 2 (2015), 1.

[32] HOYER, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research 5*, Nov (2004), 1457–1469.

[33] HUANG, K., FU, X., AND SIDIROPOULOS, N. D. Anchor-free correlated topic modeling: Identifiability and algorithm. *Advances in Neural Information Processing Systems 29* (2016).

[34] HUBER, P., CALATRONI, A., RUMSCH, A., AND PAICE, A. Review on deep neural networks applied to low-frequency nilm. *Energies 14*, 9 (2021), 2390.

[35] JAAKKOLA, T. S., AND JORDAN, M. I. Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 1 (2000), 25–37.

[36] JIA, Y. W. Y., AND TURK, C. H. M. Fisher non-negative matrix factorization for learning local features. In *Proc. Asian conf. on comp. vision* (2004), Citeseer, pp. 27–30.

[37] JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S., AND SAUL, L. K. An introduction to variational methods for graphical models. *Machine learning 37*, 2 (1999), 183–233.

[38] KAPLAN, R. M., BURSTEIN, J., HARPER, M., AND PENN, G. Human language technologies: The 2010 annual conference of the north american chapter of the association for

computational linguistics. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2010).

[39] KATO, T., CHO, H. S., LEE, D., TOYOMURA, T., AND YAMAZAKI, T. Appliance recognition from electric current signals for information-energy integrated network in home environments. In *International Conference on Smart Homes and Health Telematics* (2009), Springer, pp. 150–157.

[40] KELLY, J., AND KNOTTENBELT, W. Neural nilm: Deep neural networks applied to energy disaggregation. In *Proceedings of the 2nd ACM international conference on embedded systems for energy-efficient built environments* (2015), pp. 55–64.

[41] KELLY, J., AND KNOTTENBELT, W. Does disaggregated electricity feedback reduce domestic electricity consumption? a systematic review of the literature. *arXiv preprint arXiv:1605.00962* (2016).

[42] KOLTER, J. Z., AND JOHNSON, M. J. Redd: A public data set for energy disaggregation research. In *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA* (2011), vol. 25, pp. 59–62.

[43] KOREN, Y., BELL, R., AND VOLINSKY, C. Matrix factorization techniques for recommender systems. *Computer 42*, 8 (2009), 30–37.

[44] LAI, Y., MA, X., XU, Y., LING, Y., DU, C., DU, J., ZHANG, Y., AND PING, Y. Positive data modeling using a mixture of mixtures of inverted beta distributions. *IEEE Access 7* (2019), 38146–38156.

[45] LEE, D. D., AND SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature 401*, 6755 (1999), 788–791.

[46] LEE, D. D., AND SEUNG, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (2001), pp. 556–562.

[47] LEFÈVRE, A., BACH, F., AND FÉVOTTE, C. Online algorithms for nonnegative matrix factorization with the itakura-saito divergence. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2011), IEEE, pp. 313–316.

[48] LI, L., LEBANON, G., AND PARK, H. Fast bregman divergence nmf using taylor expansion and coordinate descent. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012), pp. 307–315.

[49] LI, L., WU, L., ZHANG, H., AND WU, F. A fast algorithm for nonnegative matrix factorization and its convergence. *IEEE Transactions on Neural Networks and Learning Systems 25*, 10 (2014), 1855–1863.

[50] LI, S. Z., HOU, X. W., ZHANG, H. J., AND CHENG, Q. S. Learning spatially localized, parts-based representation. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (2001), vol. 1, IEEE, pp. I–I.

[51] LI, X., CUI, G., AND DONG, Y. Graph regularized non-negative low-rank matrix factorization for image clustering. *IEEE transactions on cybernetics 47*, 11 (2016), 3840–3853.

[52] LÜ, L., AND ZHOU, T. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications 390*, 6 (2011), 1150–1170.

[53] MA, Z., TESCHENDORFF, A. E., LEIJON, A., QIAO, Y., ZHANG, H., AND GUO, J. Variational bayesian matrix factorization for bounded support data. *IEEE transactions on pattern analysis and machine intelligence 37*, 4 (2014), 876–889.

[54] MAKONIN, S., POPOWICH, F., BARTRAM, L., GILL, B., AND BAJIĆ, I. V. Ampds: A public dataset for load disaggregation and eco-feedback research. In *2013 IEEE electrical power & energy conference* (2013), IEEE, pp. 1–6.

[55] MIYASAWA, A., FUJIMOTO, Y., AND HAYASHI, Y. Energy disaggregation based on smart metering data via semi-binary nonnegative matrix factorization. *Energy and Buildings 183* (2019), 547–558.

[56] Mnih, A., and Salakhutdinov, R. R. Probabilistic matrix factorization. *Advances in neural information processing systems 20* (2007), 1257–1264.

[57] Mohamed, S., and Lakshminarayanan, B. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483* (2016).

[58] Nalmpantis, C., and Vrakas, D. Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation. *Artificial Intelligence Review 52*, 1 (2019), 217–243.

[59] Okun, O., and Priisalu, H. Nonnegative matrix factorization for pattern recognition. In *Proceedings of the 5th IASTED International Conference on Visualization, Imaging and Image Processing* (2005), Citeseer, pp. 546–551.

[60] Ozerov, A., and Févotte, C. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing 18*, 3 (2009), 550–563.

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

[62] Paatero, P., and Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*, 2 (1994), 111–126.

[63] Paisley, J. W., Blei, D. M., and Jordan, M. I. Bayesian nonnegative matrix factorization with stochastic variational inference. In *Handbook of mixed membership models and their applications*. CRC press, 2014, ch. 11, pp. 205–224.

[64] Rahimpour, A., Qi, H., Fugate, D., and Kuruganti, T. Non-intrusive energy disaggregation using non-negative matrix factorization with sum-to-k constraint. *IEEE Transactions on Power Systems 32*, 6 (2017), 4430–4441.

[65] Rendle, S., Krichene, W., Zhang, L., and Anderson, J. Neural collaborative filtering vs. matrix factorization revisited. In *Fourteenth ACM conference on recommender systems* (2020), pp. 240–248.

[66] Revuelta Herrero, J., Lozano Murciego, Á., López Barriuso, A., Hernández de la Iglesia, D., Villarrubia González, G., Corchado Rodríguez, J. M., and Carreira, R. Non intrusive load monitoring (nilm): A state of the art. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (2017), Springer, pp. 125–138.

[67] Rezende, D., and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning* (2015), PMLR, pp. 1530–1538.

[68] Ruthotto, L., and Haber, E. An introduction to deep generative modeling. *GAMM-Mitteilungen 44*, 2 (2021), e202100008.

[69] Salakhutdinov, R., and Mnih, A. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning* (2008), pp. 880–887.

[70] Samaria, F. S., and Harter, A. C. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE workshop on applications of computer vision* (1994), IEEE, pp. 138–142.

[71] Schirmer, P. A., and Mporas, I. Statistical and electrical features evaluation for electrical appliances energy disaggregation. *Sustainability 11*, 11 (2019), 3222.

[72] Schirmer, P. A., and Mporas, I. Multivariate non-negative matrix factorization with application to energy disaggregation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 3285–3289.

[73] Schmidt, M. N., Winther, O., and Hansen, L. K. Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation* (Berlin, Heidelberg, 2009), Springer, pp. 540–547.

[74] Shahnaz, F., Berry, M. W., Pauca, V. P., and Plemmons, R. J. Document clustering using nonnegative matrix factorization. *Information Processing & Management 42*, 2 (2006), 373–386.

[75] Shi, J., Zheng, X., and Yang, W. Survey on probabilistic models of low-rank matrix factorizations. *Entropy 19*, 8 (2017), 424.

[76] Sorzano, C. O. S., Vargas, J., and Montano, A. P. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877* (2014).

[77] Sra, S., and Dhillon, I. S. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in neural information processing systems* (2006), pp. 283–290.

[78] Stone, J. V. Independent component analysis: an introduction. *Trends in cognitive sciences 6*, 2 (2002), 59–64.

[79] Sun, C., Zhu, Q., and Wan, M. A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition. *Speech Communication 60* (2014), 44–55.

[80] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research 43*, D1 (2015), D447–D452.

[81] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical dirichlet processes. *Journal of the american statistical association 101*, 476 (2006), 1566–1581.

[82] Vasilescu, M. A. O., and Terzopoulos, D. Multilinear independent components analysis. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 1, IEEE, pp. 547–553.

[83] Vasiloglou, N., Gray, A. G., and Anderson, D. V. Non-negative matrix factorization, convexity and isometry. In *Proceedings of the 2009 SIAM International Conference on Data Mining* (2009), SIAM, pp. 673–684.

[84] VAVASIS, S. A. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization 20*, 3 (2010), 1364–1377.

[85] VLACHOS, M., DOMENICONI, C., GUNOPULOS, D., KOLLIOS, G., AND KOUDAS, N. Non-linear dimensionality reduction techniques for classification and visualization. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (2002), pp. 645–651.

[86] VOULODIMOS, A., DOULAMIS, N., DOULAMIS, A., AND PROTOPAPADAKIS, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience 2018* (2018).

[87] WALL, M. E., RECHTSTEINER, A., AND ROCHA, L. M. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.

[88] WANG, Y., AND BLEI, D. Variational bayes under model misspecification. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 13357–13367.

[89] WANG, Y., JIA, Y., HU, C., AND TURK, M. Non-negative matrix factorization framework for face recognition. *International Journal of Pattern Recognition and Artificial Intelligence 19*, 04 (2005), 495–511.

[90] WANG, Y.-X., AND ZHANG, Y.-J. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering 25*, 6 (2012), 1336–1353.

[91] WISHART, D. S., FEUNANG, Y. D., GUO, A. C., LO, E. J., MARCU, A., GRANT, J. R., SAJED, T., JOHNSON, D., LI, C., SAYEEDA, Z., ET AL. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research 46*, D1 (2018), D1074–D1082.

[92] WOLD, S., ESBENSEN, K., AND GELADI, P. Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1-3 (1987), 37–52.

[93] WU, J.-S., LAI, J.-H., AND WANG, C.-D. A novel co-clustering method with intra-similarities. In *2011 IEEE 11th International Conference on Data Mining Workshops* (2011), IEEE, pp. 300–306.

[94] XIAO, Z., GANG, W., YUAN, J., ZHANG, Y., AND FAN, C. Cooling load disaggregation using a nilm method based on random forest for smart buildings. *Sustainable Cities and Society 74* (2021), 103202.

[95] XU, W., LIU, X., AND GONG, Y. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), pp. 267–273.

[96] YUE, X., WANG, Z., HUANG, J., PARTHASARATHY, S., MOOSAVINASAB, S., HUANG, Y., LIN, S. M., ZHANG, W., ZHANG, P., AND SUN, H. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics 36*, 4 (2020), 1241–1251.

[97] ZEIFMAN, M., AND ROTH, K. Nonintrusive appliance load monitoring: Review and outlook. *IEEE transactions on Consumer Electronics 57*, 1 (2011), 76–84.

[98] ZHANG, C., NIE, F., AND XIANG, S. A general kernelization framework for learning algorithms based on kernel pca. *Neurocomputing 73*, 4-6 (2010), 959–967.

[99] ZHANG, C., ZHONG, M., WANG, Z., GODDARD, N., AND SUTTON, C. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.

[100] ZHAO, Q., MENG, D., XU, Z., ZUO, W., AND YAN, Y. l_1 -norm low-rank matrix factorization by variational bayesian method. *IEEE Transactions on Neural Networks and Learning Systems 26*, 4 (2015), 825–839.

[101] ZHAO, X., LI, X., ZHANG, Z., SHEN, C., ZHUANG, Y., GAO, L., AND LI, X. Scalable linear visual feature learning via online parallel nonnegative matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems 27*, 12 (2016), 2628–2642.

[102] Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics 10*, 5 (2021), 593.

[103] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*, 2 (2005), 301–320.